COMITÉ CONSULTATIF NATIONAL D'ÉTHIQUE
POUR LES SCIENCES DE LA VIE ET DE LA SANTÉ

NATIONAL PILOT COMMITTEE FOR DIGITAL ETHICS (CNPEN)

Reflections and warning points on digital ethics issues in situations of acute health crisis

Ethics oversight bulletin 2

Ethical issues in the fight against disinformation and misinformation

Tuesday, 21 July 2020

Disinformation and misinformation are two phenomena that have been exacerbated during the crisis caused by the SARS-CoV-2 epidemic. This situation has pressed digital platforms like social media platforms, search engines or video sharing networks to develop algorithmic practices and digital tools to help them in fighting the damaging effects of disinformation and misinformation both on the individual and social level.

The purpose of this bulletin is to identify the ethical issues and challenges arising from the widespread use of these different algorithms and tools, which are part of a complex phenomenon with wide-ranging implications. Among others, some questions arise: what does action or inaction in this domain mean in the context of COVID-19? Is it simply a quantitative shift, or are we seeing a more profound change in the nature of the digital solutions designed to fight online disinformation and misinformation? More generally, how do we face the complexity of this phenomenon, which requires an analysis that seems to go beyond ethics, or even to challenge the notion of ethics itself? Indeed, ethical questions do not arise in the same way depending on whether one is dealing with actors who act consciously to deceive their target or, on the other hand, whether one is dealing with actors who simply get caught up in the flow of information in digital format and, in particular, participate in the virality of this information often in an unconscious way. In the first case, ethical reflection questions responsibility, while in the second case it consists mainly in moving towards awareness. In either case, the requirement is to identify – specifically in the digital domain – the economic, legal, social, political or philosophical dimensions of disinformation or misinformation.

This bulletin is part of the oversight activity that the CNPEN has been pursuing since the beginning of the health crisis.[1] Its objective is to contribute to that oversight process from an ethical perspective by analyzing and discussing the measures that the social media platforms have undertaken or not during the COVID-19 crisis to fight the spread of disinformation and misinformation. In the light of that specific context, this bulletin formulates recommendations and identifies several highlights for continuing the reflection on these phenomena of disinformation and misinformation in the digital era.

Emmanuel Didier, Serena Villata, Célia Zolynski
Rapporteurs for the working group
Claude Kirchner
Director of the National Pilot Committee for Digital Ethics

[1] https://www.ccne-ethique.fr/fr/actualites/comite-pilote-dethique-du-numerique-bulletin-de-veille-ndeg1

CONTENTS

INTRODUCTION

While rumour – i.e., the public spread of information with uncertain provenance and doubtful veracity – has always existed, in the digital era the phenomenon takes the form of a potentially mass-scale propagation, often deliberate or automated, of information of all kinds. The intentions of the authors or the propagators of this information can be many and various. Some information is deliberately created for the purpose of subtle deception, generating confusion, or misleading people as well as organisations and public opinion, or to promote certain economic and political interests. The act of spreading this kind of information with the deliberate intention to mislead, to cause public harm, or to obtain an economic advantage, is classified as "disinformation". Other information may turn out to be uncertain, incomplete or incorrect, though presented as reliable and passed on in good faith by people. This includes a whole range of scientific information poorly understood or poorly interpreted, or concerns contents that are unfounded or lacking scientific foundation, which are disseminated on a mass scale by the social media platforms.[7] People who spread this content are generally unaware of its viral effects and of the consequences of their actions, and also that, by doing so, they contribute to the business model of these platforms. This phenomenon is usually " misinformation".[2]

Produced and disseminated over the Web by social media platforms, websites, forums or instant messaging platforms, disinformation and misinformation have expanded to unprecedented scale since 2016, notably with the US presidential election and the Brexit campaign, and in 2017 with the French presidential campaign. The health crisis caused by COVID-19 has intensified this phenomenon to the point that the United Nations and several of its agencies (WHO, UNICEF) now refer to the situation as an "infodemic".[3] Lockdown, self-isolation, anxiety, the gravity of the situation or simply the multiple factors of uncertainty, provide a fertile ground for the spread of disinformation and misinformation. There is disinformation and misinformation also about the origin and prevention of the SARS-CoV-2 virus, about possible treatments, about the consequences of the epidemic, about lockdown and its end, about track and trace campaigns, about discrimination against certain populations, about reports of shortages that cause unnecessary social disruption, or about false or malicious advertising.

---

[2] On this distinction, see the European Commission communication entitled *Tackling Coronavirus Disinformation – Getting the Facts Right*, 10 June 2020, JOIN/2020/8 final, p. 4 ff. CELEX 52020JC0008 FR TXT-1.pdf

[3] "Infodemics are an excessive amount of information about a problem, which makes it difficult to identify a solution. Infodemics can spread misinformation, disinformation and rumors during a health emergency. Infodemics can hamper an effective public health response and create confusion and distrust among people." https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200305-sitrep-45-covid-19.pdf?sfvrsn=ed2ba78b_4

This new scale of disinformation is intimately linked to the emergence of social networks,[4] search engines,[5] and video sharing platforms,[6] which we refer to here generically as platforms.[7] These platforms significantly increase the capacity of their users to enjoy individual freedom of expression, by contributing to the spread and interchange of information. Everyone is free to express her own opinions on these platforms, or at least this is how they are perceived by users. By leaning on the claim of supporting this principle of individual freedom, platforms have been able to present themselves as simple channels of information without any editorial responsibility. However, it has been apparent, at least since the mid-2010s, that this freedom still requires some sort of content management. The scenes of violence – in particular associated with terrorist acts – or pornography transmitted via these platforms have shown that this kind of content can lead to dangerous consequences for certain groups of users or for society as a whole. In the current circumstances, certain information about the epidemic, such as advertisements for fake treatments, can have serious health consequences. They can also accentuate people's mistrust in the public authorities and make the health crisis harder to manage. These fast emerging effects have prompted some platforms to put additional effort to moderate their content, or even to remove or promote certain information.

