Facebook’s policies against extremism: Ten years of struggle for more transparency
by Catherine Bouko, Pieter Van Ostaeyen, and Pierre Voué

Abstract
For years, social media, including Facebook, have been criticized for lacking transparency in their community standards, especially in terms of extremist content. Yet, moderation is not an easy task, especially when extreme-right actors use content strategies that shift the Overton window (i.e., the range of ideas acceptable in public discourse) rightward. In a self-proclaimed search of more transparency, Facebook created its Transparency Center in May 2021. It also has regularly updated its community standards, and Facebook Oversight Board has reviewed these standards based on concrete cases, published since January 2021. In this paper, we highlight how some longstanding issues regarding Facebook’s lack of transparency still remain unaddressed in Facebook’s 2021 community standards, mainly in terms of the visual ‘representation’ of and endorsement from dangerous organizations and individuals. Furthermore, we also reveal how the Board’s no-access to Facebook’s in-house rules exemplifies how the longstanding discrepancy between the public and the confidential levels of Facebook policies remains a current issue that might turn the Board’s work into a mere PR effort. In seeming to take as many steps toward shielding some information as it has toward exposing others to the sunshine, Facebook’s efforts might turn out to be transparency theater.

Contents

Introduction
Sanitized extreme-right (social media) content
Facebook’s broad content reviewing policies and international standards on freedom of expression
Materials and method
Results
Discussion and conclusion

Introduction

Even if Facebook’s content-reviewing techniques have become increasingly more sophisticated, violating content is still particularly present. In 2017, Facebook took action on 1.6 million pieces of content that breached its community standards. The platform identified a quarter of it itself; the rest was removed after Facebook users’ flagging them. By contrast, in December 2020, 97.1 percent of the 26.9 million of content...
actioned was identified by Facebook itself (Facebook, 2021b). Among this violating content, extreme-right content is still prevalent. For example, the NGO Avaaz (2019) reported over 500 far-right and anti-EU pages and groups that spread false and hateful content to Facebook, which were followed by nearly 32 million people, before the EU elections in May 2019.

Social media moderation, in other words trying to strike the right balance between removing extremist content and censorship, is certainly not an easy task. In blurring the lines between acceptable and unacceptable content, extreme-right creativity particularly challenges regulations. In the second section of this paper, we argue how sanitized extreme-right content shifts the Overton Window (i.e., the range of ideas acceptable in public discourse) rightward through mainstream elements of pop culture and cloaking strategies, both framed as funny and humorous practices. Yet, despite this challenging context, we review in the third section of this paper how Facebook’s community standards are still excessively broad in some areas, despite targeted recommendations in international standards and NGOs’ reports for the past 10 years. Facebook’s transparency dilemma can be explained, in part, by a two-fold tension. On the one hand, there’s the multiple and consistent PR and some concrete efforts that profess Facebook to be an actor deeply involved in promoting societal well-being. On the other hand, there’s the actual values of Facebook (embodied in rules, procedures, and goals) of moderation that may well be more specifically targeted to ensure Facebook’s continued financial expansion. While we understand that institutions have internal actors pulling in different directions, it is well known that the values of upper management create symbolic and material ends of reality within institutions. We examine such tensions in the sections below.

Facebook has taken some measures, though. In the third section of this paper, we outline that in addition to its Community Standards Enforcement Report, the social platform established the Oversight Board, who started their work in October 2020. Their role is to qualitatively analyze the implementation of Facebook’s community standards in specific cases.

In this context of longstanding transparency issues in terms of Facebook’s community standards, on the one hand, and recent initiatives for more transparency on the other, we sought to investigate to what extent specific elements of the 2021 Facebook’s community standards related to moderating extremist content are still vague or unclear (RQ1), and therefore, how they may still lead to inconsistent implementation. To provide an answer to this research question, we analyzed 644 text-image posts written by extreme-right actors (see the fourth section of this paper). This approach made it possible to bring to the surface potential issues in Facebook’s community standards in an empirical way. Our findings reveal two major issues related to the vagueness of Facebook’s community standards, namely the issue of ‘representing’ dangerous organizations and the issue of ‘endorsement’ from potentially dangerous individuals. Furthermore, these two results raise the more general issue of designating organizations or individuals as ‘dangerous’. In a second phase, we analyzed how Facebook’s Oversight Board positions itself in relation to the longstanding issues regarding transparency in its first reviews, published in January 2021 (RQ2). Furthermore, analyzing the Board’s first reports allowed us to identify in detail how Facebook collaborates with the Board and, more precisely, to the kind of documents which the members of the Board have access to in their reviewing process (RQ3). As we will argue in the last part of this paper, the Board’s no-access to the in-house rules demonstrates how the longstanding discrepancy between the public and confidential levels of Facebook policies remains accurate and might turn the Board’s work into a mere PR effort, as some experts fear.

Sanitized extreme-right (social media) content

An increasingly shift of the Overton window rightward

Right-wing online extremism is now mainstream: extremist actors have shifted the Overton window [1] (i.e., the range of ideas acceptable in public discourse) rightward, so that xenophobic, nationalistic, or exclusionary discourses which were not tolerated decades ago have now become more normalized
Facebook’s policies against extremism: Ten years of struggle for more transparency

(Conway, 2020; Reynolds, 2018). The traditional separation within the right-wing spectrum between the radical or the extreme right has become obsolete with the emergence of populist political movements and individuals that distance themselves from the traditional extreme right through “sanitized” rhetorics (May and Feldman, 2019), a metaphor that emphasizes the cleansing strategy in the mainstreaming process (Ahmed and Pisoiu, 2019; Atton, 2006; Fielitz and Thurston, 2019; Guenther, et al., 2020; Maly, 2019). This has direct consequences for the moderation practices on social media, insofar as the platforms wish to avoid their actions in content moderation being considered as reflecting a political bias against the right more broadly (Conway, 2020).

