La désinformation à l'ère du numérique : un code d'autodiscipline européen

Par **Paolo CESARINI**
Commission européenne

Les vues exprimées n'engagent que l'auteur.

Si l'expression *fake news* a fait irruption dans le langage courant lors de la campagne présidentielle américaine et le référendum sur le Brexit de 2016, d'autres opérations récentes de désinformation en Europe et ailleurs ont contribué à son inquiétante popularité [(1)]. L'attaque hybride qui cibla la phase conclusive de la campagne présidentielle française de 2017 en est un exemple. Dans un contexte différent, l'attentat perpétré à Salisbury en mars 2018 contre l'ex-agent des services de renseignement russes Sergei Skripal illustre les conséquences, au plan diplomatique et géopolitique, d'une diffusion orchestrée de fausses nouvelles, notamment par les réseaux sociaux.

Face à ces menaces, un appel pressant pour définir une réponse adéquate est venu tant du Parlement que du Conseil européens, ainsi que d'organisations représentatives de la société civile et des médias. Un récent sondage d'Eurobaromètre indique que les fausses nouvelles sont perçues aujourd'hui comme une menace pour la démocratie par 83 % des Européens et précise qu'il s'agit d'un phénomène très diffus dans tous les États membres de l'Union [(2)]. Comme l'a souligné la Commission européenne dans sa Communication d'avril 2018, *Lutter contre la désinformation en ligne* [(3)], la désinformation mine la confiance des citoyens dans les institutions publiques et dans les médias, numériques ou traditionnels, car elle s'attaque aux valeurs démocratiques. En limitant la capacité des individus de se former une opinion et de prendre des décisions en pleine connaissance de cause, elle entrave la libre expression. La désinformation diffusée à grande échelle est aussi susceptible de fausser le débat public sur des thèmes structurants comme l'immigration, les changements climatiques ou la santé, et peut fragiliser la sécurité interne ou nuire à l'intégrité des élections.

Opérateurs des services qui permettent aux fausses nouvelles d'atteindre une rapidité de diffusion, une pénétration et une précision de ciblage sans précédent, les grandes plateformes numériques, notamment Facebook, Google et Twitter, doivent désormais faire face à leur responsabilité sociale. Le rapport présenté à la Commission en mars 2018 par un groupe d'experts de haut niveau [(4)] avait mis l'accent sur la complexité du phénomène et sur la nécessité d'une approche multidimensionnelle, menée sur plusieurs fronts et impliquant toutes les parties concernées, publiques et privées. La Communication d'avril 2018 a largement suivi les recommandations du groupe d'experts en fixant une pluralité d'actions complémentaires et interdépendantes. Celles-ci visent notamment à responsabiliser les plateformes, à soutenir l'émergence d'un réseau indépendant de vérificateurs de faits à l'échelle européenne, une meilleure éducation aux médias et un journalisme de qualité, tout en renforçant la résilience des processus électoraux contre les menaces informatiques.

(1) "Freedom on the net 2017 report", Freedom house, https://freedomhouse.org/report/freedom-net/freedom-net-2017
(2) https://ec.europa.eu/digital-single-market/en/news/first-findings-eurobarometer-fake-news-and-online-disinformation
(3) https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:52018DC0236&from=EN
(4) *A Multi-dimensional Approach to Disinformation*, mars 2018, https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

L'approche retenue pivote sur un principe d'autorégulation. Le *Code de bonnes pratiques contre la désinformation* [5], élaboré en septembre 2018 par un Forum multipartite réunissant les sociétés technologiques, l'industrie publicitaire, les médias et les organisations de la société civile, représente une application concrète de ce principe. Premier exemple au niveau mondial d'autorégulation dans ce domaine, ce Code a été souscrit en octobre 2018 par Twitter, Facebook, Google et Mozilla et par plusieurs associations européennes représentatives de l'industrie publicitaire. Il est ouvert à tout autre opérateur concerné.

Vulnérabilités systémiques

Pour apprécier la portée et l'impact potentiel de ce Code, il faut considérer d'abord certaines vulnérabilités propres à l'actuel écosystème des médias. Plusieurs études récentes suggèrent que la transformation numérique de l'industrie des médias et la montée en puissance des plateformes en ligne sont à l'origine de cinq types principaux de failles systémiques pouvant être exploitées par des acteurs hostiles, dans un but lucratif ou en tant qu'instrument de subversion politique [6].

Micro-ciblage et personnalisation des publicités politiques ou engagées

Un profilage psychométrique précis des utilisateurs est rendu possible par la production croissante, la collecte et l'analyse de grandes quantités de données personnelles. Combinée avec l'application de techniques avancées d'analyse prédictive et de systèmes d'intelligence artificielle, cette abondance de données permet de personnaliser les publicités politiques ou engagées, afin d'en micro-cibler la distribution et d'en accroître l'impact sur des vastes audiences. L'utilisateur n'est pas nécessairement conscient de l'usage effectif de ses données personnelles et peut ainsi être induit en erreur quant à la nature ou la signification des informations reçues. Le scandale Cambridge Analytica/Facebook qui a éclaté en mars 2018 constitue la démonstration exemplaire d'un tel manque de transparence dans le système.

