Hate Speech, Fake News and Content Regulation on Social Networks in Europe

Romain Badouard

Hate Speech, Fake News and Content Regulation on Social Networks in Europe

Romain Badouard

romain.badouard@u-paris2.fr

Centre for Interdisciplinary Research and Analysis of the Media (Carism)

Institut français de presse

Université Paris 2 Panthéon-Assas

La version finale de cet article a été publiée dans l'ouvrage collectif dirigé par A. Monnier, A. Boursier, A. Seoane, *Cyberhate in the Context of Migrations*, Palgrave Publications, 2021.

Migrants are one of the main targets of hate speech on social networks. In France, according to the French agency Netino, which specializes in online content moderation, they constitute the third population category most targeted by aggressive speech, after Muslims and Jews (Netino, 2019). In Europe, more broadly, online campaigns orchestrated by far-right groupuscules focus primarily on the topic of migration (ISD Global, 2018). These networks, which are instrumental in the production of political fake news, also seek to spread a large number of rumours about the migrants arriving in host countries, so as to fuel indignation among their citizens (Badouard, 2020).

Anti-migrant discourse is not just a reformulation of "traditional" racist imaginaries; it also provides valuable indicators to assess the dynamics of "brutalization" of online public debate (Badouard, 2018), and the normalization of hate speech on the Internet. Faced with this trend, several European countries have introduced new legislation: both Germany and France passed new laws to force social network platforms to remove any racist, homophobic or misogynist statements within 24 hours of these being reported to them. Should a platform fail to comply within this timeframe, it could be fined up to 4% of its turnover. In the United Kingdom, the government has planned to create a new public agency to oversee these platforms' moderation policies. Whether in Germany, France or the United Kingdom, these bills face the same criticism*:* they would introduce a form of delegation of censorship power from States to major online platforms. With these new regulations, platforms are taking on new responsibilities, which they often exercise largely behind closed doors.

The issue of hate speech and fake news regulation, however, extends far beyond the relationship between States and platforms, and involves a wide range of actors developing alternative cultural regulation practices: the media have turned to fact-checking and information certification; activists are waging collective counter-speech actions or naming and shaming to block the funding of extremist or manipulative websites; companies are

marketing brand safety solutions; and Internet users themselves are engaging in participatory vigilance initiatives.

This article provides an overview of the issues surrounding the regulation of content on the Internet in Europe, in the particular context of the proliferation of hate speech and fake news. It does so based on the analysis of government reports, laws and interviews with civil society figures involved in the fight against online hate speech. Hate speech and fake news are both problematic contents that have their own specificities. In this paper, we choose to consider them together as they are the targets of similar mechanisms of content regulations that intend to address both disinformation and online hate. We do not discuss the specificities of these contents, which have been under scrutiny by scholars both as far as fake news is concerned (Tandoc, Lim, Ling, 2018) and hate speech (Benesch et al, 2018), but we rather take the angle of solutions that are put on the table by various actors to counter them.

Internet, hate speech and freedom of expression

Since its public debut in the 1990s, the Internet has always been perceived as a communication architecture at the service of freedom of expression. First, its openness allows any connected individual to produce, share and receive information via this network. The advent of the "Web 2.0" in the early 2000s further facilitated the publication of online content and supported the democratization of online speech (Cardon, 2010; Allard, 2005). Second, its decentralized nature complicates censorship by governments and businesses, insofar as there is no single "checkpoint" through which information passes, unlike the mass media (Benkler, 2016). Yet this decentralized structure does not strip governments of all means of control over the content circulating on the Internet. Because they control the points of access to the

international network within their territory, they can filter access to certain websites (China, Russia and Iran, for example, prohibit their citizens from accessing certain US social networks). Moreover, States have a wide range of surveillance and cyber-policing systems on their networks. In this respect, Western democracies are no different from authoritarian countries – something that the Snowden affair in particular brought to light. But the fact remains that within a national network, it is more difficult to filter and prevent the spread of information than in traditional media.

