ACADÉMIE DES TECHNOLOGIES
POUR UN PROGRÈS RAISONNÉ CHOISI PARTAGÉ

DÉRIVES DE LA COMMUNICATION EN LIGNE : CONSTATS ET REMÈDES

Nicolas Curien
Membre de l'Académie des technologies

Séance du 9 février 2022

Résumé

La responsabilité des dérives de la communication en ligne est souvent imputée à la négligence des grandes plateformes, mais ces dérives procèdent surtout, selon les psychologues évolutionnistes, de biais cognitifs peut-être hérités des lointaines époques où nos ancêtres étaient confrontés à des univers extrêmement hostiles. Cette « pensée paresseuse », qui est vraisemblablement l'une des variables expliquant la diffusion et le succès des fausses informations, ne reculera que grâce à un effort majeur pour développer l'esprit critique de nos concitoyens.

Afin d'endiguer la montée en puissance de l'infox, du complotisme, des incitations à la haine et de la discrimination, les entreprises gestionnaires de réseaux sociaux mettent d'elles-mêmes en œuvre des procédures de détection, de signalement et de traitement des contenus douteux. Les dispositifs automatisés, particulièrement efficaces pour lutter contre les appels au terrorisme ou la pédopornographie, doivent être complétés par une modération humaine pour les contenus plus difficiles à interpréter, comme les discours haineux ou le harcèlement. L'action des plateformes contre les dérives de la communication ne saurait remplacer celle de la justice mais, devant les volumes colossaux d'informations à traiter, elles assurent un premier filtrage et, par ailleurs, coopèrent avec les autorités de police et de justice pour les demandes d'identification.

Dans de nombreux pays, notamment européens, un encadrement législatif et une régulation des pratiques des plateformes numériques sont mis en œuvre. En France, plusieurs lois récentes imposent aux plus grands des acteurs en ligne des obligations de moyens en matière de lutte contre la propagation de contenus socialement indésirables. L'ex CSA, devenu l'Arcom au 1er janvier 2022, a acquis de nouvelles prérogatives en termes de collecte d'informations, de recommandations et de sanctions.

Intervenants

Gérald Bronner
Sociologue, professeur à l'Université de Paris, membre de l'Académie des technologies

Anton'Maria Battesti
Responsable des affaires publiques France, Facebook

Lucile Petit
Directrice des plateformes en ligne à l'ARCOM

Sommaire

Biais cognitifs : d'où proviennent-ils, comment en sortir ?

Gérald Bronner

Gérald Bronner est sociologue, professeur à l'Université de Paris et membre de l'Académie des technologies.

Les biais cognitifs sont une forme d'erreur systématique et rémanente dans nos raisonnements, au sens où, face au même type de problème, les individus, statistiquement, reproduiront les mêmes erreurs.

Un exemple de biais cognitif

Let's make a deal est un jeu télévisé qui a été diffusé de 1963 à 1977 aux États-Unis, avec un grand succès. Le candidat était placé devant trois portes fermées derrière lesquelles se cachaient deux chèvres et une voiture. Pour gagner, il devait désigner la porte correspondant à la voiture. L'animateur du jeu, Monty Hall, savait où étaient les chèvres et la voiture. Après l'annonce du choix du candidat, il devait ouvrir une porte derrière laquelle se trouvait une chèvre, parmi les deux portes non choisies par le candidat... Le candidat était alors invité à ouvrir une deuxième porte, soit la première qu'il avait désignée, soit l'autre porte non ouverte par le présentateur.

En 1990, un lecteur posa une question sur ce jeu à Marilyn vos Savant, une écrivaine réputée posséder le quotient intellectuel le plus élevé jamais mesuré : « *Une fois que l'animateur a ouvert l'une des portes, le candidat a-t-il intérêt ou non à changer de choix ?* » La plupart des gens estiment que, à ce moment du jeu, les chances pour le candidat de tomber sur la chèvre ou sur la voiture sont équivalentes (1/2), mais Marilyn vos Savant répondit que les chances de trouver la voiture étaient de 1/3 si le candidat ouvrait la porte qu'il avait choisie au départ, et de 2/3 s'il modifiait son choix. Cette assertion, qui provoqua une énorme controverse, était exacte. En effet, la probabilité que la voiture se trouve derrière l'une des deux portes non choisies initialement par le candidat vaut 2/3. Elle vaut clairement toujours 2/3 après que l'animateur a ouvert une porte dévoilant une chèvre, si bien que ces deux chances sur trois d'obtenir la voiture sont désormais entièrement concentrées sur la troisième porte, celle que n'a pas choisi le candidat et que l'animateur n'a pas ouverte. Cette vérité peu intuitive, communément appelée le « paradoxe de Monty Hall », a fait l'objet de vérifications statistiques.