The boundaries between what is considered as acceptable or not are not static (Fielitz and Thurston, 2019); extreme-right sanitization operates like “a sort of slow-acting poison, accumulating here and there, word by word, so that eventually it becomes harder and less natural for even the good-hearted members of society to play their part in maintaining this public good.” [2]. The extreme-right actors’ creativity, whether professional or amateur, is infinite when it comes to creating content they share, especially on social networks. A first type of mainstreaming techniques that allow to shift the Overton window rightward is the use of popular culture. The extreme right is now a site of cultural engagement which has appropriated visual styles and symbols of contemporary (youth) cultures (e.g., Forchtner and Kølvraa, 2017; Miller-Idriss, 2018; Pisoiu and Lang, 2015; Schedler, 2014; Schlembach, 2013). This cultural mainstreaming shift can serve as a potential entry point to radicalization:

By employing humorous ambiguity, ‘hipsterish’ aesthetics or references to popular culture, particularly cartoons and video games, this more subtle, not overtly political imagery may offer access points for undecided and not-yet politicized users to develop affinities with, and support for, far-right causes. [3]

**Cloaking strategies to circumvent community standards**

Cloaking strategies are a second major type of mainstreaming techniques. In using them, extremists aim to circumvent bans on explicit extremist content by using coded language. At the same time, this process can allow them to sanitize explicit content. Like with the inclusion of popular culture, coded language can reinforce the mainstreaming process by rendering these cloaking practices funny: cloaking has become a game. Through humor, fun and entertainment, sanitized extreme-right content in coded language is granted the illusion of harmless fun and becomes normalized. Of course, such cloaking strategies did not emerge with social media. For example, the American radio personality Rush Limbaugh, who hosted a very popular American conservative radio show (from 1984 until his death in 2021), was famous for his polarizing and controversial statements under a veneer of humor.

Coded language can concern specific terms, like “skypes” in order to refer to Jews (Magu and Luo, 2018), but visual symbols are also an extremists’ playing field. Furthermore, symbols are powerful “bonding icons” (Wignell, et al., 2017) which strengthen a sense of belonging among the insiders, who know the coded language.

Drawing the line between what is forbidden cloaked content and what is not is an issue in both the online and off-line worlds. Miller-Idriss (2018) examined how extremist clothing brands manage to exploit the legal grey areas in German right-wing movements which allow them to engage in doublespeak and invoke a certain “ideological innocence” [4]: the message is overtly vague and covertly implicit enough to circumvent hate speech regulations and claim it as harmless, but is, at the same time, clear enough to be understood as extremist ideas. Cloaking strategies are composed of a large variety of creative techniques: play with right-wing symbols (alphanumeric codes, stylistic references to the Nazi era, etc.), play with co-opted visuals or words (attribution of extreme-right meanings to non-right-wing items, such as to Pepe the frog), bricolages of new narratives out of old, and traveling images from one national background to others as well as through translations. Social networks amplify this phenomenon of playing with words and symbols. The Kekistani flag, used by the alt right and emerged out of playful disruptive dynamism on social
media, particularly captures this phenomenon, insofar as it takes up the graphic codes of the Nazi flag while at the same time disassociating them from their fixed historical meaning (Tuters, 2019).

Developments in certain German court decisions show to what extent drawing the line between the acceptable and the unacceptable is proving difficult not just for social media platforms. Extreme-right creativity sometimes takes advantage of the German legal bans, which are limited to symbols that have a direct association with banned organizations. The concept of association can be subject to various interpretations, in a context where some argue for the right to create new symbols out of Nazi symbols. These lose their direct reference but nevertheless remain evocative. Through implicit evocation, they no longer denote but connote meaning. Denotation designates concrete people, things, and places; connotation conveys explicit denoting, especially concrete ideas, in addition to implicit denoting (Barthes, 1977). Connotation can convey culturally shared meanings through established, or even stereotyped, representations which are “made-up elements of signification.” But connotation is also vital in creating new ideas and meanings (van Leeuwen, 2005). Instead of conveying well-established meanings, experiential connotation creates new ones by innovatively associating elements together. This “semiotic import” is key in experiential connotation. Cloaking strategies are often built on established or more experiential connotation. For example, Thor Steinar’s brand logo combined two runic symbols (used by the Nazis) that together evoke a swastika. While firstly banned by German courts, the legal ban of this logo was later overturned based on the argument that this logo is a new symbol that cannot be considered as a reference to the Nazi organization, even if it was created out of two banned runic symbols (Miller-Idriss, 2018). Such implicit meanings in new visual symbols based on banned ones allow their users to mainstream their ideas, engage in doublespeak and claim ideological innocence. For social media platforms, elaborating community standards to moderate such content remains particularly challenging, as we will see in the next section.

Facebook’s policies against extremism: Ten years of struggle for more transparency

From facilitators of open expression to key actors under pressure to combat extremism

It is not easy to moderate content, let alone extremist content; the mainstreaming of extreme-right ideologies and content described in the previous section makes digital extremist speech and visual representation even more difficult to pin down. Following the rhetoric of the democratizing power of the Internet, social media companies have, relatively speaking, protected themselves from this issue when they frame themselves as facilitators of open and egalitarian expression (Gillespie, 2018). Avoiding regulation was explicitly framed as an actual value: in 2016, Mark Zuckerberg argued that Facebook was not a “media company,” preferring it to be seen as a mere platform and a technology company in order not to be saddled with the legal obligations incumbent on media companies (Napoli and Caplan, 2017). Luckily for social media corporations, they are not considered as media companies, at least not yet: social media companies are private companies that are not required to extend First Amendment protection to the content that their users share on their platforms (Alkiviadou, 2019). This exemption gives social media tremendous power to influence the character of public discourse and of their users’ expression by drawing the line between what is and is not acceptable on their platforms. In doing so, they position themselves as “the new Governors of online speech,” in a “new triadic model of speech that sits between the state and speakers-publishers.” Facebook’s policies, which also concern Instagram, are only inspired by the First Amendment of the U.S. Constitution, which protects freedom of expression, including hate speech, as long as it is not based on fighting words, incitement to violence or true threats (Carlson and Rousselle, 2020).

However, this exemption might soon be a thing of the past: currently, five bills in the U.S. Congress, many supported by both Republicans and Democrats, seek to impose regulation and a new definition (such as that of a public utility) on these behemoths. Therefore, the legal definition of these digital behemoths as private companies may potentially change. Other proposed legislation involves a breakup of these virtual...
monopolies. In Europe, some laws already impose intermediary liability for social media platforms: for example, Germany’s Network Enforcement Act (NetzDG) requires them to take down any “manifestly unlawful” content within 24 hours [9]. The European Commission is currently preparing binding rules in its Digital Services Act.