The history of the public Internet is thus marked by controversies over the regulation of the content circulating thereon. In Europe, from the 1990s, various affairs have highlighted the limits of the law's normative power against that of technical intermediaries. In Germany, lawsuits brought against Internet access providers who allowed connection to anti-Semitic websites resulted in technical failures (the T-Online and Compuserve cases in 1996). The incriminated access providers decided to cut their connection to the servers hosting the problematic websites, but in so doing they also deprived German Internet users of thousands of other websites hosted on the same servers. In France, the bigger problem judges are facing relates to servers hosted in other countries, particularly in the United States. A well-known example is the lawsuit filed against Yahoo! in 2000 by the UEJF (*Union des étudiants juifs de France*, the union of France's Jewish students) and the Licra (International League against Racism and Antisemitism), for providing access to auctions of Nazi objects. The French judge in charge of the case initially demanded that the sale be removed, before fining the US website for not respecting his initial decision. When Yahoo! failed to respond, the UEJF turned to the Court of San José in the United States, which ruled that enforcing the French order was incompatible with the First Amendment of the US Constitution. Judge Fogel, in charge of the case in California, responded to the French courts in the following terms: "Although France has the sovereign right to regulate what speech is permissible in France,

this Court may not enforce a foreign order that violates the protections of the United States Constitution by chilling protected speech that occurs simultaneously within our borders"[1].

The Internet connects not only computers and servers, but also geographical areas and cultural systems. The major web services popular with Internet users, such as search engines or social media platforms, are generally US companies with servers located in the United States and therefore subject to US law. However, the cultural and legal tradition on defending freedom of expression is very different in Europe and in the United States. In America, the law places freedom of expression above the rights of individuals, in the name of the general interest, and requires that institutions regulate as little as possible the conditions under which this freedom is exercised (Zoller, 2015). In Europe, conversely, and in France and Germany in particular, numerous restrictions on freedom of expression are recognised by law, and the State readily gets directly involved in the organization of public debate (Girard, 2011). In the case of hate speech, for example, US law distinguishes between "hate speech" and "fighting word" (directly inciting violence), and punishes only the latter. In France and in Germany, on the other hand, making racist remarks in the public space is punishable by law. The Yahoo! case was thus the first in a string of trials in the 2000s, with similar outcomes: since the United States' First Amendment protects freedom of expression, a European judge cannot demand that forms of censorship be applied on US territory.

From the late 2000s, the French courts began to follow in the footsteps of their German counterparts by requiring that Internet service providers (ISPs) block access to certain websites, as provided for in the 2004 *Loi pour la confiance dans l'économie numérique* (LCEN, Law for trust in the digital economy). These measures have however proven ineffective, since an incriminated website need simply be hosted under another name to be accessible again. In 2018, when French ISPs blocked the racist and anti-Semitic website

[1] Lisa Guernsey, "Court Says France Can't Censor Yahoo Site", *The New York Times*, Nov. 9, 2001.

"démocratieparticipative.biz" (following a court decision), this led to a succession of websites being blocked, as the website regularly changed extension (from ".biz" to ".website") to escape filtering. Finally, the courts required that Google stop ranking the website, to make it less visible to Internet users. The website is thus still accessible if its URL is known, but it is no longer visible to Internet users making a simple query on a search engine.

These different cases illustrate a key point about censorship and freedom of expression on the Internet: without the technological intermediaries, the justice system and public authorities have little power to regulate content. On the Internet, it is not the law that has the greatest normative power, but technology: what individuals do on the Internet is the result not so much of what the law does or does not allow, as of what technological tools enable or prevent. US jurist Lawrence Lessig summed this up with his now famous phrase, "code is law" (Lessig, 1999). Thus, "governing" or "regulating" the Internet can never hinge on a single normative power; rather, it relies on a combination of different sources of normativity (law, technology, the market, and practices). Internet governance is said to be "multi-stakeholder", for it depends on collaboration between different types of actors. Regulating online content is a form of "digital governmentality" (Badouard, Mabi, Sire, 2016), where each actor must be certain of the other actors' goodwill in order to hope to be able to "conduct the conducts", to use Michel Foucault's expression (Foucault, 2004). In recent years, the regulation of problematic content, such as hate speech or fake news, has given rise to new forms of public-private collaboration between States and the major web platforms, raising fears of forms of "privatization" of Internet censorship.