Deux types de processus inférentiels

À la fin des années 1970 et au long des années 1980, plusieurs auteurs, dont les plus connus sont Amos Tversky et Daniel Kahneman, ont recensé 150 sortes de biais cognitifs. Pourquoi notre cerveau nous induit-il ainsi en erreur ? Dans son ouvrage *Thinking, Fast and Slow,* publié en 2011, et traduit en français sous le titre *Système 1 / Système 2 : Les deux vitesses de la pensée,* Daniel Kahneman décrit deux grands types de processus inférentiels dans notre cerveau.

Le système analytique, très énergivore, mobilise surtout le cortex préfrontal et produit une pensée critique et méthodique que l'histoire des sciences a progressivement ciselée, à travers des protocoles mettant à distance les illusions mentales. Aucun scientifique n'est à l'abri d'un biais cognitif, mais le fait qu'il travaille de façon collective et en s'appuyant sur des méthodes éprouvées limite le risque de le voir s'abandonner à des intuitions trompeuses.

À l'inverse, le système intuitif, beaucoup moins énergivore que le système analytique, nous permet, en court-circuitant ce dernier, de disposer de réponses plus rapides et généralement suffisantes pour gérer notre vie quotidienne. Les technologies vont devoir, elles aussi dans leurs développements, prendre en compte la nécessité de préserver la diversité biologique, génétique et spécifique. Selon l'usage qui sera fait de l'intelligence artificielle, par exemple, celle-ci peut conduire à une homogénéisation des systèmes agricoles et des paysages ou, au contraire, à leur diversification.

Les biais cognitifs, un avantage sélectif ?

Selon certains psychologues évolutionnistes, ce dispositif de résolution rapide des problèmes qui, parfois, nous induit en erreur, pourrait être un héritage de notre lointain passé. Les biais cognitifs auraient constitué un avantage sélectif dans l'environnement très hostile auquel étaient confrontés nos ancêtres.

On a constaté, par exemple, que nous avons tendance à sous-estimer la distance entre notre corps et un objet mobile s'approchant de celui-ci. Ce biais cognitif assez étrange peut s'avérer utile lorsque l'objet mobile en

question est un prédateur. Si, lorsqu'il est à 100 mètres de vous, vous avez l'impression qu'il est à 80 mètres, vous détalerez plus rapidement.

Il en va de même pour un autre biais cognitif, la pondération des probabilités selon une fonction en S inversé, qui nous fait surestimer les petites probabilités et sous-estimer les grandes. Si vous vous trouvez dans une forêt et que vous entendez un bruissement dans un buisson, celui-ci, statistiquement, a une plus grande probabilité d'être provoqué par le vent plutôt que par un serpent, un prédateur ou un des membres du clan ennemi. Il est cependant de votre intérêt de surestimer la probabilité que ce bruissement corresponde à un danger : nous sommes vraisemblablement les descendants des peureux plutôt que des téméraires…

L'accumulation du sucre sous forme de graisse permettait à nos ancêtres de traverser plus facilement la saison froide mais, dans un monde qui produit du sucre en quantités industrielles, elle engendre une société d'obèses. De même, ces biais cognitifs, qui ont pu présenter un intérêt dans un environnement naturel hostile, perturbent notre activité dans l'environnement informationnel qui est le nôtre aujourd'hui.

L'implémentation de biais cognitifs via l'observation de régularités statistiques

Il existe une autre interprétation à l'existence de ces biais cognitifs. Certains d'entre eux pourraient résulter de l'observation de régularités statistiques depuis notre plus jeune âge. Stanislas Dehaene, professeur au Collège de France, a montré, à partir d'expérimentations menées auprès de bébés, que le développement des réseaux neuronaux est fondé sur une logique inductive consistant à implémenter des inférences générales à partir de cas particuliers.

C'est ainsi que l'on peut expliquer l'un des biais cognitifs les plus répandus, la confusion entre corrélation et causalité. Celle-ci procède vraisemblablement du fait que, depuis notre petite enfance, la plupart des corrélations que nous observons autour de nous correspondent à des liens effectifs de causalité.

Biais cognitifs et diffusion des fausses informations

Quelle que soit leur origine, les biais cognitifs constituent probablement l'une des variables expliquant la diffusion et le succès des fausses informations.

Sur le marché dérégulé de l'information que nous connaissons, il peut paraître intéressant de mettre en avant des énoncés inquiétants qui constituent des « hameçons attentionnels ». Selon un célèbre article de la revue *Science*, « How lies spread » (2018), les fausses informations sont six fois plus virales sur Twitter que les vraies. Contre-intuitives du point de vue narratif, elles provoquent un effet de surprise qui est considéré par notre cerveau comme une plus-value informationnelle *(« Ah ? J'apprends quelque chose ! »)* et, d'autre part, elles sont intuitives du point de vue cognitif, c'est-à-dire qu'elles vont dans le sens de nos attentes ou, précisément, de nos biais cognitifs.

Selon l'article « Judging truth », paru dans l'*Annual review of psychology* (2020), l'une des principales variables expliquant la crédulité sur les réseaux sociaux est le *lazy thinking*, ou *pensée paresseuse*. Lorsque celle-ci s'hybride avec des intérêts politiques (on est davantage porté à croire telle ou telle chose selon que l'on est de droite ou de gauche), elle crée une formidable chambre d'écho à toute forme de crédulité.