This process of moderation is extremely complex: every piece of information can potentially become misinformation or disinformation depending on the context of presentation, the mode of formulation, or the point of view of the recipient. Indeed, information is not just true or false in the sense of being assigned with a truth value. It also forms part of a nexus of actions or a context, a pragmatics, in other words it is assessed in terms of its sources and its proven or presumed effects. This assessment therefore always includes a degree of uncertainty and a political perspective: news are necessarily received and interpreted in the light of a set of assumptions, and social and political effects that are specific to each recipient of the information. The truth status of a piece of information is thus as important as the pragmatics of its propagation, i.e., its consequences and empirical effects (who benefits from it, how it can be used to form alliances, what types of actions it triggers, etc.).[8] The problem therefore lies in the entanglement between, on the one hand, the assessment of the truth value of the information disseminated on the Web and, on the other hand, the degree of freedom of expression left to the users with respect to the consequences of their

[4] Facebook, Tiktok, LinkedIn, Twitter, WhatsApp, Mastodon, etc.

[5] Google, Qwant, DuckDuckGo, Ecosia, etc.

[6] YouTube, DailyMotion, Snapchat, etc.

[7] On this terminology, see the European Commission communication entitled *Tackling Coronavirus Disinformation – Getting the Facts Right*, or else the Report of the taskforce on "Régulation des réseaux sociaux – Expérimentation Facebook" [regulating social media – Facebook experiment], https://www.vie-publique.fr/sites/default/files/rapport/pdf/194000427.pdf

[8] See also position paper 2018-37 issued by COMETS (CNRS Ethics Committee) – *Quelles nouvelles responsabilités pour les chercheurs à l'heure des débats sur la post-vérité ?* [*What new responsibilities for researchers in the era of post-truth debate]?* 12/04/2018: https://comite-ethique.cnrs.fr/wp-content/uploads/2019/10/AVIS-2018-37.pdf

speech acts.[9] The complexity of this ethical evaluation is further exacerbated by the fact that, on social media, any individual or any group (which can be formed quickly and spontaneously, often solely online) can diffuse its own opinions at global scale. This lack of selectivity and levelling of social hierarchies is a key factor for this ethical analysis. The situation emerged during the Covid-19 crisis has further contributed to this informational relativisation: first, lockdown isolated individuals and made them more dependent from social media platforms; second, science, which always works to a slower timeframe than the news, had to incorporate elements of uncertainty into communications that are of vital import to the entire population.

In the absence of a universal criterion for assessing whether a piece of information is false or not, as the legitimacy of information depends on the perspective from which it is seen and on the society in which it is produced, the platforms have tried to establish frameworks for determining what can and cannot be broadcast. To do so, they sometimes rely on fact-checking organisations, often set up by newspapers, or on national and international health or governmental authorities. They also set their own criteria for differentiating content that is deemed illegal or dangerous from other content. This detection process often relies on the large-scale use of algorithms and automated tools. The Covid-19 health crisis showed that these resources are not enough: because of the scientific nature of the news being disseminated and the controversies around them, it became even more difficult to identify authorities with the legitimacy to assess the information spread on the Web. The crisis also made it clear that these platforms cannot on their own decide about the content selection and ranking procedures.

These phenomena and measures, sometimes intensified by the health crisis, raise different ethical questions. First, there is the need to explore the potential risks of disproportionate attacks to the freedom of expression right . It is essential, even in a time of crisis, to maintain the fundamental principles of our democracies, such as the access to information, the freedom of expression, the media independence, and an open transparent discussion.[10] Other questions arise about the legitimacy and the effects of the power – both digital and political – that these social media platforms seem to have acquired under the pretext of seeking to tackle the spread of disinformation or misinformation. The connections between this power and the power of the pre-existing authorities (e.g., governments, the courts) needs to be considered and analyzed. More generally, it seems essential to consider the ethical responsibility shared by the different actors that contribute to the dissemination of this content by means of online digital platforms .

Actions of this kind can result in different outcomes. One might take the view that any suppression is an attack to freedom, or conversely that the damage caused by the

[9] J.L. Austin, *How to do things with words*, Oxford, Clarendon Press, 1962.

[10] On this point, see the Statement on Freedom of expression and information in times of crisis by the Council of Europe's Committee of experts on media environment and reform (MSI-REF), 21 March 2020, together with *Respecting democracy, rule of law and human rights in the framework of the COVID-19 sanitary crisis*. *A toolkit for member states*, 7 April 2020, SG/Inf(2020)11.

suppression of certain forms of irony or humour specific to social media is inferior than the damage caused by the spread of disinformation and misinformation. One might also fear that such practices could have a ripple effect, leading to the same kind of censorship or self-censorship outside social media, and ultimately to the impoverishment of social life in general. Or that this change is likely to happen, but is not for us to assess because future generations will decide for themselves whether it is good or bad. The situation can be summed up as the product of tensions between three elements: first, the commitment to the fundamental right of freedom of expression; second, the identification of authorities – new or old ones – with the legitimacy to determine the scope of such right, as well as the limits that constrain the actions of those authorities; third, the practical procedures for moderating the interactions between users and platforms through which the decisions made by those authorities are implemented. From this triangle of tensions many ethical questions emerge.

The purpose of this bulletin is not to assess the truth value of certain information nor the immediate consequences of its diffusion; that is a task for the Web actors, social media platforms and the competent authorities. Its aim is to identify and analyze the ethical issues raised by the institutional choices enacted by the platforms to tackle the phenomena of disinformation and misinformation spread. In other words, the aim will not be to analyze how disinformation is produced but to consider the choices made or deliberately not made by these different actors in response to this " infodemic" phenomenon, in order to identify the ethical tensions that their actions or inactions can generate.

These tensions relate first of all to the implementation of the moderation tools and anti-virality mechanisms that the platforms employ (I); they also relate to the relations between these operators and the different governmental, judicial and scientific authorities, as well as the press (II).

