Meta's Oversight Board

Challenges of Content Moderation on the Internet

Pamela San Martín*

Abstract

The regulatory framework of international human rights law (IHRL) was originally designed to protect individuals from the power of states. However, the global expansion of private entities' power raised social awareness of the impact of business on human rights. The discussion around businesses and human rights has not been oblivious to the debates surrounding freedom of expression in the era of digitalisation. While the digital public sphere has undoubtedly democratised the public conversation, it has also generated risks and challenges around content moderation and freedom of expression. Platforms have therefore expanded their powers, creating a growing number of content rules and increasing the amount of content they remove. This has led to the rise of legitimate complaints concerning the lack of remedies for the proliferation of harmful speech and the 'excessive' content moderation that limits or excludes content protected by the freedom of expression. Within this framework, the discussion remains open as to who should regulate social media and how it must be done, what oversight mechanisms should be implemented to ensure people the protection of their rights online and prevent negative on- and offline impacts. This article provides an outline of the challenges generated by communication in the digital sphere and addresses the discussions on public/private regulation of content moderation. Additionally, this article addresses the importance of ensuring that the rules applied by the platforms are based on IHRL and discusses the creation of Meta's Oversight Board as a self-regulatory mechanism, as a novel and complementary alternative to governing social media.

Keywords: content moderation, social media, Oversight Board.

1 Introduction

The regulatory framework of international human rights law (IHRL) was originally designed to protect individuals from the power of states. However, the dramatic global expansion of private entities and corporations, the corresponding increase in transnational economic activity and the enormous power private entities wield over the exercise of people's rights and freedoms, raised social awareness of the impact of business on human rights. As a result, the issue of 'business and human rights' (hereafter BHR) became a permanent fixture on the global political agenda in the 1990s.[1] Since then, the debate on the relationship between the public and private sectors has also become increasingly relevant.

After extensive discussions, the United Nations (UN) Human Rights Council unanimously adopted the UN Guiding Principles on Business and Human Rights (UNGPs)[2] in 2011. Although these UNGPs are non-binding, they have been endorsed by numerous large-scale companies around the world.[3] The UNGPs implement a 'Protect, Respect and Remedy' framework, which rests on three pillars: the state's duty to protect against human rights abuses by third parties, the corporate responsibility to respect human rights by acting with due diligence to avoid infringing on the rights of others and address human rights' adverse impacts with which they are involved, and the need for greater access by victims to an effective remedy when their rights are violated.[4]

Since the emergence of the Internet and social media changed the way people communicate and access information, the discussion around BHR has not been oblivious to the debates surrounding freedom of expression in the era of digitalisation. In fact, in the realm of digital communications the private sector 'wields enormous power over digital space, acting as a gateway for information and an intermediary for expression', has become 'a driving force behind the greatest expansion of access to information in history' and is largely responsible for

* Pamela San Martín is a Member of the Oversight Board Department at the Public Company Accounting Oversight Board.

1 HRC, *Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, John Ruggie. Guiding Principles on Business and Human Rights: Implementing the UN 'Protect, Respect and Remedy' Framework*, UN doc. A/HRC/17/31 (2011), at para. 1 (last visited 26 April 2023).

2 The 'United Nations Guiding Principles on Business and Human Rights' were developed by the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises. The Special Representative annexed the Guiding Principles to his final report to the Human Rights Council (A/HRC/17/31), which also includes an introduction to the Guiding Principles and an overview of the process that led to their development. The Human Rights Council endorsed the Guiding Principles in its resolution 17/4 of 16 June 2011.

3 OHCHR, 'An Authoritative Global Framework on Business and Human Rights Turns 10' (17 June 2021), https://www.ohchr.org/en/stories/2021/06/authoritative-global-framework-business-and-human-rights-turns-10 (last visited 26 April 2023).

4 HRC (2011), above note 1, at para. 6.

doi: 10.5553/ELR.000253

'the contemporary exercise of freedom of opinion and expression'.[5]

And while the digital public sphere has undoubtedly democratised the public conversation in unprecedented ways, it has also generated enormous risks and challenges around content moderation and freedom of expression. Platforms have therefore expanded their powers, creating a growing number of content rules and increasing the amount of content they remove.[6] This has led to the rise of legitimate complaints concerning, on the one hand, the lack of remedies for the proliferation of harmful speech, and on the other hand, the 'excessive' content moderation that limits or excludes content protected by the freedom of expression.

Content moderation commonly draws attention to individual content that is kept up or removed from platforms, to the rules that should govern the platforms and to their duty to properly enforce these rules. However, content moderation at scale also involves difficult trade-offs and includes other elements – such as system design decisions – that influence the exercise of rights in the digital sphere and that are relevant to the analysis of platforms' compliance with their human rights responsibilities under the UNGPs.

