"Safe and Well-Informed": A Study of the New Governance of Bio-Citizens' Judgment on Twitter and Facebook

Jean-Gabriel Piguet

What are the consequences of the pandemic on citizenship in liberal democracies? While many academic discussions on this topic focus on the proportionality of the state of emergency and its respect for fundamental political rights, this article analyses another consequence of the pandemic on what the literature has called in recent years digital and informational "bio-citizenship". Since WHO designated "infodemia" as a major public health issue, many online platforms have changed their moderation paradigm of health misinformation. As our review of Twitter and Facebook policies in the first two parts of this study shows, platforms no longer hesitate to treat health misinformation as potential harm to both the quality of online information and their users' health. In the third and last part, we tackle the normative problems that the "somatisation of the bio-citizen" raises. Is it the case, as both platforms argue, that all content that could lead to risky behaviour must be removed? Does their libertarian philosophy require this type of suppression? Should we think with Zuckerberg that misinformation in health is more easily regulated thanks to recognised experts who can provide evidence-based answers? We argue that the platforms' definition of harm creates confusion between error, danger, and rejection of health recommendations, which undermines their claim to guarantee freedom of expression.

1. Introduction: The Promises of Bio-Citizenship

Nearly two decades ago, a wide range of scholars started to observe the development of a new "bio-citizenship" increasingly rooted in patients' demands for better healthcare policies since the 80s. Adopting different nomenclatures like bio-sociality[1] or bio-medicalisation[2], studies of patient activities showed one after the other that groups of patients were developing a new "medical activism",

[1] P. Rabinow, *Artificiality and Enlightenment: From Sociobiology to Biosociality. Essays on the Anthropology of Reason*, Princeton University Press, Princeton 1996.

[2] A. E. Clarke et al., "Biomedicalization Technoscientific Transformation of Health, Illness, and U.S. Biomedicine", in «American Sociological Review», 68 (2), p. 161-194, 2003. https://doi.org/10.2307/1519765.

refusing *mere patients' status* and claiming rights over public acknowledgement of their vulnerability.

Some of the most critical works in this research field led to the conclusion that patients' activism had provoked a genuine revolution in political subjectivity[3]. Rose and Novas described this revolution as a *somatisation* process. In its "individualising moments"[4], people were, according to them, shaping "their relations with themselves in terms of a knowledge of their somatic individuality. Biological images, explanations, values, and judgments thus get entangled with other languages of self-description and other criteria of self-judgment, within a more general contemporary "regime of the self" as a prudent yet enterprising individual, actively shaping his or her life course through acts of choice"[5]. Accordingly, the new regime of the self-implied "quite specialised scientific and medical knowledge of one's condition" and "a range of struggles over individual identities, forms of collectivisation, demands for recognition, access to knowledge, and claims to expertise". "Informational bio-citizenship"[6], as Rose termed it, was now transforming laypeople into activist and experts.

In its "collectivising moments", the somatisation of the self was logically supposed to give rise to "*new forms of democratic participation*, blurring the boundaries between state and society, and between public and private interests"[7]. Full of optimism, Rose wrote in *The Politics of Life Itself*:

> It is creating *new spaces of public dispute* about the minutiae of bodily experiences and their ethical implications. It is generating new objects of contestation, not least those concerning the respective powers and responsibilities of public bodies, private corporations, health providers and insurers, and individuals themselves. It is creating *novel forums for political debate, new questions for democracy,* and *new styles of activism*[8].

Bio-citizenship thus held out a promise of renewed democracy. Far from being a utopian project, bio-citizenship was supposed to develop in an ongoing historical process articulating old forms of citizenship such as "campaigning for better treatment, ending stigma, gaining access to services, and the like (*rights bio-citizenship*)" and "new ways of making citizenship by incorporation into communities linked electronically by email lists and websites, (*informational* and *digital* bio-citizenship)[9]. Above all, it allowed scholars to hope for a collective

[3] N. Rose and C. Novas, "Biological Citizenship", in A. Ong and S. Collier (eds.), *Global Assemblages: Technology, Politics, and Ethics as Anthropological Problems*, Wiley-Blackwell, Malden 2005, p. 439-463; N. Rose, *The Politics of Life Itself: Biomedicine, Power, and Subjectivity in the Twenty-First Century*, Princeton University Press, Princeton 2007.

[4] N. Rose, *The Politics of Life Itself*, cit. p. 135.

[5] N. Rose, *The Politics of Life Itself*, cit. p. 134.

[6] N. Rose, *The Politics of Life Itself*, cit. p. 134.

[7] D. Heath et al., "Genetic Citizenship", in D. Nugent and J. Vincent (eds.), *Companion to the Handbook of Political Anthropology*, Blackwell, London 2004, p. 152-167, here p. 152.

[8] N. Rose, *The Politics of Life Itself*, cit. p. 134-135.

[9] N. Rose, *The Politics of Life Itself*, cit. p. 134-135.

awareness of fragility as the universal and permanent condition of advanced societies and as a resource for guiding their public health policies.

Almost two decades later, what is left of the hope of revitalising liberal democracies through bio-citizenship in a time of pandemic? Unfortunately, the pandemic that began in winter 2020 may cause two significant disillusions.

The first disillusion stems from the state of emergency. Most of the liberal democracies have been compelled to temporarily limit the practice of citizenship to ensure citizens' health, arguing that it is a condition for the enjoyment of all their rights. Most governments' unpreparedness forced them to take measures in a hurry, thus silencing *de facto* any genuine deliberation in the public square and parliaments on this topic. The brutal, local, and short-term lockdown measures stand in marked contrast with the repeated calls since ten years from the World Health Organisation (WHO) to national policy-makers to anticipate pandemic's structural risk with long-terms and international measures[10]. Western democracies have succeeded neither in anticipating long-term health challenges and remaining fully aware of their vulnerability nor keeping health policies into the realm of parliamentary and citizens' deliberation. The "cycle of panic and neglect" that characterises public health investments during a crisis could be seen as a manifestation of the inability of liberal democracies to hear bio-citizen's lessons in the long-term, which definitely invites to temper the optimistic tone of Rose, Novas, and Health.