MODERATION TOOLS AND VIRALITY MECHANISMS

In order to counter the spread of disinformation, which is particularly high in the context of the COVID-19 crisis, the different social media platforms propose a variety of ways to deal with the content they carry, whether by removing it, reducing its visibility or promoting it.

Removing content: Most of the platforms remove content that might present a clear and concrete danger or be detrimental to public health (for example, by disputing a decision or a recommendation made by a public health authority or a scientific information) or, more generally, that might be detrimental to the integrity of another person or to public order. They can also refuse to diffuse advertisements that are identified as misleading or deceitful, or as exploiting panic. During the COVID-19 crisis, for example, some platforms have suspended fake accounts set up by users claiming to be health bodies or those identified as spreading information that is false and potentially dangerous to health.

Reducing visibility: Several platforms have decreased the dissemination of certain content by lowering its ranking in the homepage. They may also warn users about doubtful content and redirect them to articles or fact-checking pages on the topic. Others limit the number of content transfers possible or block accounts from which mass transfers are made.

Promoting content: Various platforms promote certain news by means of on-screen ribbons, and editorialize content or new threads originating from sources that they consider trustworthy, such as public health authorities, government departments or fact-checking websites. They may also promote such information by flagging it in their index or by increasing its visibility with free advertising.

While these measures, which rely both on technical and human methods, might in principle seem an appropriate way to tackle disinformation and misinformation spread, they also raise a number of ethical problems, relating first to the use of automatic tools to detect this kind of information (I.A), and second to the viral mechanisms, i.e., the rapid and unpredictable spread of this content, which contribute to its worldwide propagation (I.B).

A. Automatic tools to fight online disinformation and misinformation spread

Given the huge volume of information circulating on the Web, in particular through online social media platforms, automatic tools would seem to be the only way to make the detection of disinformation or misinformation more efficient and scalable than human fact-checkers could manage. However, these algorithms raise numerous ethical questions concerning both their reliability and their impact on the right of freedom of expression.

First of all, when totally automated, these tools may face algorithmic biases or classification errors. More precisely, these tools use a variety of disinformation or misinformation detection algorithms targeting different media (i.e., text, video, pictures). These algorithms rely in particular on the recognition of keywords in texts, on the detection of recycled images through the analysis of the chronology, or on the analysis of the angle and the expression of faces, including the lighting and other important features, to check the authenticity of videos of people.

Background on disinformation and misinformation detection

Assessing the authenticity of a piece of information is a complex and difficult task, even for qualified experts like human fact-checkers. For example, a first step in identifying content that might be classified as disinformation or misinformation is to analyze what other information sources say about this piece of information. This automatic task is called stance detection and consists in evaluating the relative position of two pieces of text on a certain subject, in order to establish the consistency of the content.

There are different labelling or classification strategies for detecting content that might contain disinformation or misinformation. In most approaches, detecting such information is formulated as a classification or a regression problem. The most common approach is to formulate the task of detecting such information as a binary classification problem (i.e., classifying it as either disinformation or non-disinformation). However, classifying this content into two classes is difficult because there are cases where only part of the content can be classified as being disinformation or misinformation. Detecting this information can also be formulated as a regression task, where the result is a numerical veracity score.

These algorithmic approaches to the automatic detection of disinformation or misinformation face a number of challenges. A first significant challenge deals with the availability and quality of data: for the (supervised and semi-supervised) classifiers to perform effectively, they need sufficient quantities of annotated data, but reliably annotating a large volume of data requires a long and complex effort by qualified experts. Context detection is another significant challenge in this area. It entails developing algorithms that are capable of effectively analyzing long-term and content transition information from basic knowledge. Finally, a third challenge is the cross-referencing of multimodal data. This is because, in some cases, the effective detection of fake content requires the cross-referencing of different types of information, such as text and images, and the metadata associated with that content.

Given these difficulties, algorithms for detecting disinformation or misinformation in online content can produce classification errors, called “false positives”, i.e., pieces of information that are wrongly classified as disinformation or misinformation. This is particularly challenging when dealing with humorous, ironic or satirical content, where accurate classification requires a high level of background knowledge. Disinformation detection can also lead to “false negatives”, where the algorithm fails to detect content that is actually containing disinformation or misinformation, which may then continue to circulate on the platforms. Moreover, any operational biases in the algorithms can influence the detection of misinformation or disinformation. These biases, which may arise from conscious or unconscious choices done by the developers or be embedded in the data themselves, can lead to indirect discrimination against certain groups of users.

Although necessary because of the large volumes of information to be analyzed, this algorithmic approaches of content checking can therefore raise risks of censorship and place disproportionate pressures on the right of freedom of expression. These risks of algorithmic bias and classification errors are even more important in a context where mechanisms of mediation and validation by humans are missing. In fact, during the health crisis, it emerged that the platforms were unable to allow their teams of moderators, who were working remotely, to access all the necessary information, because of its potentially intrusive or disturbing content (violent content, hate speech,...). Indeed, the often unanticipated conditions of teleworking could have led to the use of unsecured networks for the transfer of such (potentially criminal) content or to a moderation under conditions that could not easily be managed. As a result, this leaded to both less human control on these measures for the removal, downgrading or promotion of content, and a sharp increase in disinformation and misinformation. Moreover, the large-scale use of machine learning tools, along with the absence of subsequent human oversight, highlights a risk of automatic censorship and potentially undermines the capacity of the authors of the censured content to appeal against the withdrawal performed by the platform.

Next, the large-scale use of machine learning tools raises the question of the transparency and explainability of the algorithms used to detect disinformation or misinformation online. There are two separate aspects to be considered in discussing this problem. First, the explanation of the result generated by the algorithm and the main factors included in reaching it (e.g., the features used by the supervised classification system to classify content that constitutes disinformation, the degree of reliability of the results obtained by the algorithm, the training data with the biggest influence on the classification task, the most significant features within the layers of a neural network). Second, the criteria chosen by the platforms when setting their moderation policy (e.g., whether relating to financial reasons or relating to legal obligations). The algorithmic solutions employed in these tools can thus lead to “decision” (or deliberation) biases that influence content moderation. The question is then whether the platform should be transparent in informing its users and the

regulators about these different criteria, in line with the obligations imposed on it by the Act of 22 December 2018 on tackling misinformation at election times.[11]

Recommendations:

1.1 There should be a guarantee, including after the crisis, of the existence of human moderation to check the results generated automatically by the content analysis algorithms.