Hate speech and Fake news, new content regulation approaches

Web platforms, guided by the ideology of "informational liberalism" (Loveluck, 2015), which considers that the free flow of information is inherently good and should not be constricted in any way, have historically been reluctant to regulate the content they host. Central to the above-mentioned legal debates is the question of the status of technological intermediaries: are blog platforms, search engines and social networks hosts or publishers? In the case of blog platforms, they are exempt from any legal responsibility for the content published via their service, whereas search engines' technical activities to sort, rank and promote information are considered akin to editorial choices (Grimmelmman, 2014). While the courts have never been able to settle the issue, sometimes delivering contradictory opinions, lawmakers have tried to explore new ways to make technological actors accountable, either by creating an intermediary status between host and publisher, or by enacting new ground rules, imposing additional constraints on them with regard to content moderation.

In recent years several European countries, most notably France and Germany, have passed laws to regulate the conditions governing online public speech. This can certainly be related to the changing political and cultural context since the mid-2010s. First, with the terror attacks of the 2010s, the major web companies were accused of having a lax attitude towards jihadist propaganda. The "Russian interference affair"[2] and the Cambridge Analytica scandal[3] subsequently highlighted the way in which targeted marketing on social media could be used for propaganda. Finally, the "brutalization" of online interactions and militant cyber-bullying practices have tarnished the liberal-libertarian "imaginaire" of online debate. In this context,

[2] The "Russian affair" relates to the use of targeted marketing on Facebook by agencies close to the Kremlin to support Donald Trump's campaign.

[3] The Cambridge Analytica scandal relates to the collection of tens of millions of Facebook users' personal data for political propaganda purposes.

conducive to regulation, pressure from States and public opinion has been positively received by platforms, determined to counteract the image deficit caused by these affairs.

In France, two laws passed between 2018 and 2020 have redefined the contours of content regulation on the Internet and social networks. The first, "relating to the fight against the manipulation of information", introduced new summary proceedings, among other things, which allow a judge to require the removal of "false information" online within 48 hours. This unprecedented measure, which is applicable during election periods, is coupled with a "duty for platforms to cooperate" outside of these periods, overseen by the *Conseil supérieur de l'audiovisuel* (CSA), the French agency in charge of regulating radio and television. However, the law does not provide for specific sanctions against operators that refuse to cooperate with the courts. It also raises a number of questions about judges' ability and legitimacy to rule on the veracity of information in such a short time frame (Hochmann, 2018).

The law to combat hate speech, which was directly inspired by the NetzDG law in Germany, generalizes a system for reporting offensive content on the Internet[4] and requires platforms to remove any "manifestly illegal" content reported to them within 24 hours. Should a platform fail to comply, this time the law provides for financial penalties, including fines of up to 4% of the company's global turnover. Here again, the platforms' moderation activities are to be supervised by the CSA.

These new laws are controversial not so much because of the new measures they introduce, but because of the dynamics they induce. With these new ground rules, digital actors have every interest in systematically removing the content reported to them to avoid risking a fine, even if it means practising a form of "over-censorship". Internet users, for their

[4] French law considers as hate speech any statement that calls for violence or discrimination against a group or individual on the grounds of their origin, religion, sex, sexual orientation or disability.

part, know how to take advantage of this censorship "opportunity", which can be used against opponents. Abusive reporting practices are commonplace on social networks: collective "raids" are organized to silence opponents, reporting content to a moderator several times to encourage them to remove it. A study of moderation on Twitter estimated that abusive reports accounted for 47% of all reports, an unquantified share of which was linked to such political "silencing", particularly using bots to generate automatic reports (Matias, 2015). Moreover, platforms' moderation practices are generally shrouded in secrecy: the criteria governing the removal of content are not made public, and until recently, Internet users had no way of appealing against an abusive removal of content.