There is, however, another disappointment caused by the pandemic that might give rise to a more profound concern about bio-citizenship. For this reason, we would like to focus our attention on it in this article. Indeed, the crisis not only called into question the universalisation of vulnerability as an actual historical process but also the very possibility of "informational" and "digital" citizenship in a social network era. In a speech delivered on February 15 2020, the Director-General of the WHO famously warned about an "infodemic" that, according to him, "spreads faster and more easily than this virus, and is just as dangerous". From the outset of the pandemic, the infodemic created a new kind of public health problem, compromising populations' adherence to official recommendations, promoting false therapies for COVID-19, if not outright denying its existence. Director's speech ended with a solemn "call on all governments, companies and news organisations to work with the WHO to sound the appropriate level of alarm (…)", and concluded that "now more than ever is the time for us to let science and evidence lead policy"[11].

Undoubtedly, the hope for online bio-citizens led by evidence-based beliefs was shattered long before the pandemic. It is a well-known fact that citizens now mostly inform themselves on internet platforms and that platform algorithms favour the

[10] World Health Organisation, "What is a Pandemic?", «Diseases», February 24, 2010. Available at: https://www.who.int/csr/disease/swineflu/frequently_asked_questions/pandemic/en/.

[11] T.A. Ghebreyesus, *Munich Security Conference*. February 15, 2020.

spread of sensational information to the detriment of scientific content. More generally, a climate of mistrust towards the scientific and media elites has been developing for many years, transforming the "anti-medical" attitude and "claim of expertise"[12] that Rose observed in 2005 into beliefs in "alternative science".

According to WHO General-Director, the "search and media companies" are the right organisations to blame for this situation. The WHO succeeded in putting pressures on leading platforms and social networks. Right from the start of the pandemic, the latter announced a series of new measures to guarantee "the health of public conversation" and a "safe and informed" user experience.

In the following lines, we would like to assess the consequences of these measures on digital bio-citizenship practices. Among the measures announced, social networks explained that they would begin to consider misinformation about health as harm. This implied, among other things, that for the first time since the problem of misinformation appeared in the public square in 2016-2017, the platforms would allow themselves directly to censor misinformation in order to provide better information and prevent the risk behaviours that misinformation inevitably leads to.

Such a paradigm shift necessarily affects the expression of millions of citizens seeking information about the pandemic, discussing the measures adopted or not, and challenging them if necessary. Therefore, the bio-citizen is concerned in two ways by these measures: as a body to be protected from the virus COVID-19 and as a mind to be protected from the virus of misinformation. Above all, it affects not only their access to information but also their understanding of their duties on these platforms. In this respect, they redefine what it means to be a citizen, transforming them into individuals who have not only the right but also the duty to remain "safe" and "well-informed" so as not to contaminate others, and who are deprived of the right to access the platform if the disobeys.

A genuine evaluation of these measures' impact cannot only consist of a factual assessment. It must first test their normative coherence and unveil the foundations of the concept of citizenship they sketch out. Do these measures succeed in defining what are the rights and duties of citizens online when it comes to health misinformation involving a high collective risk? As J. Waldron states it, commenting on Kant's definition, citizens in constitutional states are "framers and law-givers: they are conceived to have made the state for themselves rather than to be merely the subjects of authoritarian imposition"[13]. Is it possible to consider the new paradigm of online moderation as an improvement of citizens' autonomy? Are they *sine qua non* for the re-enchantment of bio-citizenship?

There are three sets of questions that need to be addressed in assessing this new paradigm of moderation. The first aims to clarify the aims of moderation (why). What is its first purpose? Is it to weigh the benefits of freedom of expression

[12] N. Rose, *The Politics of Life Itself*, cit. p. 134.

[13] J. Waldron, "Citizenship and Dignity", in «NYU School of Law, Public Law Research Paper» 12 (74), 2013. http://dx.doi.org/10.2139/ssrn.2196079.

against health risks or improve critical judgement by giving people more accurate information? The second seeks to determine the appropriate means to pursue the chosen end (how). Can censorship of false content ever be legitimate? Should the term "censorship" be reserved for the deletion of content, or should it include, on the contrary, the degradation of the visibility of certain content? The third raises the problem of the identity of the moderator (who). Can a private company exercise it? Does this mean that a private company has a duty to uphold freedom of expression when most constitutions apply this rule to the state only?

In this article, rather than addressing these theoretical issues abstractly, we will study how two major platforms, Twitter and Facebook, attempted to address these challenges before (1) and during the pandemic and how their moderation measures affected the practice of digital bio-citizenship (2). In the final part of our study, we will challenge the platforms' leaders' claim that this paradigm shift is a mere application of the "liberal" rule that censorship is only legitimate when there is a proven risk of direct and imminent harm (3).

2. Remove, Reduce, Inform – the Moderation Paradigm of Misinformation Before the Pandemic

As noted above, misinformation on digital platforms is not a problem that first arose during the pandemic, but in the aftermath of the 2016 US elections. Although it initially had little to do with public health issues, it became one them since 2019. Many media outlets, the WHO and the European Union called for more significant moderation of health misinformation long before the pandemic. When the pandemic broke out, therefore, both platforms already had a policy on the issue.

In this first part of our study, after defining platform moderation (1) and restating its political and health context before the pandemic (2), we will show how the "remove, reduce, inform" strategy came to be applied to health misinformation (3).

2.1 Platforms as Moderation Services

The problem of misinformation is all the more severe as moderation is what best defines them. Indeed, we can follow Gillespie and Sanders who refer to platforms as online sites and services that "host, organise, and circulate user's shared content or social interactions for them" without producing the contents themselves[14]. In other words, "online platforms emerged to simplify the process of navigating the abundance of information available in the digital public sphere. (…) Moderation is, in many ways", *the* commodity that platforms offer" (Sanders 2020, 945).

[14] T. Gillepsie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*, Yale University Press, 2018.

In practice, platform moderation can be operationalised through two kinds of private governance rules, dictated either by corporate philosophy, public pressure, regulatory compliance, or profit maximisation[15]. Some of these rules are *implicit* and amount to forms of "architectural regulation"[16]. They are expressed in the code and algorithms that influence the types of contents distributed on the platform, "as well as how content is organised, promoted and presented to users". Other rules are *explicit*, "such as those documented in public-facing platform community standards and terms of service (...) – the interpretation and enforcement of which contribute to a body of *platform law*"[17].