Open self-expression, connection and sharing are Facebook’s key professed values: Facebook frames itself as “an open platform for all ideas, a place where we want to encourage self-expression, connection and sharing. At the same time, when people come to Facebook, we always want them to feel welcome and safe. That’s why we have rules against bullying, harassing and threatening someone. [...] A post that calls all people of a certain race ‘violent animals’ or describes people of a certain sexual orientation as ‘disgusting’ can feel very personal and, depending on someone’s experiences, could even feel dangerous”, according to Facebook VP EMEA Public Policy Richard Allan (2017). His text dates from 2017 but is still referred to by Facebook in 2021. This statement encapsulates how Facebook approaches hate speech based on the user’s sensitivity, from a personal and interindividual perspective, inspired by U.S. law. Ultimately, Facebook’s policies depend on a business model that seeks to give their customers/users a positive personal experience. They need to feel “welcome and safe” in order to engage on the platform — and generate profit. As Gillespie (2018) points out, the values behind the social media’s policies are not mostly some moral core (like freedom of speech, human rights, or fight against disinformation), but the best compromise possible between competing pressures (i.e., making a profit, reassuring advertisers and complying with international laws). Emotional content and heuristics, in particular, trigger automatic and emotive reactivity over deliberate thinking and in doing so, generate engagement and traffic on social platforms. Being built on these communication strategies, populist and extremist contents, in particular, generate engagement, and therefore profit. Consequently, these contents are involuntarily at the heart of the business model of social networks, to the extent that campaigns like the Stop Hate For Profit campaign invited companies to suspend their commercial activities with Facebook in July 2020, to ask to stop monetizing hate speech and disinformation. Over time, social networks have been increasingly urged to respond to mass pressure to combat violence, hate speech and misinformation, not least in the context of Trump’s incitement to Capitol violence on 6 January 2021. A major and recurrent criticism voiced by researchers, international institutions (like the United Nations), and civil society actors, concerns social media’s lack of transparency concerning its moderation policies. Nearly 10 years ago, Gillespie (2012) already pointed out Facebook’s opacity, due at that time, according to him, to Facebook’s wish to avoid drawing attention to the presence of objectionable content on its platform. Since then, it has remained a recurrent bone of contention, despite recommendations in international standards on freedom of expression. Article 19 (2018), a British NGO, listed a dozen international standards that have pleaded for transparency for a decade. As general context, the International Covenant on Civil and Political Rights (ICCPR) adopted by the United Nations, which considers that “any law or regulation must be formulated with sufficient precision to enable individuals to reasonably foresee whether content they generate or transmit is likely to be edited, removed or otherwise affected” (United Nations. Special Rapporteur on Freedom of Opinion and Expression, 2016). And even more concretely, the Joint Declaration on ‘Fake News’, Disinformation and Propaganda recommended in 2017 that “intermediaries should take effective measures to ensure that their users can both easily access and understand any policies and practices, including terms of service [and] detailed information about how they are enforced, where relevant by making available clear, concise and easy to understand summaries of or explanatory guides to those policies and practices.” [12].
more transparency and more ‘explanatory guides’ have not been addressed yet, though. The NGO Article 19 (2018) analyzed April 2018 Facebook community standards regarding hate speech and freedom of expression. Although the 2018 version was more detailed than the previous one, Facebook still continued to fall below the international standards described above (Article 19, 2018). For Carlson and Rousselle (2020), too, Facebook’s 2018 policy update had little impact on the removal process. That said, let’s also acknowledge that Facebook’s community standards are updated regularly and sometimes follow international recommendations. For example, in January 2021, they added a definition of ‘attacks’ in their definition of hate speech, which is framed as a direct attack against people, based on their protected characteristics (Facebook, 2021a). Yet, such minor changes keep most longstanding issues unaddressed.

That said, legal responses to content moderation are also sometimes considered as being too broad. Germany’s Network Enforcement Act (NetzDG) has been criticized for being vague and over-inclusive. Furthermore, it has already been instrumentalized by 10 authoritarian states to justify restricting online speech. While explicitly referring to German law, they use vague definitions of categories of fake news, defamation of religions, anti-government propaganda, extremism and hate speech, which can be exploited to target political dissent (Mchangama and Fiss, 2019).

**Discrepancies between global theoretical policies and (localized) in-house rules**

Caplan (2018) distinguishes between three models of content moderation: artisanal, community-reliant and industrial. Companies like Facebook started by managing content moderation in an artisanal way. Given the amount of content to moderate, they then moved to an industrial model. In artisanal settings, the same people are often responsible for developing the policies and enforce them. One legal counsel who Caplan interviewed compared the artisanal model to a “common law system” [13]. The common law is a legal system used a.o. in the U.S. in which the rules are primarily made by the courts following individual decisions. Therefore, case law is the main source of law and the rule of precedent requires judges to follow previous decisions made by the courts. It comes as no surprise, therefore, that being a U.S.-based company, Facebook was not inspired by the codified rules that prevail in most European legal systems (notably under the influence of the Napoleonic code). According to a former Facebook employee, content moderation rather started with a one-page list of informal rules and some workers sitting together [14]. However, this artisanal way to inductively develop and enforce policies is not compatible with the need for specific rules that allow replicable enforcement by automation and/or moderators who have to make split-second decisions, mostly far removed from the context of the speech. In this context, artisanal flexibility made way for industrial replicability. According to the same former Facebook employee, the moderation system resembled not so much a courtroom as it did a “decision factory [...] trying to take a complex thing, and break it into extremely small parts, so that you can routinize doing it over, and over, and over again” [15]. Moderation is performed through the implementation of the broad policies into a multitude of concrete in-house rules that remain confidential. In 2017, journalists of the British newspaper *The Guardian* examined more than 100 Facebook’s internal training manuals, spreadsheets, and flowcharts using thousands of slides and pictures (Hopkins, 2017). The same year, leaked documents revealed how broad notions in the policies are sometimes implemented in odd or controversial ways in Facebook’s internal rules (Angwin and Grassegger, 2017). Among Article 19’s 16 recommendations, Article 19 pointed out how Facebook’s community standards remained overly broad, notably due to the use of vague terms (e.g., praise, support, glorification, or promotion of terrorism), vague definitions of terrorist and hate speech organizations, and the lack of case studies and examples in which Facebook applies its standards in practice.