In order to fulfil their new responsibilities, platforms are also seeking to step up their moderation efforts by significantly increasing their moderation staff. This recruitment itself raises a number of issues: digital labour and the delegation of operations to poor workers in countries of the South, the psychological problems suffered by moderators due to continuous exposure to extremely violent content, and the lack of transparency surrounding the work of these moderators, who are bound by very strict confidentiality clauses[5]. Beyond the filtering operations performed by human beings, platforms are also relying on artificial intelligence to perform moderation tasks. But while the automatic recognition of violent or pornographic images is now highly effective, the equivalent systems in place for speech analysis have much greater margins of error, which compounds the risks of abusive censorship (Gillepsie, 2018). Content subject to interpretation, such as hate speech or fake news, is more difficult to process using automated procedures, since these procedures are unable to adequately take into account the context in which the content is communicated.

Another approach involving these automated mechanisms is what is called "reducing" the visibility of information considered "problematic". Both Facebook and Google have put

[5] As clearly illustrated by the documentary *The Cleaners* by Hans Block and Moritz Riesewieck, 2018, and *The Guardian* newspaper's series of reports "The Facebook Files", 2017.

systems in place to make content less visible on command, and thus to cut off reported content from a potential public. This concealment approach causes the publishing party's audience to drop, while still allowing them to put information online. An investigation by the French newspaper Mediapart recently revealed how this approach was applied to pages associated with the radical left (Delacroix, 2019): some pages had seen their audience shrink by a factor of 1,000 overnight, as their publications no longer showed on their followers' newsfeeds. These new forms of censorship are proving extremely effective: Internet users are not prevented from speaking, but platforms let them "speak in a vacuum", rendering their content invisible to their contacts.

Finally, added to these practices are the possibilities surrounding the personalization of online debate spaces, with platforms' algorithms filtering information according to Internet users' past practices to provide them with content matching their cultural or ideological preferences. The Facebook algorithm, for example, shows users the contacts with whom they are sociologically closest on their newsfeeds, while Google's algorithm factors in the choices already made by a user to provide personalized answers to their requests. This possibility of personalization illustrates how content censorship on the Internet articulates with another controversy related to social networks in recent years: information bubbles, which produce discussion spaces that are adjusted on the scale of individuals and are ideologically homogeneous.

Counter-speech, fact-checking and activism: the emergence of alternative forms of regulation

The regulation of problematic content on the Web and social media has become a societal issue, the management of which extends far beyond the relationship between States and platforms. In the case of the controversy surrounding "fake news", for example, journalists have played a key role in verifying rumours and publishing denials. In France, dedicated teams in major redactions such as *Le Monde* ("Les Décodeurs") or l'AFP ("AFP Factuel") verify most popular posts on social networks in order to determine their accuracy, gaining popularity over the general public. Their "fact-checking" practice can be described as an alternative form of cultural regulation of content. In this specific context, journalists derive their legitimacy from a professional practice: being a journalist means observing a number of ethical principles, such as cross-checking sources and verifying information. Practising "fact-checking" therefore also involves asserting a professional identity on the information market and trying to regain control from alternative sources of information that compete directly with them on social networks, in a context of very strong mistrust of the media.

As a practice, however, "fact-checking" has various limitations. First, the publication of denials and that of fake news seem to follow different informational circuits on social networks and do not reach the same audiences (Benkler, Faris, Roberts, 2018). In other words, the Internet users who share fake news are not those who consume "fact-checking" articles. Moreover, the sociological and ideological profiles of Internet users who share dubious information make them likely to express a very strong mistrust of traditional media (Le Caroff, Foulot, 2019). In this context, the publication of a denial by a "major media channel" will have little impact on their opinion. However, the work of a "fact-checker" journalist is not so much about convincing someone who "believes" fake news as about informing the silent majority (who see the information on a news thread without interacting with it) as to the dubious nature of that information.