The *explicit* rules are put in force either before or after the publication. When moderation occurs after content has already been published, it relies on a combination of community and automated flagging to detect potentially impermissible content:

> Ex post review may be conducted automatically by software and/or manually by human moderators. (…) Facebook automatically removes posts where the tool's confidence level indicates that its decision will be more accurate than human reviewers. For all other posts, the score system enables Facebook's team of human moderators to prioritise reviewing content that receives the highest scores[18].

2.2 From Political to Health Misinformation

Information has not always been regulated on platforms through official standards and rules. Long before the pandemic, social networks were singled out as the prominent architects of two events that most academics and traditional media considered dramatic, namely Brexit and Donald Trump's election. Although it is difficult to establish a direct causal link between misinformation and these two events, they drew attention to social networks' responsibility for misinformation. Indeed, the Pew Research Center showed a few months after Trump's election that 44% of Americans got their news from Facebook[19]. Another MIT study established later that false information had a 70% greater chance of being retweeted than real information[20]. Many voices were thereupon raised against the "filtering bubbles"

[15] B. Sanders, "Freedom of Expression in the Age of Online Platforms", in «Fordham Inter-National Law Journal» 43 (4), p. 939-1006, 2020. https://ir.lawnet.fordham.edu/ilj/vol43/iss4/3.

[16] T. Gillepsie, *Custodians of the Internet*, cit., p. 179.

[17] T. Gillepsie, *Custodians of the Internet*, cit., p. 179.

[18] B. Sanders, "Freedom of Expression in the Age of Online Platforms", cit., p. 947.

[19] Pew Research Center, "News Use Across Social Media Platforms", 2016. Available at: https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.

[20] S. Vosoughi et al., "The Spread of True and False News Online", in «Science», 359 (6380), p. 1146-1151, 2018. Doi: 10.1126/science.aap9559.

of social networks, as Pariser had called them[21], which prevented citizens from being confronted with opinions they do not share[22]. Above all, the very objective of information dissemination through social network algorithms (to retain users as quickly and as long as possible with the most emotional content that requires the least intellectual effort) began to be considered by many scholars as almost incompatible with the survival of the public space.

Therefore, it was not only the passivity of the platforms that was indicted but their active contribution to the distribution of misinformation. In response to these accusations, while refusing to be considered as anything other than digital hosts, Facebook and Twitter rush into a massive communication plan to show both their determination to combat misinformation actively *and* their absolute commitment to freedom of expression.

First, a Facebook spokesperson in April 2017 explained that Facebook could not "become an arbiter of truth" itself, because it was "not feasible given Facebook's scale", and not its "role"[23]. These comments echo social network standards that state they want "people to stay informed without stifling productive public debate", and that Facebook is aware that there is "only one step between misinformation and satire or personal opinions"[24]. Similarly, shortly after the 2016 presidential election, Twitter CEO Jack Dorsey, creates an internal group dedicated to "Trust and safety"[25]. He thus tries to show the group's involvement in the fight against misinformation while continuing to present Twitter as "the free-speech wing of the free-speech party"[26], a self-description that dates back to 2012, when Twitter's UK managing director, Tony Wang, sought to justify that Twitter takes a "neutral" view of the messages posted by its users.

Following these announcements, the platforms' algorithmic and internal rules were modified in 2016-2017 to mitigate the spread of misinformation, as our next section will show. Two years later, new circumstances prompted another move in the same direction. In 2019, many observers began to warn against the growth of

[21] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Press, London 2011.

[22] D. R. Grimes, "Echo Chambers Are Dangerous", in «The Guardian», December 4, 2017. Available at: https://www.theguardian.com/science/blog/2017/dec/04/echo-chambers-are-dangerous-we-must-try-to-break-free-of-our-online-bubbles; A. Hess, "How to Escape Your Political Bubble for a Clearer View", in «*New York Times*», March 3, 2017. Available at: https://www.nytimes.com/2017/03/03/arts/the-battle-over-your-political-bubble.html.

[23] A. Mosseri, "Working to Stop Misinformation and False News", «News Feed», April 7, 2017. Available at: https://about.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news.

[24] Facebook, "False News (21)", in «Community Standards». Available at: https://www.facebook.com/communitystandards/false_news.

[25] Twitter Blog, "Strengthening our Trust and Safety Council", in «Blog», December 13, 2019. Available at: https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-trust-and-safety-council.html.

[26] J. Halliday, "Twitter's Tony Wang: 'We Are the Free Speech Wing of the Free Speech Party'", in «The Guardian», March 22, 2012. Available at: https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.

vaccine hesitancy and the growing appeal of conspiracy theories. Many Western countries feared that vaccination coverage would fall below the rates recommended by the WHO[27]. For example, the number of requests for vaccine exemptions increased in 2017-2018 for the third consecutive school year in the United States. Once again, the responsibility of social networks was pointed out, as vehicles for misinformation[28]. For example, one of the pioneering studies on Polish-language social media showed that 40% of the most shared Twitter links concerning health between 2012 and 2017 could be qualified as misinformation[29]. In April 2019, when two Washington Post journalists typed "cure for cancer" into the search bar on YouTube (with no research history), the 6th video offered, viewed 1.4 million times, claimed to eliminate cancer risks with baking soda[30].

By the same token, misinformation became a public health issue, as WHO announced[31]. Media pressure forced Facebook to take more robust and specific measures against health misinformation in February 2019, after two articles published in *The Washington Post* and *The Wall Street Journal* highlighted the existence of numerous groups dedicated to conspiracy theories linked to health issues. Twitter soon followed suit. On May 19 2019, Twitter thus declared it was now "committed to protecting the health of the public conversation on Twitter" and considered that "ensuring individuals can find information from authoritative sources" was "a key part of that mission"[32].

2.3 Remove, Reduce, Inform

After making such announcements, the platforms had to show how they intended to give users access to the best possible information. As this section will show, the two waves of regulations (2016-2017 and 2019-early 2020) that followed these announcements belonged to a shared paradigm. This paradigm could be summarised in three terms: remove, reduce, inform[33].