Comment en sortir ?

Formuler les problèmes différemment

Une des solutions pour s'affranchir des biais cognitifs consiste à formuler les énoncés différemment. L'un des biais cognitifs les plus connus est la négligence des « taux de base » ou probabilités *a priori*. Soit l'énoncé suivant : « *Une maladie, qui touche une personne sur mille, peut être détectée par un test. Celui-ci a un taux d'erreurs positives de 5 % c'est-à-dire qu'il y a 5 % de faux positifs. Un individu est soumis au test. Le résultat est positif. Quelle est la probabilité pour qu'il soit effectivement atteint ?* » La bonne réponse est « *2 %* », car le taux de base est 1/1000, qu'il convient de multiplier par 20 (100% divisé par 5%) en raison de la positivité du test ; mais la majorité des élèves et des médecins auxquels cette question a été posée dans les années 1970 ont répondu « *95 %* ».

Le psychologue Gerd Gigerenzer a montré qu'il en va tout autrement lorsque le problème est formulé à l'aide de « fréquences naturelles », c'est-à-dire exprimées avec des nombres entiers, par exemple sous la forme suivante : « *Parmi 1 000 Américains, on en trouve un, en moyenne, qui est atteint de la maladie X. Pour chaque millier d'Américains en bonne santé, on trouve 50 personnes, en moyenne, qui sont positives au test. Imaginez que nous prenions 1 000 Américains au hasard, combien, parmi ceux qui ont été positifs au test, ont-ils réellement contracté la maladie ?* » Dans ce cas, 76 % des sujets donnent la bonne réponse.

Développer l'esprit critique

Plusieurs études montrent que tout ce qui permet de stimuler la pensée analytique réduit mécaniquement la probabilité que le sujet adopte des théories du complot. Le rapport *Les lumières à l'ère numérique*, que j'ai rendu à la Présidence de la République en janvier 2022,

comprend un chapitre sur le développement de l'esprit critique et de l'éducation aux médias et à l'information. L'une des trente recommandations du rapport consiste à faire du développement de l'esprit critique une grande cause nationale, non seulement en ce qui concerne les enfants et les jeunes, dans le cadre de l'Éducation nationale, mais, plus largement, en s'adressant à l'ensemble de nos concitoyens, notamment à travers la formation continue.

Infox et discours haineux sur les réseaux sociaux, comment lutter ?

Anton'Maria Battesti

Anton'Maria Battesti est responsable des affaires publiques France chez Facebook.

Les applications du groupe Meta (Facebook, Instagram, WhatsApp) sont utilisées par 3,5 milliards de personnes dans le monde, et 87 % d'entre elles vivent en dehors des États-Unis et du Canada. En France, environ 40 millions de nos concitoyens possèdent un compte Facebook. Cela représente un volume de contenus à réguler vraiment colossal. Au niveau mondial, environ 40 000 collaborateurs travaillent à cette régulation, avec des investissements qui, depuis 2016, représentent 16 milliards de dollars, dont 5 milliards pour la seule année 2021.

Ces collaborateurs se répartissent entre différentes équipes : *Content Policy* définit les règles, *Community Integrity* développe la technologie permettant de faire appliquer les règles (automatismes, détection d'images, détection de mots…) ; *Community Operation* rassemble des personnes de toutes nationalités et issues de professions très diverses (police, justice, ONG en charge des droits de l'homme, grandes entreprises…) qui assurent la modération des contenus en complément des interventions automatisées. Ces personnes doivent avoir un profil leur permettant de comprendre des enjeux extrêmement sensibles. Modérer un contenu, c'est-à-dire le supprimer, n'est jamais un acte anodin, dans la mesure où cela revient à empêcher quelqu'un d'exprimer quelque chose.

Les règles

Nous avons constitué un vaste ensemble de règles très détaillées, qui sont publiques. Elles couvrent le harcèlement, l'usurpation d'identité, les propos incitant à la haine, la traite d'êtres humains, la fraude, le spam, etc. Ces règles ont été élaborées et renforcées au cours du temps afin d'assurer la sécurité des utilisateurs, exigence qui marque la limite de la liberté individuelle.

Le processus d'élaboration de ces règles passe par la consultation, en amont, de différents experts, en interne mais aussi, de plus en plus, en externe : des juristes, mais aussi des experts en matière culturelle, politique ou sociale. Ce processus, qui est dynamique, nous amène parfois à modifier une règle, ou à définir des exceptions à la règle. Il y a quelques années, par exemple, nous avons rencontré un problème avec la célèbre photo de « la petite fille au napalm » prise par Nick Ut lors de la guerre du Vietnam. Nos règles sont très strictes concernant la pédopornographie : toute photo d'enfant nu est systématiquement censurée, et cela avait été le cas également pour cette image, alors que chacun comprend que cela n'a aucun sens, en l'occurrence. À partir de cet incident, nous avons établi une nouvelle série de règles pour traiter le cas de photos historiques ou de contenus ayant un sens particulier justifiant qu'ils ne soient pas retirés de la plateforme.

