

Société

REPRENDRE LE CONTRÔLE DES RÉSEAUX SOCIAUX

Paula Forteza

08/12/2020

Assassinat de Samuel Paty, documentaire Hold-Up, allégations de fraude lors des dernières élections américaines... L'actualité récente a remis en évidence la question majeure de la diffusion de contenus problématiques sur les réseaux sociaux. À l'occasion de la mise au débat du projet de loi confortant les principes républicains en France et du Digital Services Act au niveau communautaire, Paula Forteza propose deux axes de régulation : la régulation par la donnée, et par la société.

Ces questions ont déjà été débattues par deux fois au Parlement français depuis 2017. D'abord, en juillet 2018, lors de la proposition de loi contre la manipulation de l'information, qui a fait l'objet d'un débat agité et de réserves d'interprétation de la part du Conseil constitutionnel concernant, notamment, la définition des fausses nouvelles. Puis, en avril 2019, au moment de la proposition de loi contre les contenus haineux sur Internet, mettant en place une obligation de retrait de contenus par les plateformes, dont les dispositions ont été quasi intégralement censurées par le Conseil constitutionnel.

Parce que la troisième doit être la bonne, cette note vise à poser les termes du débat différemment. Contrairement à une approche tendant à assimiler les réseaux sociaux à des médias pour leur imposer des obligations éditoriales et les responsabiliser sur des contenus individuels, je défends l'idée de les apparenter à des « espaces publics » dont ils ont, en accueillant des millions de citoyens français sans barrières à l'entrée, adopté les principales caractéristiques. Ceci permet de préserver le meilleur des réseaux sociaux, l'expression libre et décentralisée, tout en attribuant une responsabilité d'intérêt général, susceptible de renverser le secret des affaires. Élargir au-delà des propos manifestement illégaux le périmètre des contenus spécifiques dont les plateformes sont responsables n'est pas une solution soutenable et efficace à long terme. Ce n'est qu'au niveau structurel qu'un cadre de délibération et d'expression apaisé, divers et transparent peut être consolidé.

Camille François, experte française de la manipulation de l'information et dont les travaux sur les

techniques de désinformation russes durant la campagne présidentielle américaine de 2016 font référence, propose un cadre théorique pour penser cette question. C'est ce qu'elle appelle le « ABC framework » (cadre ABC), qui représente trois niveaux d'action sur la question de la régulation des réseaux sociaux :

- A comme « Actor », soit l'acteur, celui qui crée le contenu et le partage. Cela peut être un influenceur, un groupe de particuliers ou un État. La question de l'intentionnalité et de la responsabilité de l'auteur est déjà encadrée par notre droit national et communautaire. Il reste donc ici qu'à renforcer les moyens d'action de la justice ;
- B comme « Behavior », soit le comportement de l'information en ligne, son degré de viralité et les techniques employées pour propager l'information (algorithmes, chatbot automatisés, fermes de « troll », publicité ciblée, etc.). Le focus sur le comportement est technique. Il nécessite d'avoir une connaissance extrêmement poussée des technologies employées et des dynamiques algorithmiques qui font qu'un contenu devient viral ;
- C comme « Content », soit le contenu partagé, le message véhiculé. La régulation du contenu, *a priori* ou *a posteriori*, est le niveau le plus délicat en matière de respect de la liberté d'expression.

C'est le niveau d'action B qu'il nous faut dorénavant investir. Je détaille, dans cette note, une approche de la régulation du comportement et de la viralité des contenus en ligne. L'objectif est de faire en sorte de lutter contre la propagation rapide de certains contenus problématiques plutôt que de vouloir les censurer dès leur apparition (niveau C).

La régulation de la viralité nécessite une plus grande transparence des algorithmes et un meilleur accès aux données des plateformes, condition *sine qua non* d'un pouvoir nouveau donné au régulateur. Ce n'est qu'en ouvrant la « boîte noire » et en comprenant l'impact des réseaux sociaux sur nos sociétés que nous pourrions détailler des spécifications techniques souhaitables. Il est indispensable, aussi, de développer des anticorps dans la société civile, en investissant le modèle de modération communautaire, à l'image de ce qui est pratiqué par les communautés de Wikipédia, par exemple. Enfin, nous pouvons décroisonner les auditoires captifs et donner plus de place au débat contradictoire en appliquant, notamment, le principe du droit de réponse prévu par la loi de 1881 aux comptes et pages en ligne.

Sur ces questions, la solution magique, pouvant tenir en quelques dispositions législatives, n'existe pas. Les propositions détaillées dans cette note sont des pistes de réflexion, qu'il faudra tester, expérimenter, enrichir de façon itérative. Il faudra pour cela organiser la collaboration des pouvoirs publics, des régulateurs, des élus, des réseaux sociaux, des chercheurs, de la société civile, des

citoyens. Nous devons accepter que ce travail s'inscrive dans la durée.

Les réseaux sociaux ne sont pas des médias, ce sont des espaces publics

La responsabilité des contenus en ligne

Quelle responsabilité devons-nous accorder à un réseau social sur les contenus partagés par les utilisateurs ? Cette question sous-tend l'ensemble des débats sur la modération des contenus en ligne. Au centre de ces débats s'articule le choix entre les statuts d'éditeur et d'hébergeur de contenus.

Les positions ont sensiblement évolué au cours des vingt dernières années.

Tout d'abord, la jurisprudence se fondait sur la responsabilité de l'hébergeur. L'affaire Estelle Hallyday contre Altern constitue un tournant. En 2000, des photos d'Estelle Hallyday dénudées publiées dans le magazine *Voici* sont postées sur un des blogs hébergés par la plateforme Altern. La plaignante demande le retrait de ses images et demande à engager la responsabilité de l'hébergeur dans la diffusion de ces images. Le juge tranche en faveur de la plaignante pour acter de la responsabilité d'Altern dans le contrôle du contenu de ses 45 000 blogs. À la suite de cette décision, Altern fera faillite.

À partir de 2000, le législateur introduit la distinction entre hébergeur et éditeur. C'est d'abord la directive européenne e-Commerce de 2000, puis la loi française sur la confiance dans l'économie numérique de 2004 qui l'encadrent. La responsabilité de l'éditeur ne désengage pas complètement les plateformes ; elles doivent mettre tout en œuvre pour enlever un contenu illégal au regard de la loi nationale dès lors qu'il lui est signalé. Des boutons de signalement (*notice*) sont mis en place pour permettre aux plateformes d'identifier et retirer les contenus (*takedown*).

À mesure que les réseaux sociaux se sont développés, la responsabilité pressentie des plateformes s'est accentuée. Les pouvoirs publics ont cherché à faire porter aux réseaux sociaux une responsabilité morale et légale sur la modération proactive des contenus qui se situent dans la « zone grise » entre le licite et l'illicite. C'est la voie adoptée notamment par la proposition de loi contre les contenus haineux sur Internet. Cette approche présente le risque de remplacer la décision du juge par celle de la plateforme. C'est ce que le Conseil constitutionnel a relevé dans sa décision du 18 juin 2020. Le Digital Services Act qui sera rendu public le 9 décembre 2020 et qui

doit venir actualiser la directive e-Commerce pourrait revoir la procédure du *notice and takedown* pour le substituer à celle du *notice and action* : anticiper le signalement pour enlever des contenus de manière préventive. Cette approche devra rester strictement réservée aux contenus manifestement illicites, au risque de tomber dans la même impasse.

Dès lors, l'opinion publique et les responsables politiques tendent à rapprocher les réseaux sociaux des médias classiques en cherchant à leur faire porter une responsabilité éditoriale, voire à leur octroyer le statut d'éditeur. Encore une fois, nous regardons le doigt et pas la lune. En nous concentrant exclusivement sur le retrait des contenus et en voulant étendre la responsabilité des plateformes sur les contenus individuels, nous frôlons constamment l'inconstitutionnalité, nous nous enlisons dans des exercices de définition, qualification et typologies de contenus sans fin, et nous risquons, à terme, de rendre invivable le modèle d'expression décentralisée des réseaux sociaux. La réponse ne pourra être que structurelle et viser les dynamiques de viralité.