[27] World Health Organisation, "Vaccines and Immunization", in «Health Topics», 2019. Available at: https://www.who.int/health-topics/vaccines-and-immunization#tab=tab_1.

[28] W. Y. S Chou et al., "Addressing Health-Related Misinformation on Social Media", in «*JAMA*», 320 (23), p. 2417-2418, 2018. Doi: 10.1001/jama.2018.16865; J. Farell et al., "Evidence-Based Strategies to Combat Scientific Misinformation", in «Nature Climate Change», 9, p. 191-195, 2019. https://doi.org/10.1038/s41558-018-0368-6.

[29] W. Y. S Chou et al., "Addressing Health-Related Misinformation on Social Media", cit.

[30] A. Ohlheiser, "They Turn to Facebook and YouTube to Find a Cure for Cancer", in «The Washington Post», June 25, 2019.

[31] World Health Organisation, "Ten Threats to Global Health in 2019", in «Spotlight», 2019. Available at: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019.

[32] Twitter Blog, "Helping you Find Reliable Public Health Information on Twitter", in «Blog», May 10, 2019. Available at: https://blog.twitter.com/en_us/topics/company/2019/helping-you-find-reliable-public-health-information-on-twitter.html.

[33] Facebook, "False News (21)", cit.

Remove. The guiding principle of Twitter's and Facebook's moderation policy was that the platforms should *only* remove content that involves some form of simple manipulation[34], or could result in imminent physical harm[35]. Such policy included anti-spam measures to prevent the repeated sending of an electronic message, often advertising, to many Internet users without their consent[36]. It included the deletion as well of "inauthentic commitments, which attempt to make accounts or content appear more popular than they actually are, and coordinated actions, which attempt to artificially influence conversations by using multiple accounts, fake accounts, automation and/or scripts"[37].

Both platforms gradually added restrictions on advertising, such as a ban on advertisers using keyword targeting sensitive categories, including[38], a ban on public media[39], or a ban on advertising deemed political[40].

The removal of most of harmful content was made easier by the fact that it did not involve any factual verification. It should be noted that, in this paradigm, even when platforms had to carry out fact-checking (to delete wrong declarations about official elections for example) the reason for deletions remained the potential harm created[41].

If Facebook and Twitter categorically refused to remove content just because it was wrong, why did both platforms promote the removal of certain content as a way to combat the spread of misinformation? The reason is simple: massive misinformation was (and still is) often made by spam or trolling or false accounts. Destroying trolls and fake accounts was a way to destroying the very source of misinformation for platforms without setting themselves up as arbiters of the truth.

[34] Facebook, "Manipulated Media (22)", in «Community Standards». Available at: https://m.facebook.com/communitystandards/manipulated_med; Twitter Rules and Policies, "Manipulated Media", in «Rules and Policies». Available at: https://help.twitter.com/fr/rules-and-policies/manipulated-media.

[35] Facebook, "Credible Violence (1)", in «Community Standards». Available at: https://www.facebook.com/communitystandards/credible_violence; Twitter Rules and Policies, "Hateful Conduct Policy", in «Rules and Policies». Available at: https://help.twitter.com/fr/rules-and-policies/hateful-conduct-policy.

[36] Twitter Rules and Policies, Platform Manipulation", in «Rules and Policies». Available at: https://help.twitter.com/fr/rules-and-policies/platform-manipulation.

[37] Facebook, "Spams", in «Community Standards». Available at: https://www.facebook.com/communitystandards/spam.

[38] Twitter Ads Policies, "Keyword Targeting", in «Ads Policies», 2014. Available at: https://business.twitter.com/fr/help/ads-policies/campaign-considerations/policies-for-keyword-targeting.html; Facebook, "Controversial Content", in «Ads Policies». Available at: https://fr-fr.facebook.com/policies/ads/prohibited_content/controversial_content#.

[39] Twitter Ads Policies, "State Media", in «Ads Policies», 2014. Available at: https://business.twitter.com/fr/help/ads-policies/ads-content-policies/state-media.html.

[40] Twitter Ads Policies, "Political Content", in «Ads Policies», 2014. Available at: https://business.twitter.com/fr/help/ads-policies/ads-content-policies/political-content.html; Facebook, "What's Facebook's Strategy for Stopping False News?", in «Hard Questions», May 23, 2018. Available at: https://about.fb.com/news/2018/05/hard-questions-false-news/.

[41] Twitter Rules and Policies, "Platform Manipulation", in «Rules and Policies». Available at: https://help.twitter.com/fr/rules-and-policies/platform-manipulation.

Reduce. What about false content that did not cause any harm to anyone? It would be wrong to believe that moderation boils down to a binary choice between authorisation and prohibition. To keep the balance between quality information and freedom of expression, platforms prefered to make undesirable but harmless content gradually invisible. Facebook in particular took action against entire websites that repeatedly shared false news, reducing their overall News Feed distribution , and lowered their ranking in keyword search results. Since the targeted contents were less seen and shared, without the user necessarily being informed[42], the virality of undesirable content that did not violate Facebook's rules was reduced, like misinformation.

The reduce strategy implied a much tighter fact-checking than the remove strategy. On Facebook, fact-checking could be carried out by the user community, either directly by way of signposting or indirectly through user's mass refusal to share a post, or via an internal fact-checking process.

Professional internal fact-checkers had several rating options (false, modified, partially false, missing context, satire, true). They were not to engage with opinions that belonged to the realm of the unverifiable, such as value judgments or speculations. To help these fact-checkers, Facebook partnered with the International Fact-Checking Network, thanks to which it was able in July 2019 to "reduce the access to "posts with health-related claims that are exaggerated and glamorous"[43].