L'application des règles

Historiquement, la modération des contenus se faisait à partir des signalements individuels. Un utilisateur estimant qu'un contenu est inapproprié peut nous le faire savoir. Nos équipes de modération prennent alors la décision de le retirer ou non.

Nous disposons également d'importants moyens de détection automatisée. En particulier, 99 % des textes et des images promouvant le terrorisme sont détectés de cette façon. L'identification des discours de haine est plus délicate mais le taux de détection est supérieur à 90 %, et l'analyse est complétée par la modération humaine. Le plus difficile est d'identifier les situations de harcèlement. Nous y travaillons énormément en ce moment.

L'auteur d'un contenu qui a été retiré peut « faire appel » (expression que je mets entre guillemets, car nous ne sommes pas dans le domaine judiciaire) s'il conteste cette décision. Nous avons également mis en place, depuis deux ans, un comité de surveillance afin que la décision en dernier ressort ne dépende pas des dirigeants de l'entreprise mais d'experts ayant suffisamment de recul, voire d'indépendance vis-à-vis de la direction pour déterminer en toute neutralité si la décision était justifiée ou non. Ce comité a un véritable pouvoir et il est déjà arrivé à Facebook de devoir rétablir des contenus qui avaient été supprimés. Il a également

interpellé la direction sur les décisions prises à l'encontre de Donald Trump, afin de s'assurer qu'elles n'avaient pas été arbitraires mais reposaient bien sur les règles affichées par l'entreprise. Nous sommes la seule plateforme à avoir instauré ce genre de dispositif externe qui, naturellement, ne se substitue nullement à l'autorité des tribunaux mais vient seulement compléter le dispositif interne.

Nous coopérons d'ailleurs, dans tous les pays, avec les autorités de police et de justice. Nous publions chaque année, pays par pays, un rapport dans lequel nous recensons l'ensemble des réquisitions (c'est-à-dire des demandes d'identification) reçues, au nombre d'environ 25 000 par an pour la France. Les progrès accomplis dans ce domaine sont très importants et ont abouti à de nombreuses condamnations. Le temps où certains pouvaient estimer qu'ils bénéficiaient, sur les réseaux sociaux, d'un anonymat leur permettant de faire et dire n'importe quoi est révolu. Les régulations en cours de négociations actuellement, à Bruxelles, autour du DSA (*Digital Services Act*) ont fait l'objet, en 2021, d'une « pré-transposition » en France, dans le cadre de la Loi confortant le respect des principes de la République. Celle-ci énumère les moyens que les plateformes ont l'obligation de mettre en œuvre.

Le bilan actuel

Dans le domaine des contenus explicitement violents, terroristes ou haineux, la situation s'est considérablement améliorée. Des marges de progrès existent pour les contenus « gris », c'est-à-dire offensants sans être complètement illégaux, et néanmoins susceptibles d'encourager le passage à l'acte. En voici un exemple pris dans l'actualité : des groupes Facebook ont été créés pour organiser des « convois pour la liberté », à l'instar de ce qui s'était passé pour les manifestations des Gilets jaunes. Certains de ces contenus relèvent du domaine de l'expression ; d'autres constituent des *fake news ;* d'autres encore peuvent être considérés comme des appels à la violence. Les distinguer les uns des autres n'est pas toujours simple.

La régulation de la communication en ligne

Lucile Petit

Lucile Petit est directrice des plateformes en ligne à l'ARCOM.

L'Arcom (Autorité de régulation de la communication audiovisuelle et numérique) est née, le 1er janvier 2022, de la fusion du CSA (Conseil supérieur de l'audiovisuel) et de l'HADOPI (Haute Autorité pour la diffusion des œuvres et la protection des droits sur internet). Cette transformation a entraîné de nombreux changements, mais la plupart des attributions de l'Arcom dont je vais vous parler faisaient déjà partie de celles du CSA. En revanche, je n'évoquerai pas la partie découlant de la directive sur les droits d'auteur, à la fois parce que je ne suis pas compétente dans ce domaine et parce que je me suis concentrée sur la lutte contre ce que l'on appelle le « désordre informationnel ».

Les compétences de l'Arcom

Je vais passer en revue les thèmes sur lesquels nous intervenons, et je le ferai dans l'ordre chronologique de publication des textes légaux correspondants.

2018

Le premier est la loi du 22 décembre 2018 relative à la manipulation de l'information. Pour la première fois, ce texte donne des compétences au CSA vis-à-vis des plateformes en ligne, à condition qu'elles accueillent au moins 5 millions de visiteurs uniques par mois en France, quel que soit leur pays d'établissement en Europe. Cette mesure concerne à la fois les réseaux sociaux comme Facebook ou Instagram, les plateformes de partage de vidéos comme YouTube ou Dailymotion, les moteurs de recherche comme Google ou Bing, les encyclopédies participatives comme Wikipédia, ou encore les forums. Les moyens d'action de l'Arcom consistent à émettre des recommandations sur la façon dont les plateformes doivent lutter contre la manipulation de l'information et à effectuer des bilans périodiques. Comme on le voit, il

s'agit d'une régulation par le *name and shame*, sans pouvoir de sanction.