Within this framework, the discussion remains open as to who should regulate social media and how it must be done, what oversight mechanisms should be implemented to ensure people the protection of their rights online and prevent negative on- and offline impacts. It is precisely in this context that the Oversight Board emerged as an independent self-regulatory mechanism. As this is a novel exercise that was created less than three years ago, amidst the Covid-19 pandemic, it is worth analysing its scope, the opportunities it offers and the challenges it faces, in view of the numerous regulations that have arisen and that will arise in this area in the coming years.

In order to address these issues, this article will be divided into three parts. Part I will provide a general outline of the challenges generated by communication in the digital sphere, the changes that the emergence of social media have brought about in communications and access to information, as well as the context in which online content moderation occurs, emphasising some of the particularities and complexities of content moderation on a global scale.[7] Part II will briefly address discussions on public/private regulation of content moderation, recognising that, at present, it is precisely private companies that govern the platforms, with very limited state controls. Without going in depth into the discussion of how platforms should be regulated and by whom, it will also address the importance of ensuring that the rules applied by these are based on IHRL, as well as the risks and opportunities for state regulation of these platforms. Part III will discuss the creation of Meta's Oversight Board (hereafter OSB), as a self-regulatory mechanism, external and independent, but created by a private company, as a novel and complementary alternative to governing social media in a more transparent, coherent and consistent manner, and with emphasis on IHRL. This part will also address the scope of the Oversight Board, the guarantees with which it was built, and some of the results of its work.

2 Complexities of Content Moderation at Scale

The digital public sphere has generated new challenges as a result of the predominant role of private entities in the exercise of freedom of expression – especially in closed information environments and places where platforms are synonymous with the Internet – as well as the global nature of platforms and the speed, reach and large volume of content flowing across them.

Just to give an idea of the magnitude of the content involved in social media: in Q2 2022, Facebook removed 914,500,000 pieces of content:[8] in Q3 2022 YouTube removed 5,820,978 channels (involving 207,833,024 videos, due to channel-level suspensions) and 5,603,794 videos,[9] while TikTok removed 110,954,663 videos (52,287,839 by automation).[10] It is important to note that these figures do not account for the total amount of content posted by users, or the number of times that these platforms decided not to remove reviewed content. According to the information provided by Meta[11] for the Oversight Board's Policy Advisory Opinion on the company's cross-check programme, by the end of 2021, the company 'was performing about 100 million enforcement attempts on content every day'.[12]

5 HRC, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, UN doc. A/HRC/32/38 (2016), at paras. 1-2.

6 T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (2018), DOI:10.12987/9780300235029; K. Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech', 131 *Harvard Law Review* 1598 (2018), https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/; J.L. Zittrain, 'Three Eras of Digital Governance' (2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3458435 (last visited 26 April 2023).

7 In this article, the term 'content moderation' is used to refer to '*platforms' systems and rules that determine how they treat user-generated content on their services*', as proposed by E. Douek, 'Content Moderation as Systems Thinking', 136 *Harvard Law Review* 526 (2022), https://harvardlawreview.org/2022/12/content-moderation-as-systems-thinking/ (last visited 26 April 2023).

8 Facebook, 'Community Standards Enforcement Report' (2022), taken from Douek, above n. 7.

9 YouTube, 'Community Guidelines Enforcement', https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB (last visited 11 May 2023).

10 TIKTOK, 'Community Guidelines Enforcement Report July 1 2022-September 30 2022', (19 December 2022), https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2022-3/ (last visited 26 April 2023).

11 On 28 October 2021, Facebook, Inc. changed its name to Meta Platforms, Inc. For consistency, this article uses 'Meta' to refer to the company, while references to 'Facebook' apply only to that specific social media platform.

12 Oversight Board 2021, PAO-2021-02 (*Policy Advisory Opinion on Meta's Cross-Check Program*) PAO-2021-02, https://oversightboard.com/decision/PAO-NR730OFI (last visited 11 May 2023).

Therefore, when analysing content moderation in the digital environment – particularly in social media – several elements must be taken into account: a) The important developments on the interpretation of the principles, scope and limits of the right to freedom of expression at the universal, regional and local levels; b) The changes in the way people are informed, communicate and interact, as well as the challenges, risks and opportunities that have arisen with this new forum; and c) The specific particularities of moderating content at scale – recognising that there are differences between platforms, as some, due to their size and position in the market, can significantly limit freedom of expression, while others do not have this power.

2.1 The Right to Freedom of Expression

In the Universal Human Rights system, the right to freedom of expression is provided for in Article 19 of the International Covenant on Civil and Political Rights (ICCPR) .[13] The scope of this right is broad, and it is guaranteed to all people without discrimination.[14] It includes 'freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice'.

Since, according to Article 19(3) ICCPR, the exercise of this right also 'carries with it special duties and responsibilities', it admits restrictions under certain strict and limited conditions. The three-part test of legality, legitimacy and necessity (which also includes an assessment of proportionality) determines whether or not the right can be rightfully limited. These categories have been widely developed by universal and regional bodies and mechanisms for the protection and promotion of human rights.[15] Thus, any restriction must:

a. Be provided for by law and formulated in a sufficiently clear and precise manner 'to enable an individual to regulate his or her conduct accordingly', and must be made accessible to the public. Similarly, laws must 'provide sufficient guidance to those charged with their execution … to enable them to ascertain what sorts of expression are properly restricted and what sorts are not', so that these rules do not confer 'unfettered discretion'.[16] This requirement prevents arbitrary censorship.