Twitter did not pursue this approach immediately, and the difference between the two platforms became soon apparent. In 2019, a study published by Stanford researchers showed that the number of likes, shares or comments of false content on Facebook had fallen by 50% since the 2016 US elections, whereas it had remained the same over the period on Twitter[44]. Hence, Twitter partnered with the *US Department of Health Services* and enforced a series of measures focused explicitly on health misinformation. Notably, Twitter launched "a new tool so when someone searches for certain keywords associated with vaccines, a prompt will direct individuals to a credible public health resource", hoping "to expand it to other important public health issues in the coming months". Additionally, it would stop auto-suggesting "queries that are likely to direct individuals to non-credible commentary and information about vaccines"[45]. While the effectiveness of these measures has never been evaluated, they demonstrate at least Twitter's willingness to slow down the flow of misinformation on health issues.

[42] G. Delacroix, "Facebook anéantit l'audience d'une partie de la gauche radicale", in «Médiapart», August 29, 2019.

[43] Facebook, "Adressing Sensational Health Claims", in «News», July 2, 2019. Available at: https://about.fb.com/news/2019/07/addressing-sensational-health-claims.

[44] H. Alcott et al., "Trends in the Diffusion of Misinformation on Social Media", in «Research and Politics», 1 (8), 2019. https://doi.org/10.1177/2053168019848554.

[45] E. Birnbaum, "Twitter Launches Tool to Combat Vaccine Misinformation", in «The Hill», May 5, 2019.

Inform. The final part of Twitter and Facebook's moderation strategy consists of informing the user of the origin of the message he or she is about to read[46] that it has been evaluated negatively by fact-checkers and, if necessary, redirecting him to sources considered more reliable. On health issues, Twitter goes so far as to offer "Ads for Good" to non-profit organisations so that they can set up campaigns aimed at "disseminating reliable health information to as many people as possible"[47].

3. The Fight Against the "Infodemic" as a Shift in Paradigm

On the eve of the pandemic, misinformation had already become a public health problem for the WHO and platforms, compelling the latest to strengthen the policies that were essentially aimed at political misinformation. From 2019 onwards, Twitter and Facebook started to target advertisements for fake medicines and unfounded claims about human health.

However, the foundations of the moderation paradigm had remained unchanged since 2016-2017. It was still based on the "remove, reduce, inform" triptych described above. In the absence of legal guidance and constitutional review, the legislative role naturally fell to Facebook and Twitter, which began to create new user's rights and duties in an utterly unilateral manner: misinformation could be withdrawn only when it leads to imminent danger; when this is not the case, misinformation is made less visible and confronted to best sources of information[48]. Hence, the platforms had taken a stance close to the first amendment of the US constitution: the fact that content is not accurate is *not* sufficient grounds for censorship[49].

However, at the beginning of the pandemic, this paradigm was confronted with its failure to ensure "the health of public conversation". The WHO therefore urged them again to enforce more robust measures at the end of February 2020. Soon after, Twitter and Facebook announced a range of new measures. While many of these consisted of strengthening the "reduce and inform" strategy, one point indicated a global change of philosophy: from now on, health misinformation could be *removed*.

Such a radical paradigm shift requires an explanation. Why did the pandemic change the apparently liberal and well-framed paradigm, that, according to

[46] Twitter Blog, "Nouveaux Labels pour les comptes gouvernementaux", in «Blog», August 6, 2020. Available at: https://blog.twitter.com/fr_fr/topics/company/2020/nouveaux_labels_pour_les_comptes_gouvernementaux_et_les_medias_d_Etat.html.

[47] Twitter Blog, "Our Actions to Protect Public Conversation", March 05, 2020. Available at: https://blog.twitter.com/fr_fr/topics/company/2020/notre-travail-pour-proteger-la-conversation-publique-autour-de-c.html.

[48] R. Badouard, *Les Nouvelles Lois du Web*, Seuil, Paris 2020.

[49] T. I. Emerson, "Toward a General Theory of the First Amendment", in «Yale Law Journal», 72 (5), p. 877-956, 1963.

platforms, secured citizen's freedom of speech? Above all, how could they pretend that the new paradigm was a mere application of the precedent?

In the second part of our study, we will detail the reasons that made the broadening of the concept of harm necessary in the eyes of the platform's leaders to meet the challenges of the pandemic (2.1) and why they should imply a lasting shift in paradigm (2.2).

3.1 Two Reasons for the Paradigm Shift

As outlined below, the liberal moderation paradigm adopted by Twitter and Facebook before the pandemic implied that the fact that content was wrong, even intentionally, was never deemed a sufficient reason to remove it. Thus, until the pandemic, misinformation about health was not removed, except when incorporated into advertising.

The new context of the pandemic pushed platforms to change their policy on this point and censor misleading information to fight against the "infodemic". At Facebook CEO Mark Zuckerberg's press call of March 1 Judge Brender asked, "why is it so easy for Facebook to do so much now against false news and why is it so difficult in a political context to remove misleading political information", and whether the pandemic was a test" for platforms to change their "position on political misinformation in the future"[50].

Zuckerberg mentioned two reasons to explain this change, and why it did not involve anything like global censorship against political misinformation:

> Well, I think that this is a different case for two reasons. One is we've always had our policy that doesn't allow content that's going to cause imminent danger or physical risk. (…) We've never allowed things that would lead to imminent physical risk and as I've mentioned in my opening remarks, even countries like the United States that have the strongest traditions on free expression, the standard here is you can't yell fire in a crowded theatre, basic idea. It's that you want to allow a wide range of expression, but you don't want to support things that are going to lead to imminent physical harm work for yourself or others.
>
> So in the case of a pandemic like this where we're seeing hoaxes that are basically encouraging people who are sick to not get treatment or to not act in ways that are going to protect the people around them or in some ways to do things that could be actively harmful, right, so I mean, there's one hoax going around that if you think you have this, drink bleach and that will cure it.
>
> And that's terrible that's obviously going to lead to imminent harm if you do that. That is just in a completely different class of content than some of the kind of back-and-forth accusations that a candidate might make about another, for example, during an election[51].

[50] M. Zuckerberg, *Facebook Press Call*, March 18, 2020.

[51] M. Zuckerberg, *Facebook Press Call*, cit.

It seemed indeed impossible to consider misinformation about health like any other type of misinformation. On the one hand, its truth can have direct consequences on the health of users. This consideration had already justified banning health advertisements (Twitter) or subjecting them to fact-checking as a sine qua non for publication (Facebook). The logic implied that content that was not financially motivated should be controlled in the same way. On the other hand, in the context of a pandemic, false beliefs produce risky behaviour that is dangerous for everyone, not only for those who hold such beliefs.