1.2 Instruments should be maintained, including during the crisis, which enable a content provider (i.e., a user) to appeal in the case of a decision taken by the platform to remove or promote content on the basis of algorithmic processing.

1.3 The disinformation or misinformation detection and the content recommendation algorithms used by the platforms should be transparent and explainable, and therefore auditable, in particular during crisis. Users should be informed about the criteria underlying the algorithmic decision making in order to protect the right of freedom of expression with respect to the three actions proposed by the platforms: content removal, downgrading of content visibility, content promotion.

Another problem concerns the inequality of means made available to public research and private social media platforms for the task of detecting content containing disinformation or misinformation. This is mainly due to the very limited access to data of the social media platforms. These data are extremely useful, both to human fact-checkers and for the improvement of automatic detection systems. The annotations used to identify the type of disinformation or misinformation (e.g., a truncated quotation), meta-data such as the source, date and time of publication or sharing rate of online content, can help to improve the performance of detection algorithms. This in turn raises questions about the control of data associated with disinformation or misinformation identified and collected by the platforms and the benefits of encouraging the sharing of these data between different actors.[12] The question arises in particular when the government, for example during the COVID-19 crisis, wishes to involve the platforms in public anti-disinformation policy. It would be useful, for example, to develop better targeted and synchronized sharing mechanisms

[11] Act 2018-1202 of 22 December 2018 on tackling information manipulation, Art. 11. See also CSA recommendation no. 2019-03 of 15 May 2019: "The Council encourages the platforms to give an assurance to every user regarding:
- the traceability of their data use for purposes of recommendation and content ordering, whether they are provided knowingly or collected by the online platform operator;
- clear, sufficiently precise and easily accessible information about the criteria used in the ordering of the content provided and the classification of those criteria according to their weighting in the algorithm;
- clear and precise information on its ability, if that ability exists, to undertake adjustments for the purpose of customizing content indexing and recommendation;
- clear and sufficiently precise information on the main changes made to the indexing and recommendation algorithms, together with their effects;
- an accessible communication tool that offers real-time interaction between the user and the operator, and gives the user the possibility of obtaining personalized and precise information on how the algorithms work."
With regard to the relations between the platforms and the users/consumers, see also the obligations imposed on operators under Articles L. 11-7 ff. of the Consumer Code.

[12] See Article 7-III of the Act against the dissemination of online hate content (version passed by Parliament but deemed unconstitutional by Constitutional Council ruling no. 2020-801 DC of 18 June 2020) encouraging the Internet platforms, under the aegis of the CSA, to implement open format cooperation and information sharing tools between these operators, in order to tackle the offences targeted by the law (hate content).

in Europe or internationally, while offering scientists, citizens and the civil society the possibility of contributing to their development and improvement.

Focus point:

1.a As recommended in the report of the "Social Media Regulation – Facebook Case Study"[13] taskforce and by the European Commission,[14] an extensive discussion should take place on the establishment of common databases in order to improve digital tools for tackling disinformation and misinformation, and to encourage the platforms to share the metadata associated with the data they collect for this purpose (e.g., source, subject, citations, sharing on platforms, counter-arguments published as commentary on this content). These databases would also boost scientific research in this domain.

Beyond this, another question concerns how the platforms actually implement these automatic and human methods in order to tackle disinformation and misinformation detection. In particular, a certain ambiguity has been observed in the attitude of certain platforms which, while making announcements about the measures taken to tackle the spread of online disinformation related to the COVID-19 crisis, continued to advertise the sites where this disinformation originated.[15] With regard to the prevention of online hate content, the legislation formulated to fight such content online has proposed to proceed by requiring the platforms to inform the CSA about the automatic and human methods they employ to limit the spread of the such content (see the list of offences cited in the legislation).[16] The Act of 22 December 2018, which requires the platforms to take measures to combat the spread of fake news disturbing public order or damaging the integrity of the elections, specifies that these measures and their implementation procedures must be made public.[17] Moreover, in June 2020 the European Commission announced that the platforms would be asked to submit a monthly report on the policies and measures they employ to limit the diffusion of disinformation around the COVID-19 health crisis, notably by providing data on the advertising streams associated with disinformation. The oversight authorities and their users will therefore be able to assess the effectiveness of these measures and the concreteness of the promises made by the platforms.[18]

---

[13] Report of the "Social Media Regulation – Facebook Experiment" taskforce – https://www.vie-publique.fr/sites/default/files/rapport/pdf/194000427.pdf, p. 18.

[14] On this subject, see the aforementioned Communication entitled "*Tackling Coronavirus Disinformation – Getting the Facts Right*", section 5.2.

[15] On this point, see in particular the reports of the NGO Tech Transparency Project: https://www.techtransparencyproject.org/articles/google-profiting-coronavirus-conspiracy-sites

[16] Legislation to tackle online hate content, Article 5 deemed unconstitutional by the aforementioned ruling of the Constitutional Council. With regard to the criticisms advanced in response to the inaction of certain platforms, see for example the action brought by four organizations (Union des étudiants juifs de France (UEJF), J'accuse, SOS-Racisme et SOS-Homophobie) against Twitter: https://www.lemonde.fr/pixels/article/2020/05/12/twitter-assigne-en-justice-pour-son-inaction-massive-face-aux-messages-haineux_6039412_4408996.html

[17] Act 2018-1202 of 22 December 2018 on tackling news manipulation, Article 11.

[18] On this subject, see the proposals made by the European Commission in its aforementioned Communication of 10 June 2020, p. 10-11.