Réseaux sociaux, espace public et intérêt général

Henri Verdier, ambassadeur du numérique, a décrit l'arrivée des plateformes comme de nouvelles formes d'enclosures consistant à mettre des « barbelés sur la prairie d'Internet ». Cette formule est bien trouvée et permet de décrire certaines logiques de privatisation de ce bien commun qu'est Internet. Cependant, elle laisse penser qu'Internet serait un espace fini, à l'instar d'un espace physique (les prairies) et que la logique première de ces géants aurait été celle d'enfermer (les barbelés).

La dynamique est plus retorse : le choix des réseaux sociaux a été, au contraire, celui de ne mettre aucune barrière à l'entrée, leur *business model* pouvant être résumé par le fameux adage « si c'est gratuit, c'est vous le produit ». Ils ont ainsi réussi à attirer une portion très importante de nos concitoyens, qui échangent et débattent en ligne : 37 millions d'utilisateurs Français sur Facebook, 21 millions sur Instagram, 12 millions sur Twitter.

Ils revêtent, de ce fait et en droit, les caractéristiques d'un espace public. En droit public, l'espace public est une notion en constante évolution. Cette notion a été dernièrement débattue lors de l'examen de la [loi du 11 octobre 2010](#) interdisant la dissimulation du visage dans l'espace public : « l'espace public est constitué des voies publiques ainsi que des lieux ouverts au public ou affectés à un service public ». Ici, l'espace public est un lieu où se déroulent les interactions, les échanges et la circulation des biens et des personnes. Ainsi l'espace public comprend aussi des espaces privés mais ouverts au public sans restriction.

D'un point de vue philosophique, l'espace public est l'endroit où se déroule la délibération collective. Des philosophes grecs et leur notion d'agora, en passant par le serment du Jeu de paume (« Partout où ses membres sont réunis, là est l'Assemblée nationale »), jusqu'à Hannah Arendt, nous comprenons que l'espace public existe dès que citoyennes et citoyens se rassemblent pour échanger sur les affaires publiques. Dans cette approche, c'est bien la « publicité » des informations, leur circulation sans entrave, la possibilité de débattre qui définit la qualité d'espace public. Comme le résume parfaitement le sociologue Dominique Cardon, les réseaux sociaux ont remis « en cause la frontière qui a longtemps séparé les institutions de l'espace public, les médias et les industries culturelles, d'une part, de la conversation du public, d'autre part ».

Dans une perspective politique, la formation de l'espace public a aussi été définie comme un processus progressif d'émancipation du contrôle de l'autorité étatique. Jürgen Habermas qualifie l'espace public comme « le processus au cours duquel le public constitué d'individus faisant usage de leur raison s'approprie la sphère publique contrôlée par l'autorité et la transforme en une sphère où la critique s'exerce contre le pouvoir de l'État ». Inutile ici de rappeler combien Facebook en particulier, ainsi que Twitter, ou encore les différents hébergeurs de blog, ont été cruciaux dans le renversement de régimes autoritaires. On pense instinctivement aux « printemps arabes » de 2011. De Tunis à Damas, en passant par Le Caire, les réseaux sociaux ont permis la diffusion de l'information et l'organisation des révolutionnaires. Le mouvement de surveillance des réseaux sociaux provient directement de cette réaction des États dictatoriaux, identifiant leur faiblesse vitale dans la capacité d'organisation des masses en ligne.

Les réseaux sociaux sont donc des espaces publics, pas des médias. Ils doivent de ce fait répondre à des objectifs d'intérêt général. En droit du numérique, le secret des affaires prend souvent le dessus sur la notion d'intérêt général. La démonstration est faite : lorsque 53% de la population française a accédé à un lieu ouvert au public sans restriction pour s'y rencontrer et y débattre, ça s'appelle un espace public et c'est l'intérêt général qui prime. Il ne nous reste maintenant qu'à y construire des « parcs publics en ligne », que nombreux activistes décrivent comme des espaces informationnels non soumis à des logiques marchandes et protégés de toutes publicités.

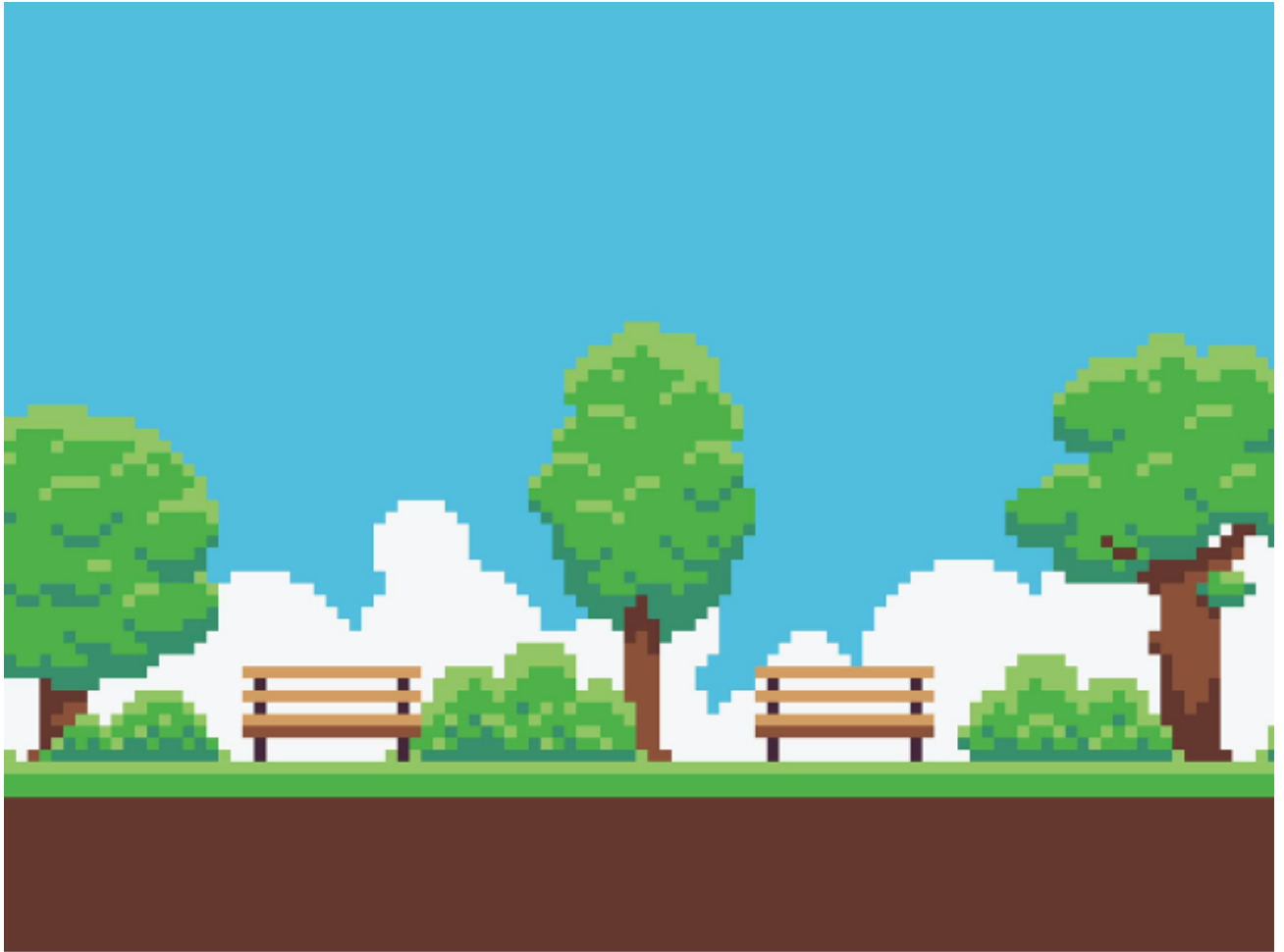


Illustration : Getty Images.

Comprendre les dynamiques de diffusion et de viralité sur les réseaux sociaux

« Bulles de filtre » et économie de l'attention

Une partie de la solution à l'exposition aux contenus problématiques – qu'ils soient à caractère haineux ou liés à la propagation de fausses informations – se situe dans la compréhension et la régulation des dynamiques de diffusion et de viralité.