Broad corporate policies combined with hidden internal rules leave significant discretion to Facebook in their implementation. As Caplan reminds us, “content policies are not law; they’re policy” [16]. In line with Caplan, Gillespie (2018) points out how these platforms take advantage of the leeway up for grabs in private policies, framed as “statements of both policy and principle,” in applying their rules when they are helpful or in sidestepping them when they are too constraining. This latitude makes for a lack of transparency regarding how the material is assessed or the way moderators make their decisions to remove materials or not, for example when profiles’ marked violations of several policies remain (Ben-David and Matamoros Fernández, 2016; Caplan, 2018; Carlson and Rousselle, 2020; Gillespie, 2018; Kreiss and
Mcgregor, 2019; Reeve, 2019). This highlights the tension between broad policies that leave leeway to the contextualized enforcement of in-house rules on the one hand, and working conditions that prompt replicable decisions, on the other.

Consequently, broad policies are likely to lead to inconsistent application (e.g., Angwin and Grassegger, 2017; Reeve, 2019). While some overt content was systematically removed after they reported it (e.g., “n-word”), Carlson and Rousselle (2020) observed that most other types of content were removed in inconsistent ways. However, it is crucial to analyze Facebook’s decisions on the basis of the most precise typologies possible. For example, content such as “Islam is cancer” was probably not removed due to inconsistent decisions as Carlson and Rousselle claim, but because this statement is a stance on a religion, which is allowed, and not hate speech towards a specific religious group. Nevertheless, we agree with Carlson and Rousselle when they note how such analysis of Facebook’s moderation decision-making can only be ambiguous, in so far as it is only based on the broad standards, without knowing the internal rules that allow to flesh them out.

Lastly, Facebook’s lack of transparency can also be attributed to the discrepancy between global theoretical policies and their concrete applications that attempt to take local contexts into account: “Our Community Standards apply to everyone, all around the world, and to all types of content” (Facebook, 2021a). Extrapolated on a global scale, Facebook policies are defined as the lowest common denominator shared by all countries and cultures; they are implemented taking the local legal and cultural contexts into consideration. However, while Facebook’s community standards and their updates are accessible to anyone, the internal local rules remain hidden. Keeping the in-house rules hidden is framed as a counter-extremism measure: Monika Bickert, Head of Global Policy Management at Facebook, argued that “we are careful not to reveal too much about our enforcement techniques because of adversarial shifts by terrorists” (Bickert, 2018).

Facebook’s most recent initiatives for more transparency

Some of Facebook’s most recent initiatives specifically address the issue of transparency. In May 2021, Facebook created its online ‘Transparency Center,’ which gathers its policies (previously consultable in the Help Center), as well as regular reports about its enforcement strategies and activities. The terms used by Facebook to describe its transparency strategies are quite revealing. We have identified three strategies. Firstly, Facebook staff members use the word ‘transparency’ in a general sense. For example, Sheryl Sandberg, chief operating officer at Facebook, refuted that the organization Women for America First called for violence and organized transport to the Capitol riots in January 2021 on Facebook. She claimed that these events were “largely organized on platforms that don’t have our abilities to stop hate, don’t have our standards and don’t have our transparency” (quoted in McNamee and Ressa, 2021). Secondly, they seem to exclusively use the word ‘transparency’ in a more concrete way when they refer to Facebook’s Transparency Reports and the number of pieces Facebook restricts in each country (e.g., Paul, 2021). Thirdly, they do not refer to ‘transparency’ but instead to ‘visibility’, when they describe the Transparency Center in claiming that it seeks to “give our community visibility into how we enforce our policies, respond to data requests and protect intellectual property, while monitoring dynamics that limit access to Facebook technologies” (Facebook, 2021c). Transparency is explicitly claimed regarding quantitative insights. For the other ones, including the policies, ‘visibility’ seems more appropriate.

Since 2018, Facebook has published its Community standards enforcement reports (CSER) on a quarterly basis. It provides metrics on how Facebook enforces its policies. In 2018, Facebook commissioned an independent Data Transparency Advisory Group (DTAG) to assess their methodological design. The DTAG was invited to review whether “Facebook is accurately identifying content, behavior, and accounts that violate the Community Standards policy, as written” [17]. This means that their mission was to review whether Facebook is using accurate and meaningful statistics; they did not evaluate the community standards. By contrast, the Facebook’s Oversight Board aims to narrowly analyze the implementation of Facebook’s community standards. It consists of a board of 40 independent members whose mission is to select and review emblematic cases and to uphold or overturn Facebook’s content decisions relating to
these cases. They started to work on cases in October 2020; the conclusions of their deliberations on the first five selected cases were published in January 2021. They concern Facebook or Instagram posts. Their decisions are meant to be binding — Facebook will have seven days to implement the Board’s decisions — even though Facebook is under no legal obligation to abide by its outcomes. For some activists, the Oversight Board is nothing but “a PR effort that obfuscates the urgent issues that Facebook continually fails to address: the continued proliferation of hate speech and disinformation on their platforms” (anonymously quoted in Perrigo, 2021), like tweaking its algorithms to reduce the rapid spread of hate speech. Other observers are more optimistic. For Daniel Weitzner, the director of MIT’s Internet Policy Research Initiative, “The ambition for the Oversight Board is for it to have a quasi-judicial role, and the key thing about any judicial institution is it has to have legitimacy to earn deference. [...] Over time, if this body makes decisions that are seen as reasonable, and Facebook follows them, I think they’ll become a part of the landscape” (quoted in Perrigo, 2021). Ultimately, the Oversight Board’s first decisions will be defining and will send a signal on how ambitious the Board will be in checking Facebook. The Board is composed of 20 academic, political, and civic leaders from all over the world. Their identity is public (e.g., Culliford, 2021). Each selected case is examined by five members of the Board. When they reach a majority decision, their decision goes to the full 20-member Board for a vote.

In the context of these longstanding issues and Facebook’s recent initiatives for more transparency, we sought to empirically investigate what aspects of Facebook’s community standards still raise issues in the moderation of extremist practices and to what extent they address longstanding recommendations of improvement. To provide an answer to this research question, we analyzed 644 Facebook and Instagram posts written by extreme-right actors, as we explain in the next section.