Recommendation:

1.4 Some mechanisms should be introduced to ensure that the platforms publish a regular activity report, accessible by the oversight authorities and their users, which gives a clear, fair, accurate and transparent account of their policy on tackling disinformation and misinformation, the algorithmic and human methods employed to this end, and the data related to the advertising streams linked with disinformation.

Focus point:

1.b Extensive future discussions are needed on the allocation of resources to human moderation and the resulting costs, the training of human moderators, the handling of cultural biases and perhaps the criteria on which moderatos' decisions are based, and the possibility of monitoring by an independent authority (in particular judges) as guarantors of fundamental freedoms.

B. Virality mechanisms

The current scale of disinformation and misinformation is attributable to an increase in the spread of such content by viral mechanisms that are activated through the tools provided by social media platforms and search engines. Their business model can amplify this phenomenon when it depends on the attention economy that promotes virality as a profit-making mechanism (1.2.1). Users also play a role in this phenomenon since micro-actions (forwarding, sharing, etc.) are the initial triggers of the virality of such content (1.2.2).

a) Platforms' business model encourages virality

The business model of some platforms is based on remuneration based on the number of clicks – hence on the promotion of "clickbait" – and relies on capturing and monetizing their users 'attention for advertisers. This model is based on an algorithmic system for promoting content which generates the most reactions and conversations. As a result, it contributes heavily to viral mechanisms. The effects of such processes can be detrimental: they spotlight content that attracts more attention through viral mechanisms (often violent or hate content or fake news),[19] by comparison with the mainstream press content. The latter is often ranked downwards to increase the visibility of viral content that generates higher advertising revenue.[20] These "engagement metrics", whose primary goal is to capture, retain and monetize the attention of social media users, are encoded in the algorithms, which accentuates the risk of promoting disinformation, if such information attracts more attention. Moreover, this model seems to lead to a kind of hyper-customization of content, which encloses users in social media bubbles and intensifies

[19] D. Cardon, *A quoi rêvent les algorithmes. Nos vies à l'heure des big data,* Seuil, 2015, p. 91.

[20] B. Patino, *La civilisation du poisson rouge. Petit traité sur le marché de l'attention*, Grasset, 2019, p. 141.

their sociological and cognitive biases– the tendency to form bonds with people who are like us – and confirmation bias – the tendency to favor information that confirms our own beliefs.[21] The measures introduced by the platforms to help tackling the Covid-19 crisis should not be used to hide the flaws in their business models.

Focus point:

1.c It would be desirable to undertake a more in-depth analysis of the mechanisms underpinning online advertising markets, whose origins and price setting criteria are currently opaque and raise a number of ethical questions.

b) Viral mechanisms and the role of users

While the platforms undoubtedly play an amplifying role in the dissemination of content, the different causes of viral mechanisms need to be analyzed, in particular the role of individual users or groups of users in this process.[22]

Users in fact play a role at two levels: on the one hand, as recipients of information, and on the other hand, as agents of virality when they contribute to the spread of information by posting, reposting and commenting on text or visual content (memes, GIFs, etc.). In this respect, we need to distinguish between two types of user, while recognizing that an individual or group can move between these two types according to circumstances. On the one hand, there are those who take part intentionally in this propagation process, usually for ideological or financial reasons. This type of behavior raises the question of liability and the legal consequences applicable to such practices. On the other hand, individuals may contribute to virality through simple negligence or ignorance about the damaging effects that it can produce.

In the latter case, the goal should be to encourage people to carefully thinking before deciding to share some information and thereby contributing to its viral spread. While in ethical terms, they are responsible for this sharing as agents of the spread of misleading information or fake news, promoting a more responsible behavior requires the platforms to give these users the tools to become aware of – and hence control – the role they play in the viral information chain. In this domain, the Act of 22 December 2018 requires the platforms to set up appropriate tools to inform users about the nature and the origin of online content, and how it is spread. The CSA thus advises online platforms to ensure that they: [23]

---

[21] The latter phenomenon is nevertheless disputed: see F. Tarrisan, *Au coeur des réseaux*, Le Pommier, 2019, p. 115 ff.

[22] On this subject, see the report on reinforcing the fight against online racism and anti-Semitism, submitted to the Prime Minister in September 2018 by L. Avia, K. Amellal and G. Taieb, https://www.gouvernement.fr/sites/default/files/contenu/piece-jointe/2018/09/rapport_visant_a_renforcer_la_lutte_contre_le_racisme_et_lantisemitisme_sur_internet_-_20.09.18.pdf, pt. 6.1.

[23] CSA recommendation no. 2019-03 of 15 May 2019 to online platform operators within the framework of the duty of cooperation in preventing the spread of fake news, no. 5.

- clearly distinguish between sponsored content and other content, and foster the development of tools that enable users to identify the criteria that prompted the platform to provide them with such content;
- - advise users to be vigilant with respect to flagged content;[24]
- clearly identify and indicate the origin of the disseminated content;
- specify how the content is spread, as far as possible indicating the conditions of its publication, such as the existence of financial remuneration, the scale of the dissemination (e.g., number of views, targeted population type, etc.), and whether this content is automatically generated.

In order for users to be able to measure and manage their role in the viral chain, social media platforms should also be asked to offer their users the possibility of:

- measuring, in their use of social media, their role in promoting content that they spread directly (by posting it on their own profile or by retweeting or sharing), or indirectly (by liking).
- being aware that the news they post or share conveys information that is notably used by social media platforms to profile them more accurately.

<u>Recommendations to the platforms:</u>

1.5 The recommendations formulated by the CSA on establishing appropriate systems for informing users about the nature, the origin, and the mechanisms of dissemination of online content should be promoted, and social media platforms should be asked:

  - to inform their users explicitly when they propose them a piece of information that has been widely shared;
  - to be vigilant before sharing flagged content.

1.6 Tools should be developed and made available to inform social media users about their role in promoting content that they spread, directly or indirectly, through the platforms.

<u>Focus points:</u>

1.d While it is important, as the CSA underlines, to promote tools that enable users to identify the criteria that prompted the platform to recommend certain contents (to understand what they see), another ethical focus will lead to the development of tools that will allow users to specify their criteria about the content to be made visible, so that their own interests are taken into account.