Therefore, the quarantine of contagious discourse seemed to be the right thing to do. But for this to be consistent with platforms' policies, the notion of harm had first to be broadened, i.e. associated with misinformation. Twitter is the most explicit on this point. After announcing greater use of automation, it states that it will "broaden its definition of harm to act on content that directly contradicts advice from authoritative global and local public health information sources"[52].

The second reason mentioned in Zuckerberg's press conference is formulated as follows:

> The other piece here is that there are broadly trusted authorities who people across – including governance, who people are just across society would all agree can arbitrate which claims are conspiracy theories or hoaxes and what's trustworthy and what's not, which makes this a very different dynamic than trying to be referee of political speech.
>
> I mean and the WHO for example, or CDC, just do have broad trust and a government mandate (…) So I think that that's – but you're basically asking about things that I think are on opposite ends of the spectrum in terms of difficulty of operationalising the response.
>
> I think for health misinformation that's probably – during a pandemic or outbreak like this, that's probably one of the most black and white situations that you could expect. And I think political speech is probably the most difficult in terms of how you arbitrate and kind of govern that kind of speech[53].

Thus, the second reason for social networks to resort to censorship is their belief that "evidence-based" solutions can address health issues. Sound health policies are only the practical transcription of scientific data. In this sense, there is no such thing as *political* public health; there are only self-evident measures.

It should be noted that neither Facebook nor Twitter bans content solely because it is considered to be erroneous by their internal fact-checking services. Facebook censures "posts that make false claims about cures, treatments, the availability of essential services or the location and severity of the outbreak or "claims that physical distancing doesn't help prevent the spread of the coronavirus", but only reduces the visibility of "claims that don't directly result in physical harm, like conspiracy theories about the origin of the virus"[54].

[52] Twitter Rules and Policies, "Misinformation Policy", in «Rules and Policies». Available at: https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy.

[53] M. Zuckerberg, *Facebook Press Call*, cit.

[54] M. Zuckerberg, *Facebook Press Call*, cit.

3.2 A Lasting Change

Several remarks must be made to appreciate the scope of this change.

First of all, we must stress that this change will not disappear with the crisis. Social networks have indeed tried to legitimise their paradigm shift through new circumstances, but they never claimed that these measures were temporary. The reasons they have given to broaden the scope of moderation may apply to any health misinformation.

Second, it must be emphasised that this is indeed a change. Indeed, platforms claim that they are merely applying to new circumstances the same liberal rule that harm to others is the only possible ground for censorship. However, two elements oppose this communication strategy.

First, even if one concedes that the same criterion prevailed before the pandemic, it must be noted that it was not applied in the same way to the issue of vaccine hesitancy in 2019, even though it had many points in common with the pandemic: it was already a global threat to public health, where misinformation endangers even those who do not adhere to it. Hence, something has changed in the moderation paradigm.

Second, the broadening of the notion of harm upsets misinformation and fact-checking on two levels. On the first level, now, Twitter, for example, includes as misinformation "statements which are intended to influence others to violate recommended COVID-19 related guidance from global or local health authorities to decrease someone's likelihood of exposure to COVID-19, such as: "social distancing is not effective", or "now that it's summertime, you don't need a mask anymore, so don't wear your mask!"[55]. Therefore, the fight against misinformation is no longer just a question of checking facts but distinguishing false content that is more or less dangerous and supporting populations' adherence to global and local health authorities' guidelines. The distinction between an erroneous and dangerous fact on the one hand, and an appeal to disobey health recommendations on the other, is completely erased.

On the second level, Twitter includes in the category of misinformation "misleading claims that unharmful but ineffective methods are cures or absolute treatments for COVID-19, such as "Coronavirus is vulnerable to UV radiation – walking outside in bright sunlight will prevent COVID-19"[56]. Therefore, in the previous paradigm, fact-checkers could differentiate between false content and "not sufficiently proven" content. From now on, "not sufficiently proven" becomes synonymous with "misinformation".

[55] Twitter Blog, "Definition Covid-19", in «Blog», April 1, 2020. Available at: https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#definition.

[56] Twitter Blog, "Definition Covid-19", cit.

4. A Liberal Paradigm?

As we outlined, the strategy to curb the spread of health misinformation during the pandemic divides into three branches: remove, reduce, inform. In every case, social networks claim to apply a liberal rule. On the one hand, reducing the visibility of content ensures the fundamental right to quality information, and does not imply any speech freedom limitation, according to the platforms. On the other hand, the removal of content obviously limits freedom of speech, but for a legitimate reason. "Even in the most free-expression, friendly traditions like the United States, you've long had the precedent that you don't allow people to yell fire in a and that – I think it's similar to people spreading dangerous misinformation in the time of an outbreak like this" explained Zuckerberg.

Many critics have already pointed out aspects of platform moderation policy little consistent with this liberal self-image. Some have highlighted the opacity of algorithms and fact-checking criteria[57], while others have denounced the absence of any governmental control[58]. Crucial and legitimate as they are, most of these critics do not discuss the fundamental justification for the paradigm shift during the pandemic and its claim to be part of the consensual "American liberal tradition". In this last section, we will try to shed some light on this blind spot.

4.1 The Reduce Strategy as a Hidden Censorship

In order to distance themselves from the unflattering image of a ministry of truth, Facebook and Twitters use a rather dubious distinction between simple reduction of false but harmless content content's visibility, and genuine censorship, reserved for cases of potential harm. Contrary to what this distinction suggests, reducing the visibility of content through an invisible algorithm is a form of soft and hidden censorship insofar as it aims at precisely the same goal: preventing the reading and sharing of content by people who would otherwise have done so.

More broadly, this hypocrisy reflects an overall hesitation. If harmless misinformation does not violate any rule, why do the platforms fight against it? The platforms' generic answer is that they are committed to "ensuring the quality of the online discussion". This answer seems unsatisfactory for two reasons. First, it does not specify how the platforms could defend the "quality of the discussion" online without setting themselves up as arbiters of the truth, even if they collaborate with external and allegedly independent fact-checker. Secondly, this response does not clarify on what grounds platforms should be concerned about the quality of the discussion if its low quality does not violate any user rights, and if social networks are truly only hosts and not publishers.