---

**Materials and method**

Our empirical research is based on the analysis of 644 multimodal posts (text + still visual content) published by two Facebook and two Instagram profiles between September 2018 and January 2020. We set up our corpus according to three criteria. The first criterion was research innovation: since research focuses on the most established extreme-right political parties or social movements, we chose instead to analyze practices carried out on a smaller scale, by relatively small groups whose communication practices are not designed by political communication strategists. The second criterion was the profiles’ popularity: we focused on groups who are small but who, nevertheless, enjoy some popularity on Facebook or Instagram. They publish content quite regularly and are followed or liked by a relatively high number of people (see Table 1). Lastly, heterogeneity was our third criterion. With this research, we sought to qualitatively identify extremist content in their greatest possible diversity; we did not seek to compare the frequency of specific extremist content between homogenous profiles, with quantitative methods. For this reason, the profiles’ variety was important, in terms of style, platform (Facebook and Instagram) and language (French, English, and German).

For feasibility reasons, we selected four profiles among an initial dataset comprising 33 extreme-right EU-based accounts which were selected through chain-referral sampling (e.g., Dillon, et al., 2020; Forchtner and Kolvyraa, 2017; Klausen, 2015; Rowe and Saif, 2016). Visual communication is key on social media, and in extremist propaganda, in particular. As evident from earlier in this paper, visual techniques allow bans to be more easily circumvented. For this reason, all the multimodal posts published by these groups between September 2018 and January 2020 were analyzed (N = 644). The selected four accounts are the following:

- Action Nationale et Radicale/ANR, meaning ‘National and Radical Action’ is a group of French nationalist activists who advocate “racial nationalism as opposed to pseudo-patriotism that is multi-colored” (ANR, 2020).
- Le Parti nationaliste français — branche de Lyon /PNF, meaning the French nationalist Party — City
Lyon branch, is not a political party but a French political movement that notably considers that “each human being can only find their magnitude, strength and fulfilment in the group from which they come” (PNF, 2020).

- Europa Invicta/EI, meaning ‘Undefeated or unconquered Europe’, is an Instagram account in English. Europa Invicta was created to celebrate Europe’s landscapes, cultural heritage, and technological advances in response to the negative feelings caused by the immigration threat to the Old Continent, a.o.
- Patrioten_nrw/NRW is a German group of extreme-right activists, who a.o. organize actions and demonstrations with neo-Nazis, partly in North Rhine-Westphalia.

### Table 1: Our corpus of extreme-right content on Facebook and Instagram.

<table>
<thead>
<tr>
<th>Group’s initials (platform)</th>
<th>Number of multimodal posts in the time period covered</th>
<th>Number of followers in May 2020</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANR (FB)</td>
<td>96</td>
<td>731</td>
<td>French</td>
</tr>
<tr>
<td>PNF (FB)</td>
<td>46</td>
<td>1,067</td>
<td>French</td>
</tr>
<tr>
<td>EI (I)</td>
<td>407</td>
<td>7,317</td>
<td>English</td>
</tr>
<tr>
<td>NRW (I)</td>
<td>95</td>
<td>1,062</td>
<td>German</td>
</tr>
<tr>
<td>Total</td>
<td>644</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For ethical and GDPR-related reasons, these four profiles were public at the time of collecting the data. Europa Invicta became a private Instagram account some months after the collection process. Both ANR’s Web site and Facebook profile have been defunct since the fall of 2020.

We first analyzed the logo of the four groups. Then, we qualitatively analyzed their 644 posts following grounded theory (Strauss and Corbin, 1997), to identify how they potentially play with Facebook’s policies. Grounded theory is an inductive qualitative approach that is frequently used by social scientists when they seek findings and ideas that emerge from empirical data. In a first phase, three independent researchers analysed the dataset twice. After this analysis and review, they discussed their findings together and grouped them into three categories: 1) references to extremist figures; 2) violence incitement; and, 3) explicit or cloaked hate speech against particular groups or individuals. In the next section, our findings are presented and completed with Article 19’s and Facebook Oversight Board’s observations and recommendations on the first five cases that they selected, and especially on the three hate-speech related ones (FB-I2T6526K, FB-QBJDASCV, FB-2RDRCAVQ). The other two cases concern nudity in the context of prevention against breast cancer (IG-7THR3SI1) and COVID19-related disinformation (FB-XWJQBU9A), respectively.

---

### Results

*Extremist views through organizational logos: The issue of “representing” “dangerous” organizations*
ANR and EI have clearly understood the evocative power of logos: this tool for visual identity is embedded in every post they published. The logos of the four groups can be classified in two categories: 1) those that use visual codes belonging to Western cultural background and are frequently used by the extreme right (the lion imagery for PNF; eagle or phoenix imagery for EI); and, 2) those that are based on the visual repertoire of banned ideologies, in our case Nazi ideology (NRW and ANR). NRW used the eagle of the service flag of the free State of Prussia, used by the Nazis, and replaced the original swastika on the eagle’s chest with the iron cross. NRW also added the green and red colors of North-Rhine Westphalia. Probably deliberately, they decided not to make use of the specific Nazi logo of the eagle with the iron cross, which is forbidden (Figure 1) but they decontextualized these well-known Nazi visuals instead. This way, NRW created a symbol through only slightly modifying Nazi symbols.

ANR’s logo is topped with the wolfsangel, a banned symbol in Germany which was used by the Nazis, notably the Hitler Youth. In ANR’s logo, the wolfsangel is visually separated from the other two symbols, unlike the runic symbols used in the Thor Steinar brand logo for example, the blending of which was considered as a new logo and considered legal by German courts (see the second section of this paper). These two examples highlight the infinite creative possibilities that make the dividing line between what is and is not acceptable particularly difficult to draw.

![Figure 1: From left to right, PNF — branch of Lyon’s logo; Europa Invicta’s logo; ANR’s logo; NRW’s logo; and, the Nazi symbol combining an eagle and a swastika.](image)

In their community standards related to dangerous individuals and organizations, Facebook declares the following: “we do not allow symbols that represent any of the above organizations or individuals to be shared on our platform without context that condemns or neutrally discusses the content” (Facebook, 2021a).

The idea of “representation” is the key idea in this standard but is rather vague and does not clarify whether representation means using existing symbols in their totality or partially, or whether associations, and more generally creative uses of symbols, are allowed or not, like in German law (see the second section of this paper), and, if that should be the case, to what extent associations have to be direct to be banned.
Neither Article 19’s recommendations nor the Board’s first five selected cases showcase extremist visual content or symbols. On a more general note, the Board addresses the issue of using vague terms, including ‘representation’.