1.e Moreover, and given the volume and speed of propagation of the information disseminated on social media platforms, some considerations could be addressed to discuss the desirability of slowing down this dissemination process by digital means.

[24] For example, the flagging of fake news by the platforms.

The aim would be to encourage users to be more careful and to analyse content before spontaneously sharing a piece of information.

More generally, it would be desirable to promote the development of critical thinking among users, so that they can share content in a mindful way. Platforms could, for example, remind users about the desirability of assessing the accuracy and reliability of their information sources, for example by comparing a piece of information with other information they hold, and by conducting a search, albeit short, on the sources and articles related to that information.

In particular, users would be encouraged to:

1.f try to identify the source of the information, to question its trustworthiness, to check its legitimacy and cross-check different sources on the same subject.

1.g think, before sharing, about the potentially uncertain nature of a piece of information, especially in the circumstances of a health crisis characterized by multiple unknowns.[25]

Improving users 'critical thinking skills would above all entail developing their digital culture both in scientific terms[26] and with respect to digital tools. In this respect, innovative new resources have been developed by different institutions connected with the health crisis, which would need to be complemented by longer term measures.[27]

Moreover, education on digital tools should be made available to the most vulnerable groups in the society, in particular young and elderly people, as part of a lifelong learning approach.[28] In this respect, particular attention should be paid to teaching users about the interfaces employed by the platforms to identify disinformation and misinformation and to enable them to recognize the alert signs used by the platforms to identify this kind of content. The government might, for example, initiate large-scale education campaigns for social media users, in which these platforms would also be involved.

<u>Recommendation for the Government and the platforms:</u>

1.7 Clear and universally understandable infographics should be developed showing the sequence of steps to be addressed in assessing the quality of information before sharing it.

---

[25] On this subject, see for example the infographics disseminated by the International Federation of Library Associations and Institutions (IFLA): french_-_how_to_spot_fake_news_0.pdf

[26] In its resolution of 21 February 2017 on science and progress in the Republic, the National Assembly called for "scientific education to be the essential medium of growth for enlightened and responsible citizens".

[27] For example, AMCSTI (professional network of scientific, technical and industrial cultures – www.amcsti.fr).

[28] On this subject, see CNCDH, Position on the bill to tackle online hate content, 9 July 2019, p. 8: final_avis_relatif_a_la_ppl_lutte_contre_la_haine_en_ligne.pdf. See also aforementioned CSA recommendation no. 2019-03 of 15 May 2019, no. 6.

Focus points:

1.h The COVID-19 crisis has reinforced the importance of raising awareness and educating citizens on the issue of disinformation and misinformation, which are particularly amplified by the use of digital tools. In consequence, there is a need for an extensive reflection on the topic.

1. i It seems particularly important to develop education campaigns on the use of digital tools, both in schools and universities and in lifelong learning contexts, including for elderly people. This education should help users to better assess the quality of information sources and to manage their role as potential agents of virality.

THE ROLE OF THE AUTHORITIES

While moderating content and controlling virality plays a dominant role in the practical handling of disinformation and misinformation, these procedures raise other ethical questions over the role played by the different authorities. The first question concerns the authority that the platforms have acquired and the oversight that should follow (2.1). The second question concerns the need for these procedures to be overseen by the institutions responsible for identifying those contents that are acceptable or unacceptable to diffuse online. Different questions then emerge over the legitimacy of these institutions, insofar as they are seen by the platforms as helping to establish the truth value of the information (2.2).

A. The authority acquired by the platforms

As the entities primarily responsible for identifying disinformation or misinformation and developing approaches to fight them, social media platforms have acquired great authority over online information sharing. However, their usual moderation practices have been partly modified by the COVID-19 crisis (see above). This raises questions at a number of levels.

The multiple measures employed by the platforms in tackling disinformation spread during the health crisis (e.g., removal of content, visibility reduction and content promotion) raise questions about the future of their role as technical intermediaries. Indeed, these measures significantly challenge the long claimed position by these platforms, which is that – as simple carriers of content without any editorial role – it is not up to them to interfere with what their subscribers (i.e., the users) publish, which means that everyone is free to express their ideas without moderation. However, this argument has been undermined following the wave of terrorist acts in recent years. With the "Christchurch Call" in May 2019, these companies essentially undertook not to disseminate terrorist content on their platforms. This shift has been consolidated by the COVID-19 health crisis. The highlighting and flagging of certain content has reached levels never attained in the past. The reinforcement of these editorial practices could have several consequences. For example, if editorial choices such as the decision to promote official information are not made visible to users, this could cast doubt on the neutrality of the search engines.

Recommendation to the platforms:

2.1 It should be made clear to users that certain content suggestions are the result of editorial choices, which can change in times of crisis.

More generally, the authority that the platforms could potentially exercise in setting their content moderation policy can give rise to a number of ethical tensions.

One view is that every platform should be free to act as it wishes, subject to be compliant with its legal obligations. This might lead them to take different positions, depending with their own economic and political interests, while claiming to be acting as guardians of democracy or of their users 'freedom of expression (cf. the differences in the handling of President Trump's messages by Facebook and Twitter during the current US presidential campaign). With regard to this, it should be recalled that these companies are not defined as media and are therefore not bound by an obligation of pluralism. They also differ from electronic communications operators, which are bound by an obligation of neutrality in the dissemination of content.[29]

Another view is that certain platforms – such as the big social media ones – constitute new digital *agoras*, places of public expression. In addition, these agoras are a locus of interaction for a growing number of "bots" (artificial users), which have great power to persuade through the content they generate, and in some cases create a risk of the mass dissemination of content possibly intended to cause economic and political instability. This might justify a review of the status of these platforms, in particular now that they have become one of the instruments of the diffusion of content impacting public health , as during the COVID-19 crisis.