2020

La loi du 30 juillet 2020 visant à protéger les victimes de violences conjugales donne au président de l'Arcom la compétence pour vérifier que les sites donnant accès à des contenus pornographiques s'assurent que leurs utilisateurs sont majeurs. Les moyens d'action sont plus coercitifs : possibilité de mettre un site en demeure, de saisir le juge judiciaire, d'édicter des lignes directrices sur les procédés techniques à mettre en œuvre. Cette loi s'applique quel que soit le pays d'établissement.

La loi du 19 octobre 2020 vise à encadrer l'exploitation commerciale de l'image d'enfants de moins de seize ans sur les réseaux sociaux et les plateformes de partage de vidéos. Les articles 4 et 5 de cette loi concernent les plateformes, et s'appliquent également selon le principe du pays de destination, et non celui du pays d'établissement. Les moyens d'action consistent à exiger des plateformes qu'elles adoptent des chartes sur la protection des mineurs participant à des vidéos, qu'ils soient seuls, par exemple en tant qu'influenceurs, ou en compagnie de leurs parents, par exemple dans le cas des influenceurs stars mettant en scène leur progéniture, et qu'elles établissent un bilan annuel.

L'ordonnance du 21 décembre 2020 portant transposition de la directive européenne SMA (Services de médias audiovisuels) concerne notamment les plateformes de partage de vidéos. Elle ne vise que celles établies en France, comme Dailymotion, et non celles implantées dans d'autres pays européens, comme YouTube, par exemple, qui relève de la compétence de l'Irlande. Les moyens d'action sont principalement l'édiction de lignes directrices et de codes de bonne conduite sur la communication commerciale, la protection des mineurs, l'absence d'incitation à la haine ou à la violence ainsi que d'apologie du terrorisme.

2021

La loi du 24 août 2021 conforte le respect des principes de la République en matière de haine en ligne et se veut une « pré-transposition » du projet de DSA (*Digital Services Act*). Le principe retenu est, cette fois, celui du pays de destination. Les plateformes potentiellement concernées sont les réseaux sociaux, les plateformes de partage de vidéo, les moteurs de recherche, les encyclopédies participatives et les forums. Il leur est demandé de prendre des moyens humains, procéduraux et technologiques proportionnés (ce terme est très important) et adaptés à la lutte contre la diffusion de contenus haineux. Les moyens d'action de l'Arcom sont la possibilité d'édicter des lignes directrices, l'accès aux principes de fonctionnement des outils automatisés déployés par les plateformes et à des données via leurs API (interfaces de programmation d'applications) ou de la collecte automatisée (*scraping*), la définition des indicateurs et des modalités de publication des rapports de transparence, ou encore la publication d'un rapport annuel. La grande nouveauté de ce texte est le fait que l'Arcom dispose d'un pouvoir de mise en demeure et de sanction des plateformes qui ne se conformeraient pas à leurs obligations légales.

Enfin, la loi du 22 août 2021 portant sur la lutte contre le dérèglement climatique et le renforcement de la résilience face à ses effets concerne les services de communication audiovisuelle et les plateformes en ligne selon le principe du pays de destination. Les moyens d'action de l'Arcom sont la promotion de codes de bonne conduite et l'exigence faite aux services ciblés de présenter au Parlement un rapport annuel sur leur mise en application.

2022

Deux compétences supplémentaires vont être attribuées à l'Arcom prochainement. À partir du 7 juin 2022, en application de l'article 6-1 de la Loi pour la confiance dans l'économie numérique, la mission de la personnalité qualifiée chargée de s'assurer de la régularité des demandes administratives du Pharos (Plateforme d'harmonisation, d'analyse, de recoupement et d'orientation des signalements) de retrait, de déréférencement ou de blocage de contenus terroristes et pédopornographiques dépendra désormais d'un membre du collège de l'Arcom, et non plus de celui de la CNIL (Commission nationale informatique et libertés).

Par ailleurs, en application du règlement européen de lutte contre le terrorisme entrant également en vigueur le 7 juin 2022, une proposition de loi devrait accorder à l'Arcom la compétence d'apprécier la régularité des injonctions de retrait, dans un délai d'une heure, de contenus terroristes faites à des fournisseurs de services d'hébergement établis en France par l'autorité d'un autre État membre, ainsi que la supervision des moyens mis en œuvre par les fournisseurs de services d'hébergement établis en France pour lutter contre la diffusion de contenus terroristes.

Un premier bilan

Comme on le voit, les compétences de l'Arcom sont nombreuses et variées. Elles portent aussi bien sur la désinformation que sur la pédopornographie ou le négationnisme. Certains de ces champs étaient déjà connus du CSA. D'autres sont nouveaux et l'Arcom a dû rapidement monter en compétence sur certains de ces sujets.