Quotes from dangerous organizations or individuals: The issue of endorsement

Our corpus contains quotes by various WWII collaborators and by David Lane, the member of the white supremacist terrorist group The Order, who was sentenced for conspiracy and participation in the murder of a Jewish journalist. Lane’s quote consists of the translation of his well-known statement “We must secure the existence of our people and a future for white children,” which has become a rallying cry for right-wing extremists, sometimes coded as “the fourteen words”. These quotes are embedded in an image comprising the actor’s (dignified) portrait and the organization’s logo. In most cases, the quote is the only textual content of the post.

Facebook claims that they remove “content that expresses support or praise for groups, leaders, or individuals involved in [e.g., terrorist activity, organized hate] activities” (Facebook, 2021a). Firstly, this community standard raises the issue of using vague terms like ‘support’ and ‘praise’. Both the Board and Article 19 before it refer to this issue. Secondly, interestingly, the issue of endorsement is absent from this rule; it is only mentioned in the section about visual symbols, in which Facebook does not allow “symbols that represent any of the above organizations or individuals to be shared on our platform without context that condemns or neutrally discusses the content” (Facebook, 2021a). Yet, endorsement is a key issue in extremist propaganda; extremist ideas are very often referred to in quotes, notably to legitimize ideas and increase credibility (e.g., Ahmed and Pisoiu, 2019).

A case analyzed by the Board raises similar policy issues as the quotes we observed in our corpus. The Board overturned Facebook’s decision to remove a post which included an alleged quote from Joseph Goebbels, the Reich Minister of Propaganda in Nazi Germany. The Board claimed Facebook’s community standards were not clear enough, preventing users from regulating their content accordingly, in line with the international standards mentioned earlier (Facebook Oversight Board, 2021). Interestingly, in its response to the Board, Facebook provided them details concerning three aspects of this standard: the type of quoted content considered as objectionable, their approach to endorsement, and the organizations and individuals designated as dangerous. The Board reported that “Facebook states that it treats content that quotes, or attributes quotes (regardless of their accuracy), to a designated dangerous individual as an expression of support for that individual unless the user provides additional context to make their intent explicit” (Facebook Oversight Board, 2021). Facebook based its decision on the designated individual, regardless of the accuracy of the quote, but also, regardless of the content of the quote: “Facebook removed the post because the user did not make clear that they shared the quote to condemn Joseph Goebbels, to counter extremism or hate speech, or for academic or news purposes” (Facebook Oversight Board, 2021). In doing so, Facebook applies the same standard of automatic endorsement, except for the explicit counter-endorsement for the visual symbols, without it being stated in their standards.

In Facebook’s policy approach to quotes based on their alleged author exclusively, posts that do not support extremist ideologies, like the Nazi party’s ideology or the regime’s acts of hate and violence in the case of Goebbels’ one, are also removed. For the Board, “the removal of the post clearly falls outside of the spirit of the policy” insofar as “it poses no risk of harm” (Facebook Oversight Board, 2021). This case highlights how Facebook’s current approach to endorsement in reported speech is not accompanied by an analysis of the content itself in order to determine whether or not it promotes violence incitement or hate speech. It also points out to what extent a more precise definition of positive terms like ‘praise’ or ‘support’ of dangerous individuals and organizations are necessary, including a clear distinction between the nature of the praised content, ideology-related or not, like in this Goebbels’ case. Lastly, Facebook claimed Goebbels was included in their internal list of dangerous individuals, “in response to the Board” (Facebook Oversight Board, 2021).

The general issue of designating ‘dangerous’ organizations and individuals
Our two findings above are related to the more general issue of designating organizations and individuals as ‘dangerous’. Dangerous organizations and individuals can be “terrorist organizations and terrorists” and “hate organizations and their leaders and prominent members,” along with “mass and multiple murderers (including attempts),” and “criminal organizations and their leaders and prominent members” (Facebook, 2021a). The Facebook Oversight Board raises the issue of designating organizations as ‘dangerous’ and recommends that Facebook provides a public list of these organizations and individuals “or, at the very least, a list of examples” (Facebook Oversight Board, 2021). These are longstanding issues. For example, Article 19 recommended three years ago that Facebook should identify the organizations that are considered as hate or terrorist ones and the way Facebook complies with various governments’ lists of designated terrorist organizations, especially in those cases where some of them can be considered terrorist ones by some actors, but as legitimate fighters (for freedom, f.i.) by others (Article 19, 2018). In October 2020, Facebook added a list of two types of organizations which they prohibit from maintaining a presence on the platform: “Militarized Social Movements (MSM), such as militias or groups that support and organize violent acts amid protests; and Violence-Inducing Conspiracy Networks, such as QAnon” (Facebook, 2021a). It is interesting to note that these additions remain very limited and are not transparently based on existing lists of dangerous organizations, as recommended. This contrasts with the explicit mention of only one organization, namely QAnon. Of course, lists will never prevent some whack-a-mole, since actors that got their content removed to account closed can easily recreate others under other names. Lists are not a panacea, but as long as they are used, they must be transparent and not left to the discretion of the platform. For the moment, Facebook has decided to reveal some of the names of these lists, following a totally confidential strategy.

Discussion and conclusion

Content moderation is not an easy task, not least when extreme-right actors strengthen mainstreaming strategies that blur these lines even more, like “a sort of slow-acting poison, accumulating here and there, word by word” [18]. This process is now commonly based on cloaking strategies that are often built on established or more experiential connotation. Of course, the controversial cases of content reviewing should not overshadow the millions of cases that Facebook has managed successfully. Yet, with its 2.7 billion users in the second quarter of 2020, Facebook plays an increasingly significant role as cultural intermediary, establishing the content and character of public discourse, and thereby determining our understanding of free speech (Ben-David and Matamoros Fernández, 2016; Caplan, 2018; Gillespie, 2012; Klonick, 2018).

Social media platforms like Facebook benefit from a significant degree of freedom by being allowed to simply draw inspiration from legislation like the First Amendment. This freedom allows them to frame themselves as “custodians” to their users’ content, with the consequence that “content moderation is treated as peripheral to what they do — a custodial task, like turning the lights on and off and sweeping the floors. The custodian approach also allows them to keep the cost of moderation to a minimum. It is occasionally championed in response to criticism, but otherwise it is obscured, minimized, and disavowed.” [19] However, over time and particularly during Trump’s mandate, social media have been increasingly urged to respond to mass pressure toward more transparency.