Aujourd'hui, notre parcours utilisateur et notre consommation d'informations en ligne sont majoritairement façonnés par des algorithmes de recommandation, de classement ou de tri. Les utilisateurs ne sont pas assez sensibilisés à l'impact des algorithmes dans leur quotidien. Cet impact est aujourd'hui de plus en plus documenté, mais les principales clés de compréhension

restent abritées sous le secret des affaires.

Les « bulles de filtre » sont un concept développé par l'hacktiviste Eli Pariser. Les algorithmes sélectionnent les contenus que nous voyons et tendent à nous suggérer des informations qui nous ressemblent, socialement, politiquement, culturellement. Nous savons, par exemple, que l'algorithme de Twitter ne présente plus les contenus de manière antéchronologique mais selon une pondération en fonction de l'engagement du tweet (like, partage), la fréquence d'engagement avec l'auteur du tweet, le temps passé à lire le tweet, l'âge du tweet, le média contenu dans le tweet et la propension du lecteur à interagir avec le média (le partager à nouveau, le liker).

Plusieurs études et articles ont été publiés sur le rôle de Facebook dans l'émergence et la diffusion du mouvement des « gilets jaunes » en France. Fabrice Epelboin [défend l'idée que le changement d'algorithme de Facebook en 2018](#) – qui favorise le contenu local et publié par nos proches sur nos fils d'actualité – a largement contribué à la recrudescence du mouvement des « gilets jaunes » en renforçant l'effet « bulle de filtre ». D'autres rétorquent que le changement d'algorithme n'aurait eu lieu qu'aux États-Unis et que l'argumentaire serait donc faux. Mais comment en être certains ? Nous sommes ici face à de véritables boîtes noires.

Les thèses de l'« économie de l'attention » expliquent que les plateformes cherchent à captiver notre attention pour optimiser les effets de la publicité ciblée. Le documentaire *The Social Dilemma*, par exemple, présente la manière dont les principales plateformes développent des algorithmes et des fonctionnalités (comme l'*infinite scrolling* ou l'*autoplay*) pour maintenir les utilisateurs captifs.

Un autre exemple concerne la diffusion des fausses nouvelles ou des théories conspirationnistes sur YouTube. [Guillaume Chaslot](#), ancien employé de Google, [se concentre aujourd'hui sur l'identification des effets pervers](#) des algorithmes de recommandation, estimant qu'ils sont majoritairement conçus pour optimiser le temps passé par les utilisateurs sur leurs services et favorisent donc les contenus « polémiques ». Une vidéo expliquant que la Terre est plate sera plus regardée et davantage recommandée par la plateforme qu'une vidéo expliquant qu'elle est ronde. Face à ces déclarations, Google a affirmé que la méthode utilisée par leur ancien employé n'était pas fiable. Mais, encore une fois, sans transparence, comment savoir qui a raison ?

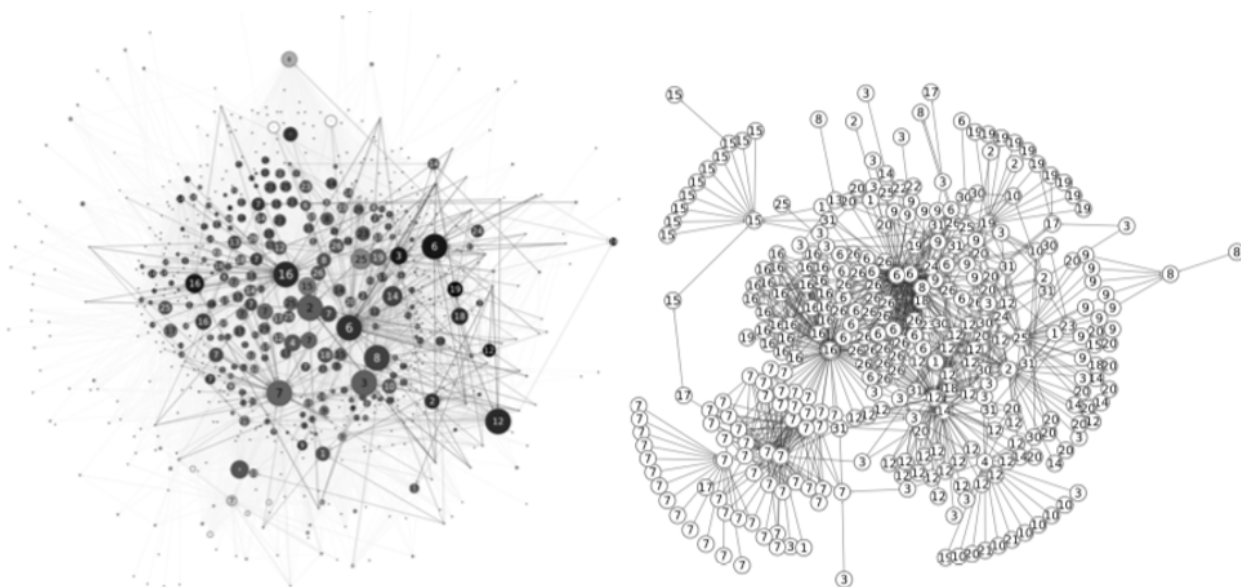
Bots et faux comptes

Une stratégie de viralité bien identifiée consiste à multiplier les faux comptes et les bots pour amplifier un contenu. Cette technique a été utilisée, par exemple, par des hackers russes lors de

l'élection présidentielle américaine ou encore pendant la campagne du Brexit.

La détection de ces techniques de manipulation – dont l'utilisation a fait l'objet d'une réflexion doctrinale très poussée en matière de cyberguerre – est permise par des modèles qui examinent le réseau des comptes qui partagent des contenus similaires, la régularité des publications (qui peut dévoiler leur caractère automatique), le vocabulaire utilisé, etc. Même les heures de publication ont permis de montrer le rôle des services secrets sud-coréens lors de la campagne présidentielle américaine de 2012, les tweets étant généralement postés aux heures de bureau des fonctionnaires.

Graphiques montrant le réseau des retweet de comptes détenus par des agents du NIS sud-coréen : ils font apparaître une forte densité, caractéristique des opérations de bots automatisés.



En France, l'ambassadeur du numérique Henri Verdier et ses équipes ont mis en place **des outils** pour lutter contre la propagation des « fake news », comme un détecteur de robot pour identifier des groupes de comptes Twitter suspects à partir d'un compte donné, ainsi qu'un chatbot collaboratif pour partager les meilleures solutions de lutte contre les bots.