Other studies have thus highlighted that “fact-checking” can curb the circulation of a rumour, provided certain conditions are met (Vraga, Bode, 2017). First, the denial must be issued rapidly and provide a consultable source. It also needs to be supported by a community of Internet users who “like” and share it to increase its visibility. Finally, both the false information and its denial must relate to a fresh news topic, about which Internet users do not yet have a formed opinion. The difficulty of meeting all of these conditions seems rather to indicate that in order to effectively fight “fake news”, far-reaching reforms will be required, particularly surrounding the economic model of platforms: limiting “likes”, sharing, or regulating targeted marketing, for example, in order to slow down the spread of information and the “media warming” that it causes (Boullier, 2019).

Given the difficulties of fact-checking on the ground, editorial teams have shifted towards an information certification approach. This is the case, for example, of the *Décodeurs* (decoders) of the French newspaper *Le Monde*, who in 2017 used the Décodex tool to publish an interactive directory of French-speaking information sources, ranking websites according to their reliability. This directory sparked a controversy regarding the legitimacy of *Le Monde* journalists as both judges of and parties to the information market, evaluating their direct competitors on social networks. Meanwhile, the organization Reporters Without Borders is leading a project to introduce an ISO news quality standard to enable newsrooms to obtain official certification. The organization’s goal is also to negotiate with platforms for sources that obtain certification to rank higher in search engine results and be more visible on social media than non-certified sources. This initiative has been criticized for endorsing a principle of two-track access that goes against the fundamental principles of “net neutrality” and equal access to information, thus discriminating between sources considered as “reliable” (but by whom, and how?), and those deemed to be “unreliable”, or that have not been evaluated, and are doomed to be kept at a distance from their potential audience.

Journalists are not the only ones who have engaged in alternative forms of content regulation. In the fight against hate speech, civil society organizations are for example seeking to mobilize Internet users by broadcasting "counter-speech". The strategy here consists in occupying the debate space by contradicting hate speech, so as to raise awareness among the silent majority about the issues at stake in the fight against cyberhate. This counter-speech strategy made waves with the action of the collective #jagärhär ("I'm here") created by a Swedish journalist in 2016. The goal of this collective is to offset the hateful comments found on online discussion threads, particularly on press websites' social media pages. Its members come together in a Facebook group where they report problematic threads and then take joint action to enter into discussion with aggressive Internet users, by opposing their comments with positive and benevolent ones, or simply fact-checking them when the claims made are untrue. The initiative has been very successful in Sweden, reaching over 74,000 members in early 2020. The collective also has several offshoots in Europe and around the world. The French group #JeSuislà, created in January 2019, already had over 5,000 members one year later, and is active on the Facebook pages of French media on a daily basis.

This form of action was theorized and popularized in the early 2010s by US researcher Susan Benesch (Benesch et al., 2018). In her work on what she calls "dangerous speech", that is, discourse which legitimizes and encourages violence against specific population groups, Benesch identified the factors that determine the level of harmfulness of hate speech. She built a theoretical model combining the popularity of the author, the characteristics of the target audience, the cultural context in which the discourse is spread, and the type of medium used. Based on this model, she established response strategies, known as counter-speeches, which aim to counter hate speech by undermining its apparent legitimacy in the eyes of its audiences. Her initiative aimed at civil society organizations, the Dangerous Speech Project,

shares “best practices” with anti-racism organizations interested in the counter-speech approach.

In Europe, the Institute for Strategic Dialogue (ISD), founded in 2006, pursues a similar objective. The London-based think tank has produced both a number of studies on online verbal violence and concrete tools to “coach” civil society organizations involved in counter-speech initiatives. The institute produces toolkits, advises organizations on methods to mobilize online communities, and develops training programmes in collaboration with the major web platforms. In France, the think tank Renaissance Numérique has also been active in the counter-speech space, particularly through the creation of the online platform Seriously. Launched in 2015 following the Charlie Hebdo attacks, Seriously provides a bank of arguments and offers Internet users factual information aimed at countering the racist, sexist or homophobic arguments they may encounter on the web and on social networks.