Astroturfing

Il s'agit d'un ensemble de techniques malicieuses, manuelles ou algorithmiques, permettant de simuler l'activité d'une foule dans un réseau social. Des systèmes automatisés (*bots*) et des bataillons de mercenaires (*trolls*) peuvent être engagés dans des opérations coordonnées visant à déformer le paysage socio-médiatique en diffusant des spams convaincants, en ouvrant des comptes fictifs sur des réseaux sociaux (*fake accounts*) pour accréditer et normaliser des contenus extrêmes ou carrément faux. Le but est de faire croire que ces messages polarisants proviennent de différentes sources et bénéficient d'un large soutien. En raison des biais cognitifs propres à chaque individu, les utilisateurs sont portés à y adhérer et à les partager sur les réseaux sociaux. Au final, cela peut favoriser la montée d'idées populistes ou la diffusion de fausses conceptions autour de sujets sensibles comme les effets des vaccins.

Clickbaiting

Ce type de vulnérabilité résulte du modèle actuel de la publicité numérique. Il repose sur des réseaux publicitaires garantissant le placement en temps réel d'annonces qui, grâce à des processus

(5) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

(6) CHRISTIE E. H. (2018), *Political Subversion in the age of Social media*, https://martenscentre.eu/publications/political-subversion-age-social-media, octobre. KAVANAGH J. and RICH M. (2018), *Truth Decay*, Rand Corporation. JEANGÈNE VILMER J.-B. et al. (2018), *Les Manipulations de l'information : un défi pour nos démocraties*, Centre d'Analyse, de Prévision et de Stratégie et Institut de Recherche stratégique de l'École militaire, Pars, août. MATZ S. et al. (2017), *Psychological Targeting as an Effective Approach to Digital Mass Persuasion*, Proceedings of the National Academy of Sciences 114/48 (November), 12714-12719. DEL VICARIO M. et al. (2016), *Echo Chambers: Emotional Contagion and Group Polarization on Facebook*, Nature, Scientific Reports 6:37825. WARDLE C. and DERAKHSHAN H. (2017), *Information disorder*, Council of Europe report, DGI 09.

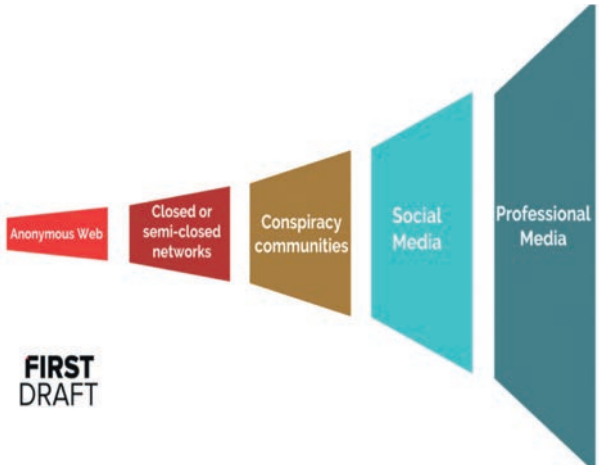
décisionnels automatisés, monétisent les sites-hôtes sur base du nombre de clics effectués. Ce modèle facilite le placement d'annonces sur des sites web qui capturent l'attention par des contenus sensationnalistes, jouant sur les émotions du public, y compris la désinformation. À titre d'exemple, ce modèle a été exploité dans un but lucratif par un groupe d'adolescents de Veles, en Macédoine, très actif pendant la campagne présidentielle américaine de 2016.

Distorsions algorithmiques

Les plateformes numériques agrègent et filtrent les contenus sur la base d'algorithmes qui exploitent les traces numériques des internautes pour identifier de façon granulaire leurs préférences individuelles. Du point de vue des plateformes, le but est de fournir du contenu pertinent et engageant afin de retenir l'attention des utilisateurs et prolonger le temps qu'ils passent sur leurs sites. Cela augmente le nombre potentiel d'impressions publicitaires, ce qui se traduit par un accroissement des revenus. Le fonctionnement de tels algorithmes reste cependant largement obscur. Des recherches récentes ont constaté que, dans certains cas, leur fonctionnement serait biaisé par une forme d'auto-radicalisation, consistant à privilégier la visualisation, par chaque utilisateur, de contenus de plus en plus extrêmes, tendancieux ou faux [7].