Some longstanding clarity issues still unaddressed

In this paper, we sought to take stock of the longstanding issues of transparency in Facebook’s 2021 community standards (RQ1). Our corpus-driven analysis of Facebook’s community standards revealed two clarity issues. Firstly, the extreme-right groups’ logos raise the issue of the definition of ‘representing’ dangerous organizations. Logos play a major role as key “bonding icons” between sympathizers (Wignell, et al., 2017). In our corpus, they were either implicit, created on the basis of visual codes belonging to our
cultural background and frequently used by the extreme right or explicit, made of the visual repertoire of the Nazi ideology. In line with Miller-Idriss (2018), it remains questionable to what extent the play with symbols, as in the case of the ANR and NRW logo, can be considered the creation of a new symbol in Facebook’s policies, like by German laws, for example, and therefore, to what extent the social platform allows creativity in using new symbols out of banned ones. The second issue concerns endorsement and the difficulties in determining when such endorsement starts, as well as how to draw a distinction between the quoted person and the quoted content itself. This is a key issue insofar as implicit endorsement through quotes is a common practice of legitimization and credibility among extremists (e.g., Ahmed and Pisoiu, 2019).

In order to increase its transparency, Facebook established the Oversight Board, which is composed of independent reviewers. In January 2021, Facebook’s Oversight Board reached its first decisions on five cases. In this research, we sought to analyze how the Board positions itself in relation to the longstanding issues regarding transparency in its first reviews ($RQ2$). Interestingly, four out of the five cases raised issues about the transparency of Facebook’s community standards regarding the criteria according to which content is considered as “violating”: Case One (FB-I2T6526K) made it clear that the definition of ‘generalization’ was missing in the standard related to prohibited generalized statements of inferiority; Case Three (IG-7THR3SI1) pointed out the inconsistencies between Instagram’s Community Guidelines and Facebook’s Community Standards, the latter taking precedence; Case Four (FB-2RDRCAVQ) highlighted how the approach to endorsement, the content of quotes and the notions of ‘praise’ or ‘support’ are vague. This case also stressed that the individuals or organizations designated as dangerous were not specified either in lists or with examples (except QAnon since October 2020). Lastly, Case Five (FB-XWJQBU9A) illustrated how the misinformation and imminent harm rule was vague. While optimists will be pleased that the Board identified several possible improvements, others will notice that most of these recommendations are longstanding issues, many of which have been addressed in international standards for 10 years. One can hope that Facebook will follow the Board’s recommendations, although one can reasonably wonder why they would follow them now while they have been addressed in international standards and NGOs’ reports for so long.

**Confidential in-house rules and the issue of lack of transparency**

Furthermore, our analysis reveals how the discrepancy between Facebook’s community standards and in-house rules remains unaddressed ($RQ3$). While the standards are applied to “everyone, all around the world, and to all types of content” (Facebook, 2021a), the internal regulations make it possible to implement the standards taking the local legal and cultural contexts into consideration. This discrepancy is a major issue, insofar as Facebook has been claiming that while the community standards are broad common denominators for their community around the world, the internal rules are concrete and contextualized *modus operandi*, reviewed through weekly audits of content reviewers’ work to ensure that the rules are implemented consistently, according to Monika Bickert, Head of Global Policy Management at Facebook (quoted in Angwin and Grassegger, 2017). Despite the importance of the internal rules, the Board only base their decisions on Facebook’s community standards and values as well as on international standards. Their members only consult documents that are publicly accessible. These documents, which do not include the in-house rules, are also widely available to any Facebook user. It seems unsatisfactory that the Board is pleased with the information on these internal rules that Facebook provides to them only reactively, “in its response to the Board” (Facebook Oversight Board, 2021). When they reviewed Facebook’s methodological design for its Community Standards Enforcement Report in 2018, Facebook’s Data Transparency Advisory Group (DTAG) did not have access to some information because of the mass of data it implied (Bradford, et al., 2019). Yet, in the case of the Board’s contextualized qualitative analysis of a limited number of isolated cases, giving due consideration to the internal rules that concretize the community standards at a local level would be fully justified and would be one step closer toward transparency. In creating independent groups (DTAG and Oversight Board) but in giving them limited or no access to key information, Facebook’s efforts of transparency might turn out to be transparency theater: “When organisations professing transparency act in ways that seem transparent, but fail to be useful when scrutinized, they are performing transparency theater” (Cherkewski, 2017). The emblematic case of
Facebook’s policies against extremism: Ten years of struggle for more transparency

Trump’s suspension has already generated accusations of transparency theater:

But if this [the work of the Oversight Board about the suspension of Donald Trump] looked like governance, that is precisely the point: The process was really little more than theater, a well-tailored spectacle of self-regulation by a company eager to avoid interference from actual governments. Its audience for this performance is, in part, Congress, which has spent the first two decades of the 21st century largely abdicating its role to regulate online spaces, though the events of Jan. 6 might finally stir legislative action. It is in Facebook’s interest to ensure that doesn’t happen, and the best way to do so might be to pass itself off as a sovereign power in its own right. By dramatically appealing to the arbitration of the Oversight Board, it is attempting to achieve just that (Atherton, 2021).

Transparency theater follows a similar logic as security theater in providing a feeling through performance and illusion, like Schneier claimed in the post 9/11 context: “Security is partially a state of mind. If this is true, then one of the goals of a security countermeasure is to provide people with a feeling of security in addition to the reality. But some countermeasures provide the feeling of security instead of the reality. These are nothing more than security theater. They’re palliative at best” [20].