Similar initiatives can be taken up by social networks and web companies, which may seek to finance them in order to delegate moderation operations to third parties. Both Facebook and Google, for example, are building partnerships with the media and are funding journalists’ posts within editorial offices, which are entrusted with verifying content circulating on the Web and on social networks. These partnerships have been criticized for jeopardizing journalistic independence, in a context where editorial offices are facing an economic crisis. The two companies are also funding counter-speech and media literacy projects, with a view to improving the quality of the content circulated via their services. Internet users themselves are invited to participate in moderation schemes, particularly by reporting fake news and hateful content in circulation. The contribution of Internet users to reporting mechanisms in this way is not without problems, particularly surrounding the abusive reporting practices mentioned earlier. In this context, censorship on social networks

also has a "participatory" dimension, insofar as collaborative moderation mechanisms can be hijacked for political purposes.

Some civil society organizations are adopting a more head-on strategy towards major Web companies, particularly with regard to the regulation of the advertising market. Their mobilisations revolve around the operations of advertising agencies, and in particular Google's agency, which allows websites that publish fake news or hate speech to generate substantial revenue (sometimes several tens of thousands of euros per month) by hosting advertising. Faced with the digital giants' lack of responsiveness on this sensitive matter (which goes to the very heart of their business model), organisations such as Sleeping Giants have launched "name and shame" campaigns on social networks. Their approach has consisted in mobilising Internet users to record all the advertisements that appear on far-right websites and publish screenshots on Twitter, calling out the companies in order to show, if not their collusion, at least their lack of vigilance regarding these sites. These mobilisations have been relatively effective: fearing the "bad buzz" that such a campaign could generate, many companies have become more vigilant about the websites on which they display their advertising. This has even led to the creation of a new marketing niche around "brand safety", which offers companies "clean" communication campaigns on the Internet and social networks, working only with websites identified as "safe".

Conclusion

The new European laws seeking to regulate hate speech and fake news on the Internet entail the risk of abusive censorship on social networks. By putting pressure on platforms to remove problematic content in a very short space of time, European States are supporting the extra-judicialization of censorship (Tréguer, 2019), entrusting public speech regulation

functions, until then the preserve of judges, to private companies. But these new laws also pursue a legitimate goal. If the free expression of opinions is considered to be an inalienable human right, then all citizens should be able to enjoy it equally. Hate speech that seeks to silence (or harm) minority or contradictory voices, to the extent that it reduces access to this right for these sections of the population, can thus be limited. This European "corrective regulation" approach (Girard, 2019) is primarily concerned not with protecting targeted minorities, but with promoting an open and pluralistic approach to public debate.

How, then, can the fight against hate speech be reconciled with freedom of expression? The preferred approach should be to respect the fundamental rights of Internet users, following three principles. First, ensuring the transparency of moderation practices, by making platforms accountable for the publications they remove, the accounts they block and the content they hide. Second, notifying censored Internet users, allowing them to know precisely on which criteria their content was blocked or filtered, and informing them in real time of the implementation of a procedure concerning them, and its evolution and outcome. Finally, allowing Internet users who see their content removed or their account blocked to be able to request a second evaluation of their publications and, if applicable, to have their content re-published. These three principles are at the very core of the Santa Clara Principles[6] established by US academics and digital freedom activists. While this charter is not legally binding, it paves the way towards reconciling freedom of expression and the fight against problematic content, by guaranteeing the fundamental rights of Internet users.

References

[6] https://www.santaclaraprinciples.org/

Allard L (2005) Express Yourself 2.0 ! Blogs, podcasts, fansubbing, mashups...: de quelques agrégats technoculturels à l'âge de l'expressivisme généralisé. In Éric Maigret and Éric Macé (eds), *Penser les médiacultures. Nouvelles pratiques et nouvelles approches de la représentation du monde*, Armand Colin/INA: Paris.

Badouard R (2020, forthcoming) Fausses informations, vraies indignations ? Les « fake news » comme ressources de discussions politiques au quotidien. *RESET* journal, 2020 (10).