Dans tous les cas, il s'agit d'imposer des obligations de moyens ou de contrôler la régularité des demandes faites par une administration tierce. En aucun cas, l'Arcom n'est chargée de vérifier elle-même des contenus. Notre rôle est de nous assurer que les efforts

des plateformes sont proportionnés à l'objectif, c'est-à-dire à la fois suffisants pour lutter contre des phénomènes préjudiciables, voire illicites, sans pour autant conduire à une censure excessive de l'expression des citoyennes et citoyens.

L'Arcom a déjà publié deux rapports, en 2020 et 2021, sur les moyens déployés par les plateformes en ligne pour lutter contre la manipulation de l'information : transparence des outils automatisés, procédures de signalement, possibilités de recours, dispositif d'éducation aux médias, place accordée aux sources vérifiées et faisant autorité, etc.

Ces rapports soulignent l'existence d'une bonne coopération, en général, avec les plateformes concernées, avec quelques nuances pour certaines d'entre elles. Ils pointent le fait qu'en dépit de la richesse des informations transmises par les plateformes (le rapport rendu par Facebook en 2021 comportait, par exemple, environ 150 pages), certains renseignements manquent encore, notamment sur les moyens mis en œuvre spécifiquement pour la France, sur le nombre de signalements, sur la façon dont les recours sont traités, ou encore sur le ratio entre nombre de décisions et nombre de recours.

Notre dernière publication insiste beaucoup sur l'importance de la transparence non seulement vis-à-vis du régulateur, mais aussi, et avant tout, vis-à-vis du public, de la société civile et de la recherche académique. Face à la masse des informations mais aussi à la complexité de certains sujets, le régulateur ne peut pas représenter, seul, une solution miracle. Il est nécessaire que d'autres administrations puissent se saisir de ces questions, mais aussi que l'ensemble des parties prenantes, y compris les utilisateurs, puissent exercer leur rôle dans la chaîne des responsabilités.

Débats

Esprit critique et liberté d'opinion

Existe-t-il plusieurs sortes d'esprit critique, comme plusieurs formes d'intelligence ? Avoir l'esprit critique vis-à-vis d'une œuvre d'art, est-ce la même chose qu'avoir l'esprit critique devant un énoncé ou un graphe ?

Gérald Bronner : J'ai rédigé avec Elena Pasquinelli, pour le compte de l'Éducation nationale, un rapport intitulé *Éduquer à l'esprit critique*. En cherchant à définir cette notion, nous sommes parvenus à l'idée qu'avoir l'esprit critique consiste surtout à savoir dans quelles conditions faire confiance. En effet, ce que chacun de nous sait de source directe tient dans le creux de la main.

Selon notre définition, l'esprit critique concerne uniquement les catégories du vrai et du faux, et non celles du bien et du mal, ni du beau et du laid. Ces dernières doivent être, à mon sens, soigneusement protégées de tout discours rationnel, car elles relèvent de la liberté d'opinion.

Biais de popularité et itération

L'un des biais les plus puissants sur Internet n'est-il pas le biais de renforcement, qui nous conduit à être attirés par les contenus allant dans le sens de nos propres opinions ?

Gérald Bronner : C'est effectivement un biais assez bien mesuré, que l'on appelle aussi le biais partisan. Contrairement à ce que l'on croyait jusqu'à récemment, il a cependant moins de poids que la pensée paresseuse.

L'un des biais les plus puissants sur les réseaux sociaux est le biais de popularité. Nous avons tendance à accorder d'autant plus de crédibilité à un contenu qu'il est présenté comme apprécié par un plus grand nombre de personnes, a fortiori si ce sont des personnes en lesquelles nous avons confiance, comme nos amis. Dans notre rapport, nous avons d'ailleurs suggéré que, sur les plateformes, le nombre de « like » soit masqué par défaut.

Un autre biais est l'itération. Si, en vous promenant, vous croisez dix personnes différentes qui vous disent qu'elles ont entendu une explosion, vous avez tendance à penser que c'est vrai, et ce serait le cas en écologie naturelle. Aujourd'hui, les réseaux sociaux introduisent la possibilité que ces dix personnes aient pour source un seul individu affabulateur, ce qui constitue une perturbation de notre calibrage social.

Raison et passion

Vous semblez faire une distinction entre le champ de la logique et le champ des émotions. Or, l'émotionnel influe énormément sur ce que nous pensons. Mon neveu a contracté la Covid-19 et vient de passer trois semaines à l'hôpital. Il en est néanmoins sorti en prétendant que tout cela était « du pipeau ».

Gérald Bronner : Effectivement, et contrairement à une longue tradition de pensée, raison et passion ne s'opposent pas. On peut ressentir de l'indignation sur la base d'un raisonnement, et on peut aussi rationaliser une émotion. Je n'ai pas abordé ces aspects car ils sortent de mon champ de compétences.