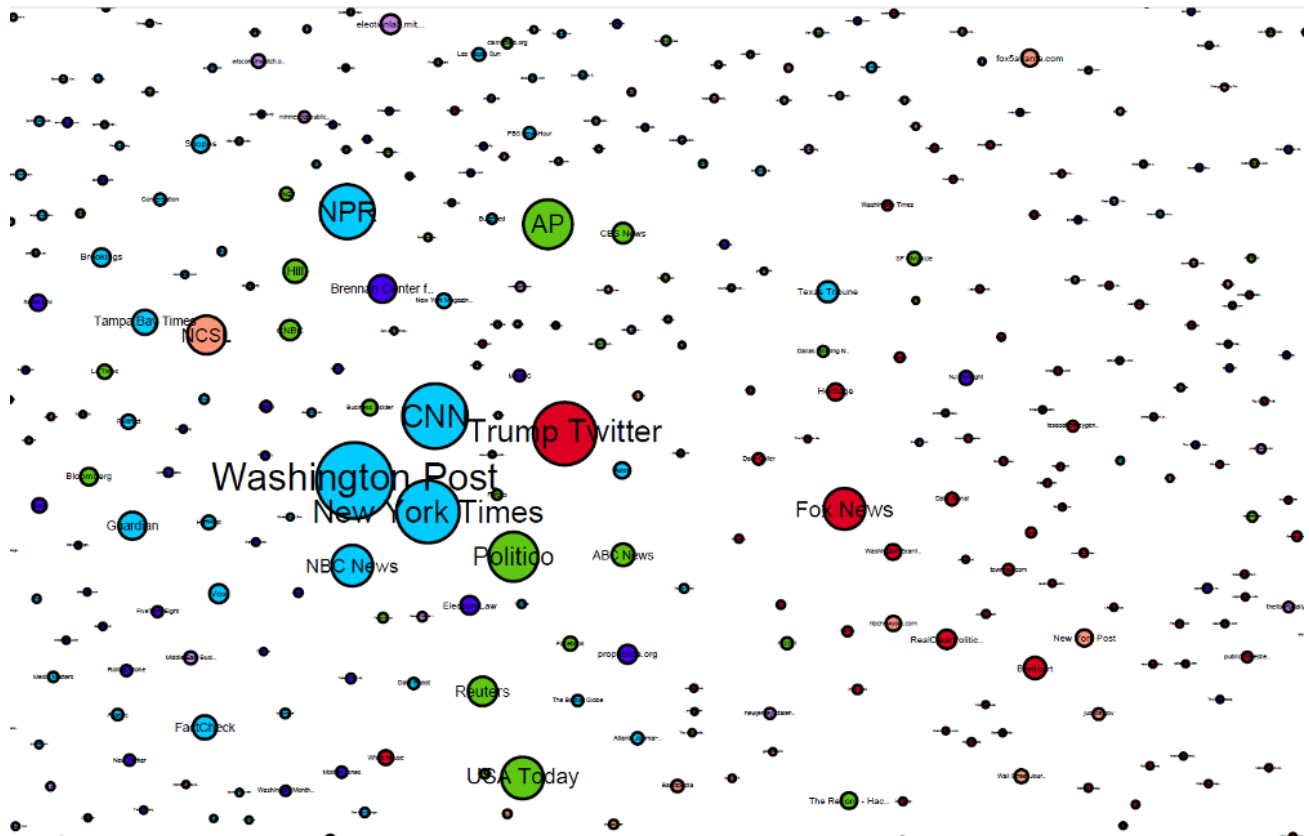
De la périphérie vers le centre : le rôle des médias *mainstream*

De manière certainement contre-intuitive, ce sont parfois les médias *mainstream* ou les comptes « certifiés » ayant une très forte notoriété qui propulsent des contenus douteux. C'est le paradoxe du « fact-checking » : alors même qu'il entend « débunker » (démystifier) une information problématique en la réfutant, il lui donne en même temps une plus grande visibilité.

Un cas a particulièrement mis en exergue cette question : l'affaire de l'« emoji fourchette ». En juillet 2020, une partie des réseaux sociaux se sont enflammés en prétendant que l'« emoji fourchette » avait été retiré par Twitter. Alors même que cet emoji n'a en réalité jamais existé, cette fausse information a été diffusée par des réseaux de la fachosphère pour discréditer la mobilisation liée à l'affaire Adama Traoré. Adama Traoré avait été accusé par un codétenu d'agressions sexuelles et de menaces en utilisant une fourchette. En réalité, les propagateurs de cette « fake news » ont construit de toutes pièces une campagne de désinformation, en anticipant le rôle d'amplificateur des « fact-checker » qui allaient ainsi relayer de manière beaucoup plus importante leur message.

Un autre exemple est l'élection américaine de 2020. Le Berkman Center de Harvard a bien mis en évidence combien les allégations de fraude électorale liées au vote par correspondance provenaient très majoritairement du compte twitter de Donald Trump lui-même et de la chaîne de télévision Fox News. La part d'influence dans la production de contenu des médias de l'« alt-right » ou encore des hackers russes reste **très minoritaire**.

Carte des médias en ligne

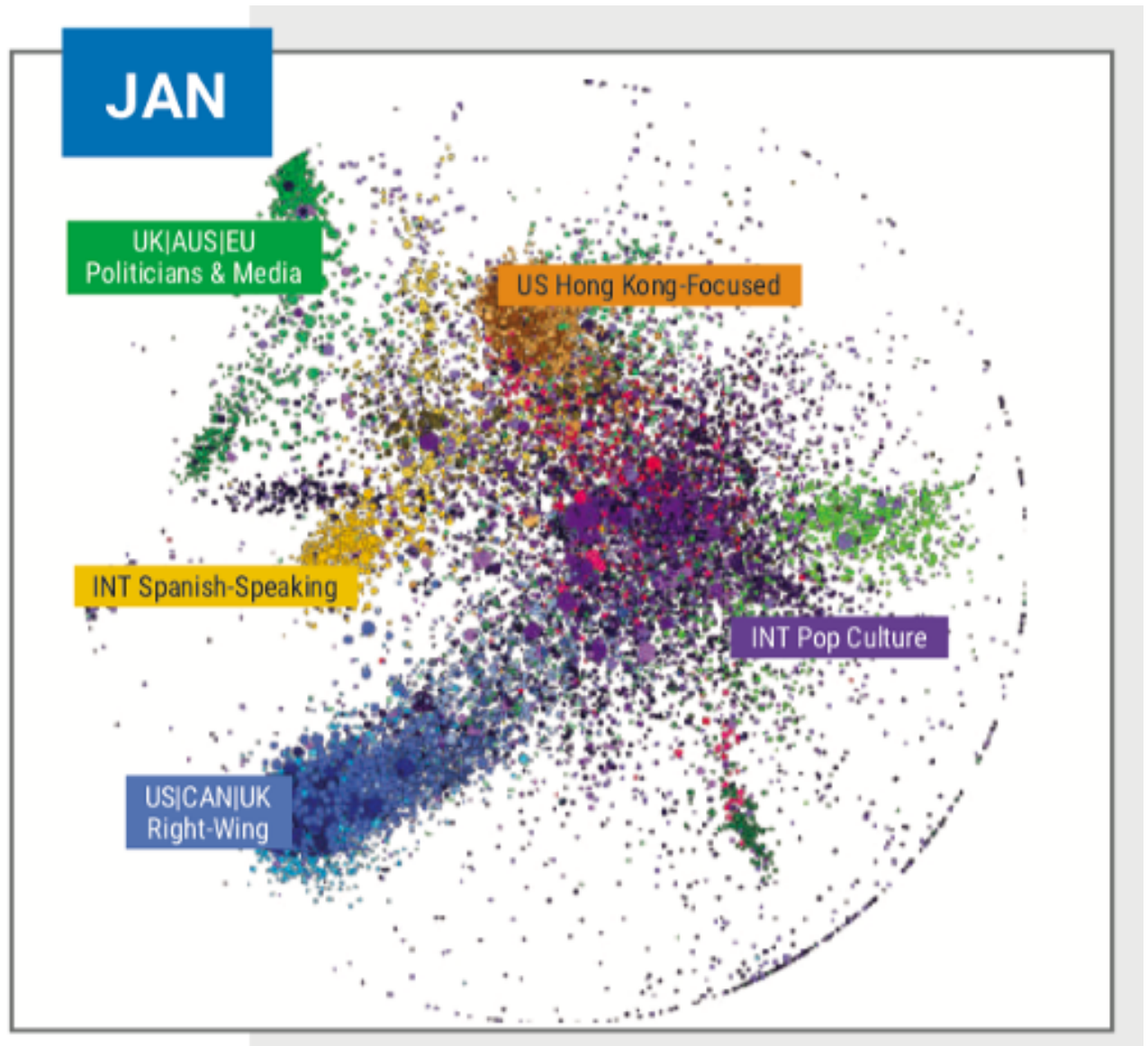


La taille des points montre l'importance des publications sur le sujet de la fraude électorale lors des élections américaines de 2020. La proximité entre les points permet de visualiser les

interdépendances entre les informations partagées. Ainsi, le compte twitter de Donald Trump et celui de CNN sont relativement proches, parce que l'un et l'autre se sont mutuellement cités à plusieurs reprises. Source : [Berkman Center, Harvard University](#).