Badouard R (2018) Internet et la brutalisation du débat public. *La Vie des Idées.* Available at : https://laviedesidees.fr/Internet-et-la-brutalisation-du-debat-public.html [Accessed 12.11.2020]

Badouard R, Mabi C, Sire G (2016) Beyond 'Points of Control': Logics of Digital Governmentality. *Internet Policy Review,* 5 (3).

Benesch S (2018) *Dangerous Speech. A practical Guide*, Dangerous Speech Project. Available at: https://dangerousspeech.org/guide/ [Accessed 12.11.2020]

Benkler Y (2016) Degrees of Freedom, Dimensions of Power. *Daedalus,* 145: 18-32.

Benkler Y, Faris R, Roberts H (2018) *Network Propaganda. Manipulation, Desinformation and Radicalization in American Politics*, Oxford University Press: Oxford.

Boullier D (2019) Lutter contre le réchauffement médiatique. *Internet Actu*. Available at: http://www.internetactu.net/2019/02/05/lutter-contre-le-rechauffement-mediatique [Accessed 12.11.2020]

Cardon D (2020) *La Démocratie internet. Promesses et limites.* Le Seuil: Paris.

Delacroix G (2019) Facebook anéantit l'audience d'une partie de la gauche radicale. *Mediapart*. 29 August 2019. Available at : https://www.mediapart.fr/journal/france/290819/facebook-aneantit-l-audience-d-une-partie-de-la-gauche-radicale?onglet=full [Accessed 12.11.2020]

Foucault M (2004) *Sécurité, territoire, population. Cours au Collège de France, 1977-1978,* EHESS, Gallimard, Le Seuil: Paris.

Gillespie T (2018) *Custodians of the Internet. Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press: New Haven.

Girard C (2019) *Délibérer entre égaux. Enquête sur l'idéal démocratique.* Editions Vrin: Paris.

Girard C (2011) Le 'libre marché des idées' et la régulation de la communication publique. *Klesis, Revue philosophique,* 21, pp. 218-238.

Grimmelmman J (2014) "Speech Engines", *University of Maryland Legal Studies Research Paper*, n°2104-11.

Hochmann T (2018) Lutter contre les fausses informations : le problème préliminaire de la définition. *Revue des droits et libertés fondamentaux.* N°16.

ISD Global (2018) Hate at the push of a button. Right-Wing troll factories and the ecosystem of coordinated hate campaigns online. Available at: https://www.isdglobal.org/wp-content/uploads/2020/04/Hate-at-the-Push-of-a-Button-ISD.pdf [Accessed 12.11.2020]

Le Caroff C, Foulot M (2019) L'adhésion au 'complotisme' saisie à partir du commentaire sur Facebook. *Questions de communication,* n°35.

Lessig L (1999) *Code and Other Laws of Cyberspace*. Basic Books: New-York.

Loveluck B (2015) *Réseaux, libertés et contrôle. Une généalogie politique d'internet*. Armand Colin: Paris.

Matias J N (2015) Reporting, Reviewing, and Responding to Harassment on Twitter. SSRN Scholarly Paper ID 2602018, *Social Science Research Network*, 2015. Social and Information Networks (cs.SI), arXiv.org > cs > arXiv:1505.03359, Cornell University.

Netino (2019) Panorama de la haine en ligne. Available at: https://netino.fr/panorama-de-la-haine-en-ligne-2019/ [Accessed 12.11.2020]

Tandoc E, Lim Z W, Ling R (2018) "Defining "Fake News": A typology of scholarly definitions", *Digital Journalism*, vol.6, n°2.

Tréguer F (2019) *L'utopie déchue. Une contre-histoire d'internet*. Fayard: Paris.

Vraga E K, Bode L (2017) I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information Communication and Society*. 21(10). pp. 1-17.

Zoller E (2015) La liberté d'expression aux États-Unis : une exception mal comprise. In Muhlmann G (ed.). *La Liberté d'expression.* Dalloz: Paris, pp. 179-224.