Le biais d'intentionnalité

Certains biais cognitifs ne sont-ils pas créés par les algorithmes eux-mêmes, qui reflètent et reproduisent les biais de leurs programmateurs ?

Gérald Bronner : Je ne crois pas que les programmateurs introduisent délibérément des biais cognitifs dans les algorithmes afin de nous induire en erreur. En revanche, certains éléments algorithmiques se fondent sur le *deep learning* à partir des données que tout un chacun produit. C'est pourquoi il existe un emboîtement intime entre le fonctionnement des algorithmes et celui de notre cerveau, en sorte que le monde numérique nous apprend souvent beaucoup de choses sur nous-mêmes mais en nous caricaturant, c'est-à-dire en nous faisant nous abandonner collectivement à certaines pentes obscures de notre cerveau.

Dans ce domaine, je m'oppose à certains de mes collègues sociologues qui estiment que toutes les dérives de la communication en ligne procèdent d'artefacts mis sciemment en œuvre par les GAFAM. Je suis convaincu que les GAFAM ont leur part de responsabilité mais qu'une grande partie de ces dérives vient du fonctionnement de notre cerveau. Si on l'oublie, on n'aura aucune chance de résoudre le problème.

C'est ce que l'on appelle le biais d'intentionnalité…

Gérald Bronner : Tout à fait. Lorsqu'un phénomène produit des externalités négatives, nous avons du mal à les considérer comme des conséquences secondaires et nous en faisons généralement des conséquences primaires fondées sur des intentions. C'est la base même des théories du complot.

Danger, risque et incertitude

La confusion entre danger et risque n'est-elle pas un facteur important de biais cognitif ?

Gérald Bronner : Le risque est un danger pondéré par une probabilité. L'incertitude est un danger qui n'est pondéré par aucune probabilité. Dans notre environnement informationnel, beaucoup d'incertitudes se transforment en risques et beaucoup de risques sont considérées comme des dangers, quand ce ne sont pas les incertitudes qui se transforment directement en dangers, ce qui nuit beaucoup à l'acceptabilité de certaines innovations technologiques. Cette tendance procède de la surestimation des coûts par rapport aux bénéfices. En termes psychologiques, il faut environ 2,50 euros de bénéfice pour compenser 1 euro de coût.

Les biais cognitifs sont-ils universels ?

Tous les êtres humains présentent-ils les mêmes biais cognitifs ?

Gérald Bronner : Les phénomènes sociaux résultent de l'hybridation entre invariants mentaux et variables sociales, c'est pourquoi les biais cognitifs ne sont pas répartis également dans la population. Ils peuvent être modifiés par le niveau d'étude, ou encore par le genre ou la classe d'âge. Les femmes et les personnes âgées, par exemple, montrent, dans certaines situations, une plus grande aversion au risque que les hommes et les personnes jeunes, ce qui peut les inciter à surestimer encore davantage les faibles probabilités. Cela dit, certains biais cognitifs semblent universels, en particulier la confusion entre corrélation et causalité. Sans cette confusion, il n'y aurait pas de magie, or la magie est présente dans toutes les cultures…

Le suivi des signalements

Les personnes ayant signalé un contenu inapproprié sont-elles informées des actions engagées par la plateforme à la suite de leur démarche ?

Anton'Maria Battesti : C'est ainsi que cela doit se passer normalement. La personne reçoit un accusé de réception, puis est informée de ce qui a été décidé et se voit proposer de « faire appel » si la décision ne lui paraît pas satisfaisante. Elle est également informée des suites données à l'appel.

La répartition des tâches

Comment s'opère la répartition des tâches entre modération humaine et modération automatique ?

Anton'Maria Battesti : Les contenus ne nécessitant pas la prise en compte du contexte font l'objet d'interventions massivement automatisées. C'est le cas, par exemple, des photos comportant le drapeau de Daesh, ou des images pédopornographiques. En effet, en aucun cas on n'aurait le droit de publier une image pédopornographique dans le but, par exemple, de dénoncer ces pratiques. Ces images sont donc purement et simplement interdites.

Le cas des discours de haine est plus complexe. Par exemple, certains internautes peuvent recourir à l'ironie ou au second degré pour tenir des discours racistes ou proférer des menaces. Identifier ce type de discours exige un travail d'interprétation qui ne peut pas être réalisé de façon automatisée. De même, une victime de propos racistes qui cite ces propos pour s'en plaindre risque d'être censurée elle-même, si la modération est réalisée de façon automatisée. Dans ce genre de cas, l'intervention humaine reste irremplaçable. L'outil peut, en revanche, pré-détecter des contenus suspects et les soumettre au jugement d'un modérateur.

Comment sélectionnez-vous le modérateur auquel vous confiez l'analyse de chaque cas ?

Anton'Maria Battesti : Nous avons cherché à construire un système très simple pour l'utilisateur, mais également très efficace pour nos équipes. Lors de chaque signalement, la personne doit indiquer si le cas relève du harcèlement, du spam, d'une fraude ou d'un discours de haine, par exemple. Ces différentes catégories nous permettent d'orienter le signalement vers les bons experts.