[57] R. Badouard, *Les Nouvelles Lois du Web*, cit.

[58] D. Kaye, *European Union Draft Directive on Copyright in the Digital Single Market*. Available at: https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-OTH-41-2018.pdf.

Consequently, the ambition of reducing fake content distribution cannot be considered part of the "liberal conception" of speech freedom invoked to give it a non-threatening appearance. Either misinformation provokes harm for other users' rights, in which case open censorship is needed and platforms have to act as "arbiters of truth", or misinformation is not harmful, and it is hard to see what makes the reduction strategy legitimate.

Therefore, not distributing false and harmless content amounts to doing too little or too much against misinformation; too much, if misinformation does not imply any harm, and too little if it does.

4.2 From Libertarian-Liberalism to Liberal Paternalism? Two Objections

Nevertheless, rather than condemning the reduction strategy, should we not say that social networks should stop referring to the libertarian tradition to justify it? Aren't there any good reasons for limiting freedom of speech when it seems so obvious to prevent an essential part of the population from accessing expertise on public interest issues?

Let us suppose for a moment that moderation applies transparent criteria, that the user has a genuine right of appeal, and that the State effectively controls the decisions of the platforms and their applications. Let us also assume that the platforms only censor statements that are very consensual and refuted by scientific authorities. Is a lesser distribution of harmless but false content still prohibited in principle, as part the libertarian tradition would claim? As Girard points out, many classical doctrines defend freedom of speech as a *mean* for collective emancipation and enlightened opinion[59]. In that case, we must ask ourselves not only whether free expression physically harms others, but whether it is actually contributing to form an enlightened opinion. Financially motivated algorithms have failed on this point, as we outlined. Therefore, would it not be the best defence of citizens' freedoms to encourage health content distribution that experts do not take down?

We must admit that a genuine debate on freedom of expression cannot be limited to Zuckerberg's vague defence of "American" freedom of expression, which is more aimed at defending the company's image in the United States than defining a sound doctrine. Another liberalism is undoubtedly *possible*, and it has been referred to in recent years as "liberal paternalism"- a paternalism that does not aim to decide on behalf of the citizen but guide him or her towards good choices through incentives rather than prohibitions. From this point of view, the reduction strategy chosen by the platforms could appear to be equivalent to the "nudging" strategies that many public institutions already use and which, for example, make the healthiest products more accessible without banning the others in cafeterias[60].

[59] C. Girard, "Pourquoi a-t-on le droit d'offenser?", in «La vie des idées», 8 December, 2020, p 7. Available at: https://laviedesidees.fr/Pourquoi-a-t-on-le-droit-d-offenser.html.

[60] C. Sunstein and R. Thaler, "Libertarian Paternalism Is Not an Oxymoron", in «University of Chicago Law Review», 70, p. 1159-1202, 2003. https://doi.org/10.2307/1600573.

However, two arguments must be raised against silencing voices that propose "alternative facts", even if one considers the value of freedom of expression to be conditional. Firstly, acknowledging that institutional health experts are trustworthy is one thing, but considering them as the only source of knowledge and giving them a right of censure is quite another. Can we exclude *as a matter of principle* that ordinary citizens may legitimately testify, for example, about the side effects of a vaccine, even if their testimony is wrongly formulated as general knowledge about the dangerousness of the vaccine; or that an independent expert might be right against the WHO?

Second, such a policy deprives users of the possibility to judge for themselves who is a trustworthy scientific authority, which seems contrary to their dignity as autonomous citizens and voters. This argument relies on a general principle: the idea of self-government that defines democratic citizenship cannot do without the idea of self-government of thought. This principle, in turn, leads us to distinguish between a legitimate means to serve the "quality of online discussion" understood as democratic deliberation, namely, the re-information of citizens and the redirection towards content validated by the experts, and an illegitimate means, the will to decide on behalf of the citizens which factual narrative is trustworthy.

4.3 Error, Danger, and Public Health Recommendations

As for the removal of content that is likely to cause imminent physical harm, it would be difficult to deny that it legitimates the immediate suppression of a potentially monitored tweet by millions of people, such as the link to a video in which US President Donald Trump explains that bleach could help cure covid[61].

However, the specific definition of a health risk provided by the two platforms raises two problems. First of all, this definition bonds danger with error. At first glance, the platforms seem well aware that these two terms do not equate, in so far as they reserve deletion for the most dangerous content, thus recognising that misinformation is not a source of imminent physical danger in its own right. However, they establish a somewhat mysterious link between the two terms by including the removal of harmful content in the description of the general policy against misinformation and by removing "misleading claims that unharmful but ineffective methods are cures or absolute treatments for COVID-19, such as "coronavirus is vulnerable to UV radiation – walking outside in bright sunlight will prevent COVID-19"[62]. Indeed, if health protection is the sole ambition that legitimates removal, this policy should be entirely dissociated from that of the fight against misinformation. Staying "safe" and staying "informed" are not two identical objectives. Moreover, the distinctions between being "known to be ineffective", "not known to be effective" and dangerous have not always been

[61] BBC News, "Coronavirus: Outcry After Trump Suggests Injecting Disinfectant as Treatment", April 24, 2020.

[62] Twitter Blog, "Definition Covid-19", cit.

evident in practice. The first category has been misused to delete many tweets articles defending hydroxychloroquine as early as the end of March 2020[63] without clearly established alternatives and at a time when the US Food and Drug Administration had issued an emergency use authorisation for chloroquine and hydroxychloroquine to treat patients hospitalised with COVID-19[64]. At the time, chloroquine was not known to be effective, but it was not known to be ineffective or dangerous. Generally speaking, it is difficult to understand what authorises platforms to remove harmless contents if what ultimately justifies removing them is the physical harm it could cause, particularly when there is no internationally recognised alternative to the promoted treatment.

Second, the definition of harm and risk proposed by the platforms creates an even more dubious equivalence between non-compliance with health instructions, dangerousness, and misinformation. Indeed, both Twitter and Facebook state in the general description of their anti-misinformation policy to remove content where it calls for distrust of health instructions[65].