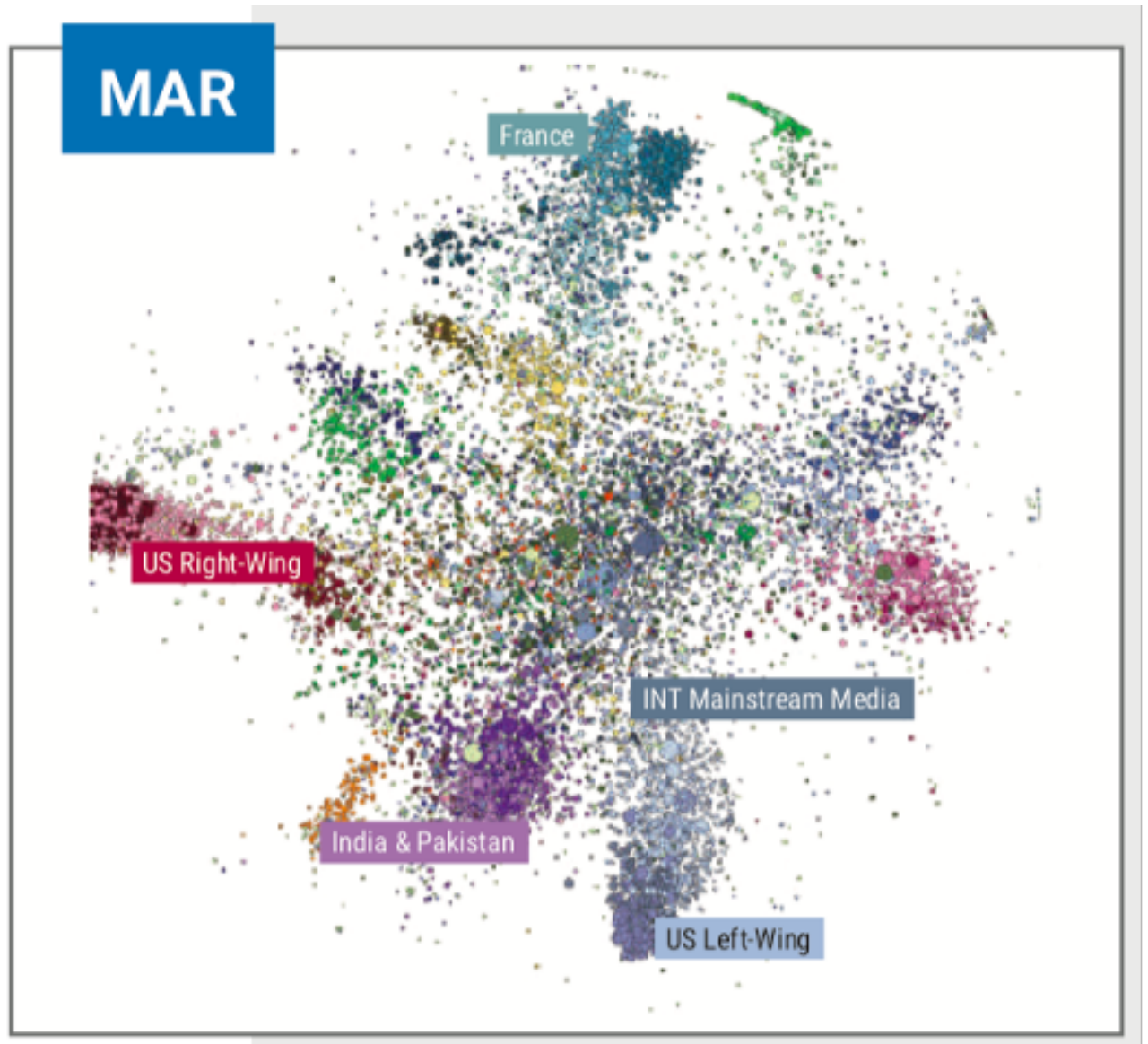
On retrouve des dynamiques similaires de viralité dans le cas de la propagation des informations liées à la crise de la Covid-19. Graphika, un centre de recherche indépendant qui cartographie les interactions sociales sur les réseaux sociaux, a rendu un rapport montrant la diffusion de fausses informations dans les premiers mois de la crise de janvier à mars. Au début de la crise en janvier, les fausses informations circulent principalement dans les réseaux de la droite américaine (« alt-right », QAnons, etc.). Progressivement, au cours des semaines qui suivent, les contenus de désinformation s'étendent à d'autres sphères et la part de ces groupes recule considérablement (de 20% à 6% de l'émission totale de « fake news »). La part des comptes et des médias *mainstream* partageant ces informations a pris le relais, comme le montrent les deux cartes ci-dessous extraites du rapport.

Carte des comptes à l'origine de la diffusion de « fake news » sur la Covid-19, en janvier 2020



Source : Graphika.

Carte des comptes à l'origine de la diffusion de « fake news » sur la Covid-19, en mars 2020



Source : Graphika.

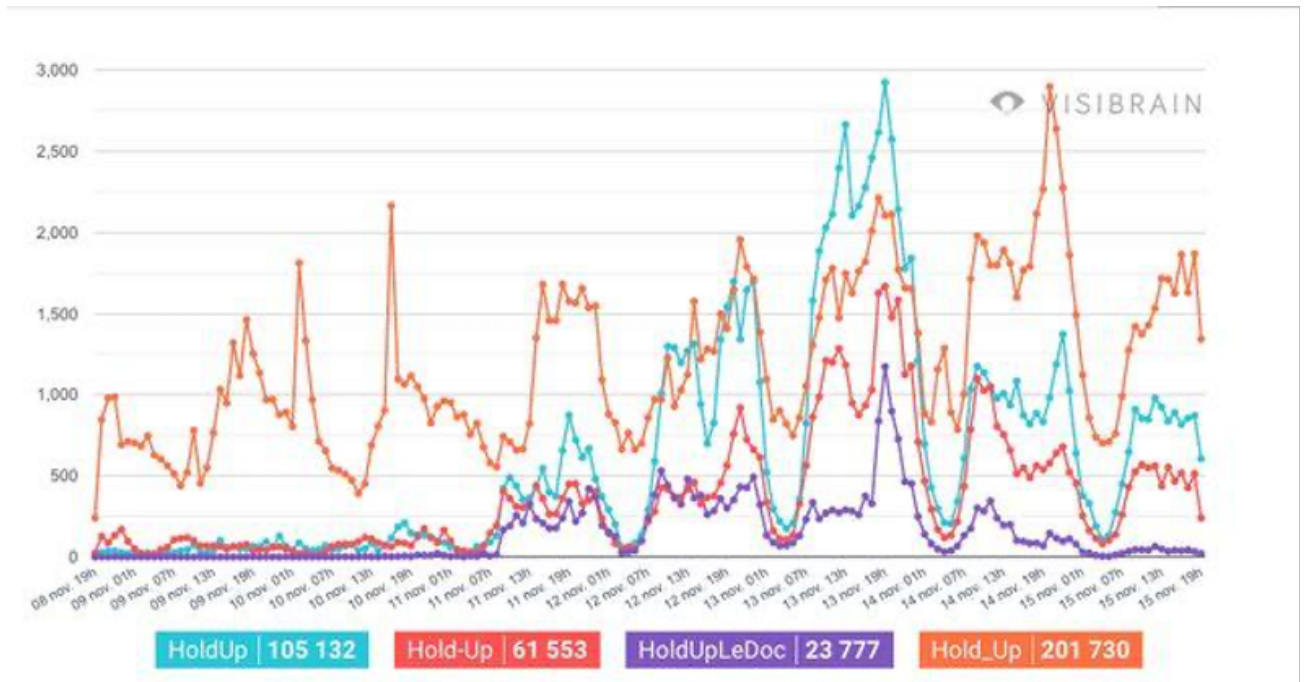
Même si les médias classiques ne reprennent pas à leur compte les thèses complotistes, le fait de les évoquer participe à leur diffusion.

Les canaux de la désinformation et de l'information ne sont pas hermétiques. Il faut, par exemple, s'intéresser à la manière dont Fdesouche.com, site emblématique de la fachosphère française, fonctionne. Ils ne créent pas de contenu propre. Leur action consiste uniquement à reprendre des contenus de médias traditionnels ou de personnalités et de les éditorialiser. Ainsi, ils entretiennent une bataille culturelle en utilisant les techniques de leurs adversaires.

Plus récemment, il est très probable que l'écho du documentaire controversé *Hold-Up* n'aurait jamais été possible si les médias ne s'étaient pas emparés du sujet et mis en place une course au

« débunking ». Grâce à un effet Streisand très fort suite à son retrait de la plateforme de vidéo en ligne Vimeo, le documentaire a été visionné plus de 2 millions de fois et fait l'objet de plus de 400 000 tweets en quelques heures.