There are also questions about whether the platforms have the legitimacy to assess the conformance of content to the law, and to be solely responsible for deciding on its removal, given that this entails accepting a form of private law and an increase in censorship, infringing on the freedom of expression. This brings into play the position of judges, whose role in guaranteeing fundamental freedoms must not be sidelined. Yet the reality of this judicial oversight is open to question, both because of the volume and speed of the propagation of disinformation, and because of the limited territorial reach of that oversight in circumstances where most of the platforms are global companies and are not confined to a single national space.

Yet another question concerns the liability of these companies, both in making this kind of editorial choices and in contributing, by default, to the spread of disinformation and misinformation.[30] At present, their liability is limited as they are classified as hosts as defined by the Confidence in the Digital Economy Act 2004-575 of 21 June 2004, which transposed into French law the European Parliament and Council Directive on e-commerce,

[29] On this point, see the annual study by the Council of State on Digital Technology and Fundamental Rights, 2014, p. 217 ff.

[30] On this subject, see the report on "Creating a French framework for social media responsibility".

2000/31/CE of 8 June 2000. In fact, a number of discussions are underway on a possible redefinition of the liability of these social media platforms, both in Europe[31] and the United States.[32] In any case, their responsibility in the case of withdrawal or non-withdrawal of content should be considered with respect to the freedom of expression, a reminder of the Constitutional Council's recent condemnation of the legislation to tackle online hate content.[33]

Focus point:

2.a Ideas should be proposed at national and European level for redefining the responsibility of these platforms with regard to freedom of expression.

Another issue concerns the oversight of these new authorities. Even if this is not actually a new issue, but it might be raised again in the light of the more relevant role of the platforms in handling information during the Covid-19 crisis. Up to now, the European Union has largely favored self-regulation, which is also advocated by the platforms themselves (see the promotion of good practice guidelines),[34] whereas certain countries, like France, have opted for public authority oversight (with respect to disinformation, see Act 2018-1202 of 22 December 2018 on tackling information manipulation). Other countries argue that this oversight should be exercised by authorities that are independent both of the platforms and of the public authorities.[35] The existence of such oversight raises numerous ethical questions. For this reason, the benefits and risks associated with oversight measures need to be assessed, especially with regard to freedom of expression and the limits set on such oversight. In addition, the success and limitations of self-regulation by the platforms in tackling disinformation and misinformation need to be assessed, as well as the question of whether journalistic ethics could form the basis for the ethical practice for social media. New forms of regulation need to be examined, notably the regulation by an independent authority, although the definition of its role could be tricky.

---

31 On this subject, see the ongoing consultation on future legislation on digital services: https://ec.europa.eu/commission/presscorner/detail/fr/ip_20_962

32 https://www.lemonde.fr/pixels/article/2020/05/28/dans-sa-charge-contre-twitter-donald-trump-veut-changer-le-regime-de-responsabilite-des-reseaux-sociaux_6041052_4408996.html

33 Constitutional Council, Ruling 2020-801 DC of 18 June 2020, above.

34 Tackling online disinformation: a European approach, European Commission Communication COM(2018) 236 final, 26 April 2018

35 On this subject, see in particular the above-mentioned report on "Creating a French framework for social media responsibility: acting in France with a European ambition", fourth pillar.

Focus points:

2.b Extensive reflection needs to take place on oversight of digital platforms and in particular on the establishment of a new authority responsible for platform regulation, or on the reinforcement of the role of an existing independent authority responsible for their regulation, such as the CSA, i.e., the Higher Council for Audiovisual and Digital Technology.

2.c Users and the civil society should be able to organize so as to become interlocutors for these social media platforms, with the aim of empowering and giving autonomy to all the actors – citizens, the civil society organizations, companies – alongside the democratic institutions.

B. The authorities on which the platforms depend

In order to monitor and check the content circulating on the Internet, especially at a time of health crisis, social media platforms need to be able to compare it with information coming from sources that are deemed reliable or legitimate: "fact-checkers", government departments, and the public statistical agency that provides the vast majority of information about the pandemic. However, the relations of the platforms with these authorities also raises a number of difficulties.

In order to discriminate between guaranteed information and the other content, platforms can, for example, rely on the work of fact-checkers. Various national actors have emerged in France, such as Décodeurs for *Le Monde*, *Checknews* for *Libération*, *Fake off* for 20 minutes, or AFP's *Factuel*. International organizations have also been formed, such as the European Union backed *European Digital Media Observatory* (EDMO).[36] They are often set up by newspapers, academic institutions or societal organizations, for the purpose of investigating viral content and attempting to assess its truth value. They try to establish the facts or, conversely, to show that the information under examination lacks foundation, and try to identify the sources and networks that usually produce and/or disseminate fake content. However, fact-checkers and their actions are subject to complex tensions too. First, some of these entities are financed by the platforms themselves, therefore undermining their independence. Fact checking is costly, since it entails maintaining large databases and paying teams with the ability to manage and exploit them. Second, these facts-checkers do not have access to all the information circulating on the platforms – in particular, they cannot see information shared within private groups or messaging networks, which leaves a large number of blind spots. In this regard, the European Union is trying to encourage smoother information flows between the platforms and the fact-checkers. Moreover, the authority under which the fact-checkers operate may itself be disputed, with the result that users reject the classification of content as "disinformation" or "misinformation". Indeed, a user may take the view that the desire to hide information,

[36] European Commission Communication "*Tackling Coronavirus Disinformation – Getting the Facts Right*", section 5.2., Support to fact-checkers and researchers, p. 11.

emanating from a disputed authority, is a proof of the authenticity of that information. Finally, many fact-checking bodies are primarily made up of journalists, who may lack the competence to deal with certain types of information or with scientific controversies, in particular during the Covid-19 epidemic.