This equation is based on three false assumptions. First, it assumes that public health recommendations always follow the expert's voice and cannot spread dangerous misinformation. Following such a rule would, for example, lead to the rapid elimination of the account of the experts who criticised a French health minister, who, in the midst of a face mask shortage, tried to convince French citizens that it was not only utterly useless to wear a mask when one respected social distances but dangerous when one did not know how to put it on[66].

Second, refusing to maintain content that in one way or another invites people to disobey health instructions, even when there is no manipulation or misinformation implied, while allowing "political discussions about the pandemics", supposes that it would be possible to strongly distinguish legitimate political debates on health recommendations and calls to disobedience and risky behaviours. Nevertheless, the difference between questioning health policies and calling for disobedience is not strong enough to draw the line between authorised and forbidden content. If a citizen is free to question instructions openly, he or she must be free to draw the practical consequences that flow from this opinion and expose himself to government sanctions. One cannot allow the expression of the premises of a speech and ban its practical conclusions.

[63] S. Pixako, "Big Tech Thought the Pandemic Wouldn't Be Political", in «Inter Press Service News Agency», May 28, 2020. Available at: http://ipsnews.net/business/2020/05/28/big-tech-thought-the-pandemic-wouldnt-be-political-think-again.

[64] R. Sandler, "FDA Authorizes Anti-Malarial Drugs Chloroquine and Hydroxychloroquine for Emergency Coronavirus Treatment", in «Forbes», March 30, 2020. Available at: https://www.forbes.com/sites/rachelsandler/2020/03/30/fda-approves-anti-malarial-drugs-chloroquine-and-hydroxychloroquine-for-emergency-coronavirus-treatment.

[65] Facebook, "Combating Covid-19 Misinformation", in «News», March 25, 2020. https://about.fb.com/news/2020/03/combating-covid-19-misinformation.

[66] R. Prizac, "Pénurie de masques : chronique d'un mensonge", in «*l'Humanité*», 8 April, 2020. Available at: https://www.humanite.fr/penurie-de-masques-chronique-dun-mensonge-687538.

In this respect, this rule is the practical translation of what Zuckerberg indicated as his second reason to combat health misinformation more aggressively than political misinformation, namely the firm belief that WHO's recommendations and opinions are not political but only "white or black"[67] and scientific. He seems to consider disobedience to expert recommendations as a mere negation of the facts and an unconscious risky behaviour. This presupposition ignores that a call to disobey health instructions and, for example, reopen restaurants is not a false assertion that could be debunked by expert reasoning, but the prescriptive part of a value judgment. Moreover, it is hard to see in which sense a call to disobedience could be characterised as "misleading assertion".

5. Conclusion

The main objective of this study was to highlight the consequences for biocitizens of the change in moderation philosophy that occurred during the pandemic on two platforms, Facebook and Twitter.

Before the pandemics, both platforms prohibited only misleading and dangerous content. During the pandemics, they started to delete and make invisible unproven claims that were, at the time, defensible and not deemed dangerous by many health authorities. Moreover, they redefined health misinformation as dangerous content at the outset of the pandemic, but in doing so, they no longer had any way of distinguishing misleading content from conscious and non-misleading calls for disobedience. They practically started to treat the latter category in the same way as calls for murder in the previous paradigm, arguing that such messages were harmful to others since they increased the risk of contagion for all.

The main question raised by this paradigm shift is not whether a genuine biocitizenship implies that the state should better control the platform's governance. A broad consensus seems to emerge on this issue. Instead, the question it asks is whether governments should regulate platforms in order to enlighten the biocitizen's opinion by protecting him or her from misinformation, to maximise the protection of his or her body by protecting it from beliefs dangerous to him and to others, or to enable him to express more freely. The platforms have chosen to give priority to the second objective, which they believe justifies new censorship, operated invisible censorship designed to achieve the first and claimed to guarantee the third.

In that respect, this paradigm shift illustrates the paradoxical development Rose and Nova called the "somatisation of the individual". In this case, we should describe it more precisely as the somatisation of the online citizen and, more specifically, the somatisation of his judgment. Indeed, it was both the concern for the vulnerability of his mind and body that legitimised the quarantine of his speech from the outset.

[67] M. Zuckerberg, *Facebook Press Call*, cit.

This security turning point raises the following normative problem: is it legitimate to prohibit an expression as soon as it calls to insubordination towards official health recommendations? Such an ambition seems inconsistent with the ideal of self-government that defines democratic citizenship. Taking down content for the sole reason that it could lead to risky behaviours raises the most critical question brought forth by this pandemic: to what extent can health take precedence over all other legitimate goods? Just as one cannot treat a call to disregard health recommendations to break the loneliness of the elderly as a call for direct violence, neither can one treat in the same way a misleading assertion that creates immediate danger, such as Trump's video suggesting that ingesting bleach could help cure Covid, and the defence of a harmless and "proven to be inefficient" remedy. To be legitimate, censorship can only remove scientifically indefensible and dangerous content. It excludes public health recommendations as a basis for censorship, which by definition are injunctions addressed to autonomous persons, and are not conclusions that would simply stem from scientific knowledge. Whether one favours risk minimisation or defends that prudence does not boil down to risk avoidance, one thing seems clear: this discussion is too critical for bio-citizens to be left to the algorithmic and regulatory governance of platforms.

Jean-Gabriel Piguet
Haute Ecole Spécialisée de Suisse Occidentale Valais-Wallis
(HES-SO Valais-Wallis), Suisse.
jean-gabriel.piguet@hevs.ch

Jean-Gabriel Piguet (M) holds a double master's degree in history of philosophy and political science (Sorbonne-Sciences Po Paris), was an assistant and scientific collaborator in moral and political philosophy (Institut Catholique de Paris, University of Fribourg) from 2015 to 2018. Since then, he has been in charge of the applied ethics department of the University of Applied Sciences of Western Switzerland (Valais). His current research focuses on data ethics, neo-liberalism and digital governance. In parallel, he is completing a PhD in political philosophy on the prudence and limits of individual rights in the liberal tradition (University of Poitiers, France).