Progression des mots-clés liés à *Hold-Up* 8 novembre-15 novembre



Source : Visibrain.

Le « dark social » : la vitalité invisible

Hold-Up constitue un cas typique des nouvelles formes de propagation de contenus. Une grande partie de la viralité s'est organisée dans les chaînes de conversations privées, dans une viralité de proche en proche.

La diffusion sur les réseaux sociaux constitue la partie émergée de l'iceberg informationnel. Selon le GlobalWebIndex, en 2019, les messageries privées sont privilégiées aux réseaux sociaux publics pour le partage de contenu. WhatsApp est utilisé par 1,5 milliard d'utilisateurs, Messenger par 1,3 milliard, WeChat par 1 milliard, Telegram par 200 millions.

L'importance du « dark social » – concept popularisé par le journaliste Alexis C. Madrigal – croît à mesure que des contenus sont censurés sur les réseaux sociaux. Ainsi, WhatsApp a été abondamment utilisé par les partisans du candidat Jair Bolsonaro pour le faire élire lors de la campagne présidentielle de 2018. Dans ces conditions, il est quasiment impossible de détecter ou

bloquer des contenus jugés comme problématiques dans des moments aussi critiques que des campagnes électorales.

Avec les messageries privées, nous abordons un autre problème juridique en termes de régulation : celui du droit de la correspondance, qui protège le secret des échanges entre deux personnes.

La régulation par la donnée

Transparence des algorithmes et accès aux données

Il existe un chemin permettant de concilier les libertés fondamentales (liberté d'expression, protection de la vie privée) avec la lutte contre les discours haineux ou dangereux pour la préservation de la démocratie. La première pierre d'une nouvelle régulation doit passer par une transparence plus importante des algorithmes et une capacité du régulateur d'auditer les données des plateformes.

Comme l'indique Dominique Cardon, la rétro-ingénierie est une technique possible de régulation des algorithmes lorsque les effets sont prévisibles par la plateforme mais non identifiables par l'utilisateur. Quand les effets de l'algorithme ne sont pas prévus par la plateforme et pas identifiables par l'utilisateur, la seule solution consiste à auditer les données. Ces dernières années, Facebook a ainsi mis à disposition de certains chercheurs, notamment ceux du médialab à Sciences Po, quelques jeux de données très ciblés. Néanmoins, ces jeux de données restent, selon les chercheurs ayant pu les consulter, assez limités et difficiles à traiter. Par exemple, les données utilisateurs, qui sont strictement anonymisées, ne permettent pas de prendre en compte le degré de visibilité des comptes et donc l'audience et l'impact des contenus partagés.

Les effets des algorithmes et leur type de régulation

EFFET...	...identifiable par l'utilisateur		...non identifiable par l'utilisateur	
	Exemple	Type de régulation	Exemple	Type de régulation
...prévisible par la plateforme	Fil d'actualité Facebook	Éducation	Google Shopping	Rétro- ingénierie
...non prévisible par la plateforme	Autocomplétion	Critique et médias	Discrimination	Audit des données

Source : Dominique Cardon, *Culture numérique, Paris, Presses de Sciences Po, 2019.*

En 2016, le législateur a introduit de nouvelles obligations de transparence sur les algorithmes utilisés par les acteurs publics. La loi pour une République numérique oblige ainsi les administrations à prévenir les usagers lorsqu'un algorithme s'est immiscé dans le traitement de leur dossier (demande de bourse, calcul d'un impôt ou d'une taxe, attribution de prestations sociales, etc.).

Dans le cadre du projet de loi de transposition du RGPD dont j'étais rapporteure, j'ai fait adopter un amendement visant à rendre intelligible pour les usagers la manière l'algorithme est intervenu dans une **décision administrative** individuelle. Sur le fondement notamment des dispositions de l'article 15 du RGPD, cette logique pourrait, par exemple, être utilement étendue au secteur privé.

Dans la loi contre la manipulation de l'information, le Parlement a adopté, à mon initiative, un **amendement** contraignant les plateformes à révéler les statistiques sur la provenance des contenus. Actuellement, des contenus sont proposés en ligne par les plateformes, soit par un accès direct (grâce à l'URL ou un moteur de recherche extérieur) soit par un accès indirect (c'est-à-dire suggéré) recommandé ou mis en avant de façon algorithmique. Connaître la part de chaque voie d'accès nous permettra de comprendre si un algorithme en particulier est plus ou moins responsable de l'affichage régulier de certains types de contenus qui véhiculent notamment des fausses informations. Ces statistiques doivent être consultables par tous en ligne et être réutilisables.

Armer le régulateur

Il faudra aller plus loin. Comme nous l'indiquions avec Benoît Loutrel et Aymeril Hoang dans une tribune, le législateur doit confier au régulateur le pouvoir d'obtenir la transparence et d'accéder aux données des grands réseaux sociaux, à l'instar de l'Autorité des marchés financiers (AMF), qui conditionne tout appel à l'épargne publique à une stricte transparence de la part de l'émetteur de titres.

Dans le rapport de la mission sur Facebook conduite par Benoît Loutrel, les auteurs plaident pour la création d'un « régulateur de la responsabilisation des principaux réseaux sociaux *via* le contrôle des obligations de transparence des fonctions d'ordonnancement et de modération des contenus, et de devoir de diligence leur incombant ». Ce régulateur, selon les rapporteurs, « ne serait ni le régulateur des réseaux sociaux dans leur globalité, ni le régulateur des contenus qui y sont publiés. Elle ne serait pas compétente pour qualifier les contenus pris individuellement. Elle coopérerait avec les services de l'État placés sous l'autorité du gouvernement et les services judiciaires. » Un régulateur ouvert sur la société civile, associant à ses travaux des communautés techniques, des scientifiques et des journalistes, par exemple, constituerait une vraie innovation.

Pour ma part, j'ai plaidé lors de la discussion sur la loi contre les contenus haineux sur Internet pour un dispositif constituant un premier pas dans ce sens. La proposition était de donner au Conseil supérieur de l'audiovisuel (CSA) la capacité de lancer des appels à projet d'intérêt général portant sur le fonctionnement des opérateurs de plateforme qui justifieraient l'accès temporaire aux données des réseaux sociaux. Pourraient y participer administrations, chercheurs, journalistes ou associations. Un cadre d'accès aux données devrait être créé, en collaboration étroite avec la Commission nationale de l'informatique et des libertés (CNIL) et de l'Autorité de régulation des communications électroniques et des postes (ARCEP). Un tel dispositif existe d'ores et déjà dans le domaine de la statistique publique où le ministre chargé de l'économie dispose de la faculté d'exiger à des personnes morales de droit privé la fourniture des données pertinentes afin de réaliser des enquêtes. En développant cette nouvelle régulation, nous pourrions ainsi ponctuellement lever le verrou du secret des affaires en construisant des accords de non-divulgaration (DNA) entre plateformes et autorités publiques.

Le régulateur doit être le bras armé de la société, tout autant que celui de l'État. Avec Sébastien Soriano, président de l'ARCEP, nous expliquions, en novembre 2019, ce que nous entendions par « régulation par la donnée » : « Nous ne voulons pas d'un régulateur qui serait là pour brider l'innovation ou décider à la place de la société. Le contrôle des Big Tech ne doit pas céder le pas à celui de l'État. Le régulateur doit être l'instrument de la société, pour contrebalancer le pouvoir des

plus forts. Cette philosophie d'action est cruciale. » L'ambition est de donner des capacités d'action aux utilisateurs pour qu'ils deviennent eux-mêmes des régulateurs à part entière.

Agir sur les algorithmes et les fonctionnalités

Les connaissances puisées dans l'étude collective des algorithmes et des données permettront à la société de détailler les spécifications techniques souhaitées ou acceptables. Certains réseaux sociaux seraient peut-être amenés à modifier leurs algorithmes ou leurs fonctionnalités. Quelques pistes d'actions peuvent dès à présent être explorées.

Par exemple, il serait possible d'envisager que les algorithmes de recommandation sur nos fils d'actualité pondèrent de manière moins importante le critère de notoriété du contenu. Ainsi, nous ne verrions pas en priorité les contenus fortement partagés ou commentés, mais plutôt d'autres types de contenus (ceux de nos proches, des contenus « fact-checkés » ou encore des contenus proposés de façon aléatoire).