Dynamiques de diffusion

L'attribution des opérations de désinformation est complexe. Il s'agit rarement de l'action d'individus isolés, mais plus souvent du résultat d'une coordination entre plusieurs acteurs agissant à différents niveaux de l'écosystème numérique. Claire Wardle a récemment publié sur le site de *First Draft* [8] une représentation visuelle de l'enchaînement de ces mécanismes d'amplification (voir image ci-contre). Une recherche empirique conduite par le MIT montre que les fausses nouvelles se propagent plus vite et plus largement sur les réseaux sociaux que les informations véridiques [9]. Les humains seraient donc plus enclins à partager du contenu faux sur les réseaux sociaux. Dans ce contexte, les utilisateurs, mais aussi les journalistes professionnels et les éditeurs de presse écrite ou audiovisuelle, peuvent participer de manière involontaire ou inconsciente au processus d'amplification.



La responsabilité des plateformes

Ces vulnérabilités jouant un rôle déterminant dans la dissémination de fausses nouvelles : l'absence de règles encadrant la responsabilité des médias sociaux dans ce domaine est souvent citée parmi les causes principales du problème.

(7) RIEDER B, MATAMOROS-FERNANDEZ A. and COROMINA O. (2018), *From Ranking Algorithms to "Ranking Cultures": Investigating the Modulation of Visibility in YouTube Search Results*, Convergence: The International Journal of Research into New Media Technologies 24/1, pp. 50-68.
(8) https://firstdraftnews.org/5-lessons-for-reporting-in-an-age-of-disinformation/
(9) VOSOUGHI S., ROY D. and ARAL S. (2018), "*The Spread of True and False News Online*", Science 459, pp. 1146-1151.

La Directive sur le commerce électronique [10] établit un régime de responsabilité limitée pour les plateformes numériques qui se bornent à héberger du contenu de tiers sur leurs services. En vertu de ce régime, les plateformes sont tenues de supprimer *uniquement* les contenus illégaux, de façon rapide dès qu'elles en ont pris connaissance, en adoptant les mesures utiles pour en limiter la résurgence en ligne. La récente adoption de la Directive sur les services audiovisuels [11] a fait évoluer ce cadre législatif en obligeant les médias sociaux à adopter proactivement des bonnes pratiques, tant pour supprimer les contenus illicites tels que les discours haineux, violents ou incitant au terrorisme que pour limiter l'accès à certains contenus préjudiciables (notamment pour les enfants). Dans ce domaine, la directive intervient en rendant contraignants les objectifs à atteindre, et ceci à l'échelon européen, afin d'optimiser les efforts entrepris ou à entreprendre par les entreprises concernées et par les États membres pour adresser efficacement ce type de problèmes dépassant les frontières nationales.

Les fausses nouvelles, qui peuvent être préjudiciables sans être *per se* illicites, ne rentrent cependant dans aucun encadrement réglementaire spécifique au niveau européen. En effet, la nature protéiforme du phénomène rend un tel encadrement particulièrement ardu. Les fausses nouvelles peuvent être un instrument de lutte politique interne ou un moyen privilégié pour orchestrer des campagnes de haine, pour intimider ou diffamer des figures publiques ou des groupes sociaux. Elles peuvent provenir d'acteurs internes ou externes à un pays qui, à leur tour, peuvent être de nature étatique ou non étatique. Mais elles peuvent aussi apparaître dans un contexte de satire, de parodie, de manifeste critique sociale, ou être le résultat de simples erreurs journalistiques.

À cause de cette diversité, l'imposition d'une pure et simple obligation de suppression des fausses nouvelles comporte un risque évident d'interférence avec les droits fondamentaux qui garantissent la liberté d'expression. En particulier, l'article 11 de la Charte des droits fondamentaux de l'UE établit le droit de tout citoyen « de recevoir ou de communiquer des informations ou des idées sans qu'il puisse y avoir ingérence des autorités publiques et sans considération de frontières ». Ce principe protège notamment les contenus satiriques, parodiques et de critique sociale légitime, indépendamment de leur nature extrême ou choquante.

En décembre 2018, le Plan d'action contre la désinformation de la Commission et du Service pour l'action externe de l'UE [12] a confirmé une définition de la désinformation qui se fonde sur deux composantes essentielles : la nature du *contenu diffusé* et l'*intentionnalité trompeuse* qui caractérise la création et la dissémination de fausses nouvelles. Une approche réglementaire basée uniquement sur une simple catégorisation de fausses nouvelles préjudiciables aurait risqué d'entraîner le cauchemar orwellien d'un arbitre de la vérité, publique ou privée.