Par ailleurs, la modération de certains contenus ne nécessite pas de compétence linguistique (par exemple, une photo de maltraitance) et pourra être adressée à des modérateurs travaillant dans n'importe quel pays, alors que, dans d'autres cas, il est nécessaire de pratiquer la langue utilisée, ce qui va également orienter l'attribution du cas à traiter.

Concrètement, combien de modérateurs employez-vous en France ?

Anton'Maria Battesti : Nous faisons appel à plusieurs centaines de modérateurs parlant français, mais les contenus publiés en France et ne nécessitant pas de compétence linguistique peuvent être modérés par des non-francophones, et beaucoup de contenus sont modérés de façon automatisée.

Cela dit, nous sommes soumis à une obligation de moyens, et non de résultats. À nous de nous organiser pour atteindre les objectifs qui nous sont fixés.

Le code pénal doit-il être complété ?

Le code pénal permet-il de sanctionner ces nouveaux types de délits ou faudrait-il inventer de nouveaux outils de régulation juridique ?

Anton'Maria Battesti : Compte tenu du volume colossal des données à traiter, le législateur a préféré demander aux plateformes d'effectuer un premier travail de régulation. Parmi les milliers de personnes ayant harcelé la jeune fille connue sous le nom de Mila, une douzaine seulement se sont retrouvées devant les tribunaux et, heureusement, elles ont été condamnées, ce qui constitue un signal très important envoyé aux utilisateurs des réseaux sociaux. Des milliers d'autres personnes ont échappé à la justice, mais la plateforme a suspendu les comptes de celles qu'elle a pu identifier.

Le délai de retrait des contenus

Dans quel délai un contenu haineux est-il supprimé ?

Anton'Maria Battesti : Certains contenus ne sont jamais publiés, dans la mesure où les mécanismes de filtrage permettent de les détecter instantanément. D'autres sont publiés et peuvent être détectés après coup, soit de façon automatisée, soit par le biais d'un signalement. En fonction du degré de menace ou d'urgence, certains contenus peuvent être supprimés en quelques heures. D'autres, malheureusement, ne sont jamais signalés, ni détectés automatiquement, ni traités.

Le processus est rendu plus complexe par le fait que certains internautes signalent tout et n'importe quoi et, inversement, que certains ne dénoncent pas des discours de haine dans la mesure où ils portent sur une autre communauté que la leur… Les algorithmes, pour leur part, visent à identifier tous les discours haineux, quelles que soient leurs cibles.

Fake news idiotes et Fake news dangereuses

Supposons que je publie sur un réseau social un message encourageant mes lecteurs à ne pas se faire vacciner contre la Covid-19, à ne pas respecter les gestes barrière et à se bourrer d'hydroxychloroquine, le tout en m'appuyant sur une vidéo du docteur Raoult s'exprimant sur BFMTV. Quelles seront ma responsabilité, celle du réseau, celle de l'orateur que je cite, celle du média qui l'a accueilli ?

Anton'Maria Battesti : Les règles de notre plateforme distinguent les *fake news* dangereuses de celles qui sont simplement idiotes. Si vous tenez des propos mettant les gens en danger, nous supprimerons votre contenu. Nous l'avons fait pour plusieurs millions de messages ces derniers mois. La mise en cause de la responsabilité du médecin invoqué ou de la chaîne de télévision qui lui a donné la parole relève d'autres sphères d'action mais, en ce qui concerne notre plateforme, nous avons décidé que ce type de propos n'y avaient pas leur place.

La coopération avec les autres pays européens

Lucile Petit, de quelle façon coopérez-vous avec vos homologues dans les autres pays européens ?

Lucile Petit : Certains textes de loi rendent nécessaires des procédures de coopération, en particulier la directive européenne SMA, fondée sur le principe du pays d'établissement et non du pays de destination. C'est ainsi que nous sommes compétents pour toute l'Europe en ce qui concerne Dailymotion.

L'Arcom est, par ailleurs membre, depuis 2013, du Groupe des régulateurs européens des services de médias audiovisuels (European Regulators' Group for Audiovisual Media Services - ERGA), qui a constitué des groupes de travail sur l'évolution des textes et notamment sur le projet de DSA. Nous pouvons aussi échanger de façon bilatérale ou multilatérale avec certains régulateurs en fonction de nos centres d'intérêt, par exemple, récemment, sur les questions de vérification d'âge.

Mots clés : biais cognitifs, esprit critique, haine en ligne, infox, management des contenus, plateformes numériques, régulation, réseaux sociaux

Citation : Nicolas Curien, Gérald Bronner, Anton'Maria Battesti & Lucile Petit. (2022). *Dérives de la communication en ligne : constats et remèdes*. Les séances thématiques de l'Académie des technologies. https://www.academie-technologies.fr/publications/derives-de-la-communication-en-ligne-constats-et-remedes/