Une autre possibilité serait d'envisager de multiples filtres permettant de sélectionner le type de contenu mis en avant. L'application Flipfeed propose cette option : elle permet de remplacer notre fil d'actualité Twitter par celui d'une personne fictive n'ayant pas du tout les mêmes orientations politiques, idéologiques que nous. Cette possibilité de reprendre le contrôle sur l'éditorialisation des contenus que nous voyons pourrait permettre de limiter les effets de « bulle de filtre » évoqués précédemment et d'augmenter la sensibilisation des utilisateurs à l'impact des algorithmes sur leur consommation d'information.

Certains réseaux sociaux permettent, de par leur design et leurs fonctionnalités, une moins grande viralité que d'autres. C'est par exemple le cas de Snapchat, qui fonctionne sur le principe des « stories » (vidéos courtes qui disparaissent rapidement). C'est un réseau social fermé où les contenus ne peuvent être vus que par des relations. À l'inverse, Twitter est le réseau social le plus ouvert : le partage (retweet) est à la base du modèle. Récemment, Twitter a changé son interface : des « frictions » ont été ajoutées avant de partager des contenus (comme le fait de proposer la lecture des articles proposés ou de rajouter un commentaire personnel sur les retweets). Un utilisateur peut aussi désormais empêcher la publication de commentaires sur son Tweet. Cela a pour but de limiter la confrontation et la polémique. Twitter a finalement mis en place la possibilité de faire des « stories » qui ne peuvent, elles, être partagées ou commentées.

Enfin, les messageries privées jouent un rôle important dans les « raids » haineux et le cyberharcèlement. Ce fut le cas notamment dans l'affaire Milla. La jeune fille a reçu plus de 50 000

menaces de mort et injures à la suite d'une vidéo postée sur Instagram dans laquelle elle critique la religion musulmane. Il est déjà possible de bloquer un utilisateur et de signaler un contenu partagé dans un message privé. Face à un niveau aussi élevé, il devrait être rendu possible de pouvoir désactiver momentanément la messagerie, tout en conservant l'activité du réseau social et ainsi stopper le déferlement de haine.

Interopérabilité

Portée de longue date par les militants du logiciel libre, l'interopérabilité permet d'envisager un Internet sans monopole des effets réseaux. L'interopérabilité est la capacité d'un système informatique à fonctionner avec d'autres produits ou systèmes d'information existants.

Le RGDP a introduit un droit de portabilité pour rééquilibrer les rapports entre les responsables de traitements et leurs usagers. Il s'agit pour l'utilisateur de pouvoir récupérer l'ensemble de ses données pour les transférer d'un service à un autre. Désormais, c'est l'interopérabilité entre plateformes qui est en jeu, par l'ouverture de leurs API (interfaces de programmation active). Cela permettrait d'éviter l'étape de téléchargement des données et d'aller vers un changement de service le plus fluide possible.

Pour le collectif La Quadrature du Net, « l'interopérabilité garantit à tout le monde de ne pas se trouver captif d'une plateforme : de pouvoir librement la quitter, sans perdre ses liens sociaux, et de continuer à communiquer avec ses contacts. L'interopérabilité permet à quiconque de lire depuis un service A les contenus diffusés par ses contacts sur un service B, et d'y répondre comme si elle y était ». L'exemple le plus connu d'interopérabilité est celui des mails. Demain, nous pourrions faire de même avec notre compte Twitter ou notre compte Facebook. L'interopérabilité pourrait ainsi permettre l'émergence de nouveaux services avec différents modèles d'affaires, différentes politiques de modération, différentes interfaces et rendre ainsi le choix aux utilisateurs.

Cette proposition a aussi été défendue par l'ARCEP dans le cadre de la consultation lancée par la Commission européenne sur le Digital Services Act.

Encadrement du micro-ciblage

Le modèle d'affaire des plateformes soulève un nombre important de critiques, en particulier en ce qui concerne la pratique du micro-ciblage publicitaire. C'est le cœur de la critique de la chercheuse Shoshana Zuboff dans son livre sur le « capitalisme de surveillance ». L'individualisation de la

publicité sur les réseaux sociaux est permise par l'agrégation par les plateformes de données personnelles. La densité et la précision de ces données permettent de construire des modèles prédictifs des comportements consuméristes des individus.

C'est l'application de ces techniques de micro-ciblage publicitaire à des fins politiques qu'a révélé le scandale Cambridge Analytica. En calculant des traits de personnalités à partir des interactions sur la plateforme, des types de contenu lus et partagés, des commentaires et messages rédigés, la société anglaise a permis au camp de Donald Trump, par exemple, de cibler très précisément des publicités politiques de manière individualisée. En d'autres termes, deux personnes pouvaient recevoir deux informations différentes en provenance de sa campagne en fonction de leur profil et de leurs attentes supposées, créant un détournement des données personnelles à des fins politiques et une distorsion de l'information en temps de campagne électorale.

Le micro-ciblage publicitaire appliqué au débat politique va à l'encontre de l'objectif d'intérêt général des réseaux sociaux d'assurer un cadre de délibération neutre, transparent, apaisé. De toute évidence, la pratique du micro-ciblage doit être très fortement encadrée, si ce n'est complètement interdite en ce qui concerne les contenus politiques.

La régulation par la société

Modération par les plateformes et par la justice

Dans son dernier entretien à *L'Obs*, Laetitia Avia a émis le souhait que les plateformes recrutent des « milliers de modérateurs ». Cette nouvelle approche est moins problématique dans le sens où elle rompt avec l'objectif de finalité de retrait rapide des contenus pour se placer sur un objectif de moyens.

La pratique de la modération des contenus a fait l'objet de plusieurs documentaires. Ils ont mis récemment en évidence la condition des « nettoyeurs du Net », ces petites mains payées au clic pour visionner des contenus haineux et violents, parfois traumatisants. C'est grâce à ces modérateurs qu'au plus fort de la guerre en Syrie contre Daesh, nos réseaux sociaux n'ont pas été envahis d'images de torture et d'exaction, par exemple.

Nous n'aborderons pas ici la question – néanmoins tout à fait centrale dans les mois qui viennent – des conditions sociales et de travail des modérateurs du Net. Le législateur devra aussi créer un statut plus protecteur pour « ces petits doigts de l'intelligence artificielle ».

La modération des contenus par les plateformes soulève au moins deux problèmes.

Le premier est d'ordre quantitatif : le nombre de contenus qui circule est trop important pour être traité de manière omnisciente. La prévalence des contenus haineux sur Facebook est identifiée de 0,10% à 0,11%, soit 10 à 11 contenus haineux toutes les 10 000 vues. Ce qui représente environ 22 millions de contenus modérés entre juillet et septembre 2020, uniquement sur Facebook. C'est pourquoi des IA sont désormais utilisées pour identifier des contenus illicites avant le signalement par un utilisateur. Facebook met en évidence que 94,7% des contenus haineux retirés le sont par une IA, avant un signalement. Ce taux n'était que de 23,6% en 2017. Cette tendance, qui vise l'efficacité, peut s'avérer problématique puisqu'elle traite de façon automatisée des décisions ayant trait à la liberté d'expression.

Le second est d'ordre qualitatif : la modération humaine est soumise à des interprétations liées à des biais culturels et sociaux. Les différents documentaires précédemment cités montrent bien comment des modérateurs des réseaux sociaux aux Philippines, par exemple, ont une approche du licite et de l'illicite très divergentes de celles d'un modérateur basé en Europe. Cela peut conduire parfois à de grandes équivoques, comme la censure d'un tableau d'une artiste américaine figurant Donald Trump nu.

Seule, la modération par les plateformes n'est donc ni soutenable ni souhaitable. Il nous faut, d'abord, renforcer la capacité d'action de la justice.

Il s'agit de consolider à la fois le mécanisme de plainte pour la présence de contenus haineux, mais aussi le mécanisme d'appel pour contester une décision de retrait de contenu pour un utilisateur qui se sentirait lésé. Pour lever l'opacité sur les décisions de modération des plateformes, nous devons créer un véritable « Digital Due Process ».