Recommendations:

2.3 The transfer of information between platforms and fact-checkers should be facilitated.

2.4 The membership of fact-checking teams should be diversified to include researchers and societal representatives.

With regard to the promotion of certain content in particular, there are also questions as to the neutrality of the government when the content it promotes is also the information that legitimizes the government action. Leaving social media platforms and government to talk exclusively to each other raises significant risks of censorship. Arguments that challenge certain political decisions by the government could be unfairly ignored or discarded. The example of "Désinfox information" was a red flag. This webpage was set up by the government, but withdrawn a few hours after a freedom injunction was filed with the Council of State by the National Union of Journalists, which saw it as a "serious attack on pluralism".[37]

Moreover, this exclusivity could have the effect of producing certain ambiguities in the relations between these platforms and the public authorities, in particular the government. Platforms could, for example, have the support of the government to approve or moderate certain information, and the government could ask them to promote certain content related to the management of the health crisis. This would lead to a real risk of complicity, providing the platforms with further influence that could be exploited to oppose any attempts to control their practices in other spheres (for example with regard to informing users about their operating methods).

Recommendation:

2.5 The content moderation mechanisms employed by the platforms during the health crisis should be published, in particular those that guarantee the transparency of the interactions between these companies and the public authorities, and retrospective oversight should be exercised on these mechanisms by the competent authority, in particular judges as guarantors of individual freedoms.

[37] refere-liberte---04-05-2020.pdf. Petition rejected by the Council of State Ruling of 8 June 2020, stating that "the Prime Minister removed this webpage on 5 May 2020, i.e. after the submission of the petition", with the result that "the conclusions of this petition have lost their purpose and there is no further need to rule on it" while ordering the government to pay the costs – see also the speech by Culture Minister Franck Riester, announcing the withdrawal of this webpage: Questions to the Government, 5 May 2020.

Moreover, the current health crisis is peculiar in that it is, to a degree rarely encountered, perceived through large volumes of scientific figures and statistics that are often produced by the public statistical agency (SSP), such as the number of deaths, of people infected, of people treated, respectively in hospitals, in nursing homes and with respect to the population in general. Most of the political measures, speeches and individual views on this epidemic are channeled and underpinned by quantitative tools that rely on definitions and methods that circumscribe their scope, for example reports on the number of deaths, which at best is an approximation of reality. The limitations in the scope of statistical tools are only rarely discussed, whereas these figures are widely reported. The failure to provide a context to these figures may be a source of interpretation that itself contributes to a form of disinformation.

<u>Recommendations:</u>

2.6 The communication of the statistics on the epidemic should be combined with methodological explanations, and discussions about the context and the limitations of such figures.

2.7 Discussions should be published not only on the methods used to produce the statistics, but also on their uses and on the changes that they undergo depending on their use: how they are exploited, communicated and sometimes distorted, how they influence the behavior of the government or the public.

Finally, the use of established authorities to promote scientific content that is presented as reliable should not result in the controversial nature of that content being downplayed.[38] The WHO, for example, seems to be accepted by the platforms as a reliable source of scientific findings, whereas other scientific authorities dispute this, sometimes with good reason.[39] Other players, such as users, scientists or societal organizations could thus be involved in selecting the information to be reported.[40]

[38] On the ethical issues associated with scientific untruths, the post-truth world and scientific communication in the public sphere, see also COMETS Position Paper no. 2018-37, "Quelles nouvelles responsabilités pour les chercheurs à l'heure des débats sur la post-vérité ?" [*What new responsibilities for researchers in the era of post-truth debate?],* published on 12 April 2018.

[39] On this question, see the op-ed published in *Le Monde* and signed by numerous authorities, which calls for the development of "the Forum on information and democracy, set up in November 2019 by eleven organisations, think tanks and research centres in nine countries, to implement the Partnership" between the platforms and social actors ("We call upon the Internet giants to make a decisive shift in favour of the right to reliable information", Le Monde, 02/05/2020. Signed among others by Joseph Stiglitz, Christophe Deloire and Shirin Ebadi).

[40] On this subject, see the previously cited "Creating a French framework for social media responsibility: acting in France with a European ambition".

APPENDICES

Individuals interviewed

- **Serge Abiteboul,** member of the network regulation taskforce (2019) and member of the ARCEP steering group
- **Lucien Castex**, Secretary-General of Internet Society France
- **Guillaume Champeau,** Director of the Ethics and Legal Affairs Department,
- **Leonard Cox,** Vice-President of Public Affairs and CSR,
- **Jean-Claude Ghinozzi,** Chairman and CEO **and Sébastien Ménard**, Strategy Adviser, QWANT
- **Guillaume Goubert,** editor of the newspaper La Croix
- **Béatrice Oeuvrard,** head of public affairs, Facebook
- **Audrey Herblin-Stoop,** head of public affairs, Twitter
- **Jonathan Parienté**, journalist at Le Monde, head of the Décodeurs unit
- **Ramón Ruti**, co-founder and CTO, Storyzy

Members of the working group that contributed to the development of this document

Laurence Devillers	Claude Kirchner
Emmanuel Didier*	Jérôme Perrin
Karine Dognin-Sauze	Catherine Tessier
Christine Froidevaux	Serena Villata*
Eric Germain	Célia Zolynski*
Alexei Grinbaum	
Jeany Jean-Baptiste	**corapporteurs*

Members of the National Pilot Committee for Digital Ethics

Gilles Adda	Christine Froidevaux	Christophe Lazaro
Raja Chatila	Jean-Gabriel Ganascia	Gwendal Le Grand
Theodore Christakis	Eric Germain	Claire Levallois-Barth
Laure Coulombel	Alexei Grinbaum	Caroline Martin
Jean-François Delfraissy	David Gruson	Tristan Nitot
Laurence Devillers	Emmanuel Hirsch	Jérôme Perrin
Karine Dognin-Sauze	Jeany Jean-Baptiste	Catherine Tessier
Gilles Dowek	Claude Kirchner - directeur	Serena Villata
Valeria Faure-Muntian	Augustin Landler	Célia Zolynski

The publication of this bulletin was approved 8 July 2020 at the plenary assembly that included Emmanuel Didier (member of the CCNE – the National consultative committee on bioethics) as a guest observer, with 16 votes in favour, 1 vote against and 2 abstentions.

Press contact: communication@comite-ethique.fr