By way of conclusion and in an effort to pave the way for future research, we wish to end this paper with the issue of Facebook’s in-house regulations. The Board pleads for concrete lists, or examples at the very least. For what reasons is it so unimaginable that internal rules would be made public, or at least scrutinized by independent committees, such as the Board? What may be the legitimate hurdles that might not be overcome? To be sure, these documents would be long, not very user-friendly and rather technical, but isn’t this a sine qua non of finely-grained regulations? Facebook’s argument that they do not want to “reveal too much about our enforcement techniques because of adversarial shifts by terrorists” (Bickert, 2018) is understandable but it does not measure up to ICCPR’s Article on Freedom of expression, according to which “a norm, to be characterized as a ‘law’, must be formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly and it must be made accessible to the public” [21]. Of course, “content policies are not law; they’re policy” [22], which can lead to leeway in applying their rules when they are helpful or in sidestepping them when they are too constraining (Gillespie, 2018). Law terminology is used to qualify the Oversight Board: its role is sometimes described as “quasi-judicial” (e.g., Weitzner quoted in Perrigo, 2021) and when he launched it in 2018, Facebook founder and CEO Mark Zuckerberg named it a “Supreme Court” of Facebook (e.g., in Salinas, 2018). But like legitimate claims for laws, we still call on Facebook to make its regulations finely-grained, fully transparent and publicly accessible. As we pointed out in this paper, concrete rules and lists are not the be all and end all; some of them can easily be circumvented through endless whack-a-mole processes. But as long as Facebook’s industrial model for content moderation is based on replicability by automation or in second-split human decisions, these in-house rules are at the heart of its moderation system and should require public access and scrutiny accordingly. Undoubtedly, in a context of increasing intolerance of opposing views, concrete rules and lists will always be sensitive and will never end in consensus. Especially today, when you have defenders of the ‘cancel culture’ disproportionately punishing people who have behaved in ways that they consider politically unacceptable, on the one hand, and some conservatives and extremists pleading against social media censorship, for example, in response to Facebook’s and Twitter’s decision to ban Donald Trump from their platform, (which was the impetus for Florida’s right-wing 2021 “Stop Social Media Censorship Law”), on the other. Based on this new law (probably Unconstitutional on its face), social media platforms will be fined in case a state-wide political candidate is banned. Furthermore, legal initiatives that seek to curb online hate speech (e.g., Germany’s Network Enforcement Act NetzDG) are sometimes criticized as being too vague and are instrumentalized by authoritarian states to justify restricting online speech through vague and over-inclusive categories (Mchangama and Fiss, 2019). As these few
examples remind us, debates can only be free and open when the criteria that led to moderation decisions are transparent. ‘Visibility,’ as claimed by Facebook regarding its policies, cannot replace transparency. For the 153 public figures who wrote and published a famous open letter in Harper’s in June 2020, “the way to defeat bad ideas is by exposure, argument, and persuasion, not by trying to silence or wish them away. We refuse any false choice between justice and freedom, which cannot exist without each other” (Ackerman, et al., 2020). Defeating bad ideas by exposure or not is beyond the scope of this paper, but we claim that the exposure of the concrete criteria that are used to moderate them appears as an absolute necessity and can no longer be silenced. Therefore, internal rules have to be integrated in the Oversight Board’s reviewing process, and not only casually in response to some of the Board’s observations. In doing so, the Board might truly become an efficient tool to reduce the tension between broad policies that give leeway to the contextualized enforcement of internal rules on the one hand, and working conditions that foster replicable decisions which do not leave much room for context, on the other. In doing so, the Board might avoid the label of being a mere “PR effort” (anonymously quoted in Perrigo, 2021). Its work has just begun; everything is still wide open.

About the authors

Catherine Bouko is an associate professor of communication at Ghent University (Belgium). Her research interests include critical discourse analyses of multimodal social media in political and societal contexts. Direct comments to: Catherine [dot] bouko [at] ugent [dot] be

Pieter Van Ostaeyen is an independent analyst and Ph.D. candidate at Katholieke Universiteit Leuven (Belgium). He published on the automatic detection of Jihadi hate-speech on social media, Koranic references in ISIS’ Dabiq magazine, the Islamic State’s strategic trajectory in Africa and Belgian Foreign Fighters. E-mail: pieter [dot] vanostaeyen [at] kuleuven [dot] be

Pierre Voué is an artificial intelligence practitioner at Textgain (Belgium). His research interests include natural language processing, online hate speech, and online polarization and extremism. E-mail: pierre [at] textgain [dot] com

Acknowledgments

We thank First Monday’s referees for their very useful suggestions.

Notes

1. The Overton window is a model named after political theorist Joseph Overton. It is a framework for understanding the range of ideas the mainstream public is willing to consider and accept.


5. Kekistan is a fictional country invented by 4chan members in the U.S. in 2015. The name was derived from “Kek”, a semi-ironic deity who is portrayed as a bringer of chaos and darkness. It has been criticized for weaponizing irony to spread alt-right ideas.
Facebook's policies against extremism: Ten years of struggle for more transparency

Figure 2: Flag representing the imaginary state Kekistan.

9. Mchangama and Fiss, 2019, p. 3.
10. Article 19, 2018, p. 6.
References


Article 19, 2018. “Facebook Community Standards: Analysis against international standards on freedom of expression” (30 July), at https://www.article19.org/resources/facebook-community-standards-analysis-
Facebook's policies against extremism: Ten years of struggle for more transparency


Lisa Bogerts and Maik Fielitz, 2019. “‘Do you want meme war?’ Understanding the visual memes of the German far right,” In: Maik Fielitz and Nick Thurston (editors). Post-digital cultures of the far right: Online actions and offline consequences in Europe and the US. Bielefeld: Transcript Verlag, pp. 137–154.


Facebook’s policies against extremism: Ten years of struggle for more transparency


Daniel Kreiss and Shannon C. Megregor, 2019. “The ‘arbiters of what our voters see’: Facebook and
Facebook’s policies against extremism: Ten years of struggle for more transparency


Ranking Digital Rights, 2015. “Corporate accountability index. 2015 research indicators,” at
Facebook's policies against extremism: Ten years of struggle for more transparency


Freedom of Expression and countering violent extremism,” at
accessed 8 August 2021.


---

Editorial history

Received 10 May 2021; revised 10 June 2021; revised 15 June 2021; revised 28 July 2021; revised 30 July 2021; accepted 30 July 2021.

Copyright © 2021, Catherine Bouko, Pieter Van Ostaeyen, and Pierre Voué. All Rights Reserved.

Facebook’s policies against extremism: Ten years of struggle for more transparency by Catherine Bouko, Pieter Van Ostaeyen, and Pierre Voué. First Monday, volume 26, number 9 (September 2021). doi: https://dx.doi.org/10.5210/fm.v26i9.11705