La création d'un parquet national numérique verra le jour avant la fin de l'année. Pour que son action soit efficace, il lui faudra des moyens techniques et humains suffisants, une capacité de coopération et de contrainte envers les réseaux sociaux pour la transmission d'informations. L'institution judiciaire est le parent pauvre des fonctions régaliennes de l'État. Son niveau de numérisation et d'équipement technologique est alarmant. La plateforme de signalement des contenus illicites, Pharos, est à cet égard totalement sous-dotée : 27 policiers ont eu à traiter 213 000 signalements en 2019.

Modération par la communauté

Jusqu'à présent, la modération des contenus a principalement été réalisée par les plateformes et par la justice. Une troisième forme de modération doit être développée : celle par la communauté, permettant une régulation par la société.

En avril 2019, je répondais dans une tribune à Mark Zuckerberg qui invitait les États à jouer un rôle plus actif dans la régulation d'Internet. Je détaillais alors des exemples concrets de la modération communautaire, modération décentralisée à l'image du fonctionnement de la communauté des contributeurs de Wikipédia ou des forums type Discourse ou Discord.

Aux États-Unis, le site skeptics.stackexchange.com est un site participatif sur le scepticisme scientifique, où les internautes peuvent poser des questions et y répondre librement. Chaque utilisateur a une note de « réputation » qui permet de se fier à la qualité de son jugement.

Un autre exemple est celui des « elfes lituaniens ». Il s'agit d'un mouvement de résistance civique contre les trolls russes qui propagent la désinformation en ligne. Plus de 5000 Lituaniens participent à cette initiative, en partenariat avec le ministère de la Défense, en vérifiant la véracité des articles partagés en ligne, répertoriant les liens considérés par la communauté comme faux et en publiant du contenu vérifié.

Taïwan est probablement le modèle le plus abouti. Grâce à l'impulsion de sa charismatique ministre des Affaires numériques Audrey Tang, « l'infodémie » de fausses nouvelles liées à la pandémie de la Covid-19 a été contenue. Le Taiwan [FactCheck Center](#) est un modèle de collaboration entre médias, chercheurs, citoyens et gouvernement pour détecter des informations douteuses et les analyser. La société civile a, par ailleurs, conçu [un bot \(Aunt Meiyu\)](#) pour lutter contre la désinformation diffusée dans des groupes de discussion privés. Environ 65 000 personnes utilisent ce bot qui détecte les textes et liens partagés sur une boucle et les compare avec une base de données construite sur la base de signalements et contributions communautaires afin d'alerter les membres du groupe sur la probabilité que le contenu partagé soit faux.

Cette modération communautaire est la base de la stratégie publique de contre-discours, qui a permis qu'Audrey Tang résume par les trois mots « Fast, Fair, Fun » :

1. « Fast » : la fausse nouvelle est rapidement « fact-checkée » et une réponse est proposée en moins d'une heure ;
2. « Fair » : toutes les informations font l'objet d'une vérification, sans discrimination de la provenance ou du type de contenu ;

3. « Fun » : les producteurs de contenus utilisent des « meme » (images humoristiques) et répondent selon le standard du « 2-2-2 » : pas plus de 2 images en réponse ; pas plus de 20 signes pour le titre ; pas plus de 200 caractères dans le corps du message. Cela permet de créer des contenus viraux. Pour Audrey Tang, seul l'humour est plus viral que la haine et que les fausses nouvelles sur les réseaux sociaux.

C'est ainsi que Taïwan a pu contenir « l'infodémie » de fausses nouvelles liées à la pandémie de la Covid-19 sans procéder à aucun « takedown » de contenu.

Ces différents modèles de régulation communautaires peuvent permettre de répondre à la crise de défiance très importante envers les médias traditionnels ou encore la parole des politiques. Ils sont aussi diamétralement opposés à la logique que souhaitait mettre en place le gouvernement au début de la pandémie avec le site gouvernemental Désinfox Coronavirus, centralisant les articles de « fact-checking » de certains médias. Ce projet a été finalement abandonné. La qualité du débat public est une responsabilité de tous. Nous pouvons tous y contribuer, nous engager, nous mobiliser. L'approche de la modération communautaire pourrait aussi permettre de passer d'une vision fondée sur le retrait des contenus à une vision fondée sur la mise en avant de contenus de bonne qualité.

Pour un droit de réponse numérique : renforcer le dialogue démocratique

Le législateur a souvent tendance à vouloir créer un nouveau droit du numérique, comme si le droit existant ne pouvait s'appliquer sur Internet. Notre loi sur la liberté de la presse encadre l'expression publique et les conditions d'un dialogue démocratique apaisé : nous pouvons transposer certains de ces principes aux débats qui ont lieu sur les réseaux sociaux.

L'article 13 et 13-1 de la loi sur la liberté de la presse encadre le droit de réponse. Toute personne nommée ou désignée dans un journal peut demander à celui-ci un droit de réponse pour défendre une opinion différente. Le directeur de la publication est tenu par la loi de lui accorder dans un délai imparti, sous peine de contravention. Nous pourrions envisager un « droit de réponse numérique ». Un utilisateur postant sur son compte un message nommant ou désignant une personne devrait accorder un droit de réponse si celle-ci le lui demande. À la différence d'une réponse en commentaire du message (qui n'atteint jamais le même niveau de visibilité) ou d'une reprise du message par le compte de la personne désignée (qui ne renforcera pas le débat mais la confrontation), le droit de réponse contraindrait l'auteur du message à le publier à son audience.

Cela pourrait contribuer à nouveau à briser les effets de « bulle de filtre » en ayant une capacité de s'adresser et d'argumenter auprès d'autres communautés. La fonctionnalité « demande de droit de réponse » pourrait être prévue par les réseaux sociaux et standardisée.

Pour un droit d'affichage public numérique : neutralité politique et rappel à la loi

Autre disposition de la loi de 1881 qu'il conviendrait d'appliquer sur les réseaux sociaux : l'affichage public, prévu aux articles 15 et 17. Chaque mairie désigne des emplacements réservés à la communication des lois et des actes émanant de l'autorité publique. Ainsi, il pourrait être envisagé que les réseaux sociaux sanctuarisent des espaces publicitaires pour la communication, par les autorités, de rappels à la loi (notamment en ce qui concerne les risques encourus pour la diffusion de certains contenus). En période électorale, l'affichage public en ligne pourrait correspondre à des panneaux électoraux : chaque candidat pourrait ainsi avoir accès à un traitement équitable sur la place publique numérique.

Conclusion

Il n'existe pas de solution magique pour lutter contre la multiplication des contenus problématiques circulant sur les réseaux sociaux. La transparence des algorithmes, le développement de nouveaux paramétrages et fonctionnalités, l'interopérabilité, l'interdiction du micro-ciblage, la modération par la communauté, le droit de réponse 2.0 peuvent cependant constituer des solutions concrètes pour enrayer les dynamiques de viralité. Cette approche ne cherche pas à choisir entre les plateformes et les États. Elle cherche, au contraire, à faire émerger un troisième acteur : les utilisateurs. Ces derniers doivent être protégés dans leur droit à une expression libre de leurs opinions, éclairés dans la façon dont les algorithmes façonnent leur consommation d'information, responsabilisés dans leurs actes de diffusion de certains discours, épaulés par le régulateur dans leur rapport de force avec les plateformes.

Dans une tribune publiée dans la revue *Foreign Affairs* en novembre 2020, le philosophe Francis Fukuyama s'inquiète du rôle croissant occupé par les réseaux sociaux dans la vie démocratique américaine :

« Le pouvoir économique et politique concentré des plateformes numériques est comme une arme chargée posée sur une table. [...] La question pour la démocratie américaine, cependant, est de savoir si elle est sûre de vouloir laisser l'arme ainsi posée, où une autre personne avec de pires intentions pourrait venir la prendre. Aucune démocratie libérale ne peut se contenter de confier un

tel pouvoir politique dans les mains des individus sur la base d'hypothétiques bonnes intentions. »

Seuls des contre-pouvoirs issus de la société civile suffisamment puissants et autonomes seront capables de maîtriser l'influence grandissante des réseaux sociaux dans nos vies démocratiques.