C'est pourquoi l'élément de l'intentionnalité intervient pour éliminer ce risque en caractérisant la désinformation en fonction de *comportements* préjudiciables pour l'écosystème de l'information, et en raison des vecteurs, des outils technologiques et des méthodes utilisés. Par exemple, des opérations de désinformation visant l'intégrité des processus électoraux par l'application de méthodes de communication abusives entrent dans cette catégorie. La loi Macron, qui criminalise les fausses nouvelles dans des conditions précises (distribution massive et artificielle) et dans un contexte électoral spécifique, semble suivre une telle logique.

Un code d'autodiscipline pour l'industrie

Dans ce contexte, les médias sociaux et d'autres acteurs responsables des effets « vulnérants » décrits plus haut doivent reconnaître la nécessité de revoir leurs pratiques de manière à ne plus

(10) https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:32000L0031
(11) https://www.droit-technologie.org/wp-content/uploads/2019/01/CELEX_32018L1808_FR_TXT.pdf
(12) https://ec.europa.eu/commission/sites/beta-political/files/eu-communication-disinformation-euco-05122018_en.pdf

contribuer à l'amplification de la désinformation et à mieux gérer l'impact de leurs technologies sur les vulnérabilités du système de la communication. Par exemple, le trucage de vidéo (*deep fakes*), la création de faux comptes ou l'utilisation de systèmes automatisés (*bots*) capables d'amplifier l'emprise sur le public de contenus faux sont des comportements qui dénotent une manifeste intentionnalité trompeuse et qui, de ce fait, devraient impliquer une action corrective efficace par les plateformes et une sanction efficace et proportionnée en cas d'inaction.

Le Code de bonnes pratiques contre la désinformation de septembre 2018 spécifie les cinq types de vulnérabilité principalement responsables de la diffusion virale de la désinformation en ligne. Il engage les plateformes à y remédier par des mesures pertinentes et ajustables à une échelle européenne. Il prévoit notamment :

- des mesures de transparence pour les annonces politiques ou engagées visant à permettre l'identification des sponsors, des montants payés et la compréhension des critères de ciblage. Cela inclut aussi la création d'archives numériques permanentes, accessibles à des fins de recherche ;
- des engagements pour mieux assurer la résilience des services en cas d'attaques au moyen de systèmes automatisés, faux comptes et autres techniques d'*astroturfing* favorisant la radicalisation et la polarisation des communications ;
- le développement d'outils de placement d'annonces en ligne qui protègent les marques des annonceurs en démonétisant les sites web qui se financent par la prééminence qu'ils donnent aux fausses nouvelles ;
- une série de mesures utiles pour assurer un filtrage algorithmique reflétant la fiabilité des sources d'information plus que la popularité du contenu, et pour lier et distribuer de façon automatisée une pluralité de contenus reflétant des points de vue divers autour de sujets clivants ;
- et enfin, pour mieux gérer la problématique liée aux dynamiques de diffusion, le Code demande aux plateformes de s'engager dans une collaboration effective avec les organisations de recherche et avec les vérificateurs de faits pour consentir un monitorage indépendant, constant et à l'échelle européenne, de ces dynamiques. Cela comporte aussi l'accessibilité des données nécessaires à mieux saisir l'émergence et les risques de diffusion virale des campagnes de désinformation, dans le respect naturellement du règlement pour la protection des données personnelles (RGPD).

L'efficacité du Code dépendra surtout de la capacité des signataires d'en poursuivre les objectifs de manière rigoureuse et constante. Leurs efforts devront être à la hauteur de la taille de chaque signataire et de la responsabilité qu'il porte. L'implémentation de ces bonnes pratiques devra enfin être évaluée régulièrement par la Commission en coopération avec les autorités des États membres et sur la base d'indicateurs de performance transparents. Un monitorage serré est en cours.

Même si nécessaire, le Code n'est cependant pas en soi suffisant pour neutraliser la désinformation dans l'environnement actuel. D'une part, l'efficacité des bonnes pratiques prévues par le Code, comme celles visant à mieux comprendre et contrôler les dynamiques de diffusion ou à mieux protéger les utilisateurs, dépend d'autres actions menées en parallèle par la Commission et les États membres ; on peut citer notamment celles destinés à soutenir l'émergence d'un réseau européen indépendant de vérificateurs de faits et de chercheurs académiques, ainsi que celles visant à favoriser l'éducation du public aux médias et à soutenir un journalisme professionnel.

D'autre part, des facteurs non liés aux technologies, comme l'ingérence directe d'États tiers, ou le taux de polarisation politique induite par les inégalités économiques, ou l'action d'autres forces qui poussent vers la radicalisation et l'extrémisme dans la société, jouent un rôle tout aussi critique. En conséquence, le Plan d'action de décembre 2018 prévoit aussi d'autres interventions, en particulier la création de mécanismes spécifiques de coordination entre autorités nationales et avec les institutions de l'UE et la création d'un système d'alerte rapide, dans le but de faciliter les échanges d'information et une meilleure analyse des menaces par les autorités compétentes. Le Code représente l'un de ses piliers.