   Applied to platform rules, the UN Special Rapporteur on freedom of expression has stated that these rules should be clear and specific.[17] People using platforms should be able to access and understand the rules, and content reviewers should have clear guidance on their enforcement.

b. Have a legitimate aim. Legitimate aims are listed in Article 19(3) of the ICCPR: respect of the rights or reputations of others and the protection of national security, public order (*ordre public*), public health or morals.

   The UN Human Rights Committee has interpreted the term 'rights' to include the human rights as recognised by the ICCPR and, more generally, in IHRL.[18]

c. Be necessary and proportionate to achieve that legitimate aim. That is to say, 'appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected'. In this sense, restrictions 'must not be overbroad'. The principle of proportionality 'must also take into account the form of expression at issue as well as the means of its dissemination'.[19]

The UN Human Rights Committee has noted that restrictions on the exercise of freedom of expression 'may not put [the right itself] in jeopardy' and has emphasised that 'the relation between right and restriction and between norm and exception must not be reversed'.[20]

As the ICCPR does not create binding obligations for platforms as it does for states –although these have the positive obligation to protect individuals from the actions of private entities[21] – the digital age has triggered new questions about the extent to which the promotion and protection of the freedom of opinion and expression should be considered the responsibility of the Information and Communications Technology (ICT) sector.[22]

In his 2018 report, the UN Special Rapporteur on freedom of opinion and expression suggested that Article 19(3) of the ICCPR provides a useful framework to guide platforms' content moderation practices.[23] He later noted that although 'companies do not have the obligations of Governments … their impact is of a sort that requires them to assess the same kind of questions

13 At the regional level, in Art. 13 of the American Convention on Human Rights and Art. 10 of the European Convention for the Protection of Human Rights and Fundamental Freedoms, with some differences.

14 Freedom of opinion and freedom of expression 'constitute the foundation stone for every free and democratic society. The two freedoms are closely related, with freedom of expression providing the vehicle for the exchange and development of opinions. Freedom of expression is a necessary condition for the realisation of the principles of transparency and accountability that are, in turn, essential for the promotion and protection of human rights…. The freedoms of opinion and expression form a basis for the full enjoyment of a wide range of other human rights….' HRC, *General Comment N° 34*, UN doc. CCPR/C/GC/34 (2011), at paras. 2-4.

15 Such as the UN Human Rights Committee, the Inter-American Commission and Court of Human Rights, the European Court of Human Rights and the Special Rapporteurs on Freedom of Expression.

16 UN Human Rights Committee HRC (2011), above n. 14, at para. 25.

17 HRC, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, UN doc. A/HRC/38/35 (2018), at para. 46.

18 HRC (2011), above n. 14, at para. 28.

19 *Ibid*., at para. 34.

20 *Ibid*., at para. 21.

21 HRC (2016), above n. 5, at paras. 8 and 9. ('States have both a negative obligation to refrain from violating rights and a positive obligation to ensure enjoyment of those rights. These positive obligations may require public authorities to take steps to protect individuals from the actions of private parties.' and 'Human rights law does not as a general matter directly govern the activities or responsibilities of private business.').

22 *Ibid*., at para. 5.

23 HRC (2018), above n. 17.

doi: 10.5553/ELR.000253

about protecting their users' right to freedom of expression'.[24]

Consequently, we face an unforeseen challenge in this new digital era: it is no longer the state, or the state exclusively, that controls the exercise of rights. Private entities, particularly the owners of the current large digital platforms, have significant control over what is said and what is heard in the public space. Currently, there is no sufficient legal or theoretical framework to address this new context.

2.2 Changes, Challenges and Risks of the Digital Era

During most of the 20th century, radio, television and print media were the main channels of information. However, with the arrival and widespread use of the Internet, public and political communication evolved, changing the way in which people access and interact with information. New digital communication spaces and new digital alternative media emerged since. Social media became the most widely used medium,[25] allowing all users to create and share content and to participate in discussions in the public space. Thus, the monopoly of communication in a few hands disappeared, giving way to deconcentrated, plural and diverse information to which people have immediate access.

In this scenario, the technological optimism that accompanied the emergence of social media, built on the values of decentralisation and horizontality, suggested that the absence of specific regulation was the best way to optimise its potential. However, over time, it became clear that although the Internet could probably be the most important democratising instrument of speech of the last centuries,[26] it has also created enormous risks and challenges that must be addressed.

Social media platforms enable people – even those historically excluded or marginalised – to connect with each other; to access knowledge, culture, progress and information; to denounce and raise awareness on human rights violations; and to debate and exercise political control over public affairs in unprecedented ways. However, it also creates a space where harmful content such as hate speech, intimidation and harassment, incitement to violence and the spread of misinformation has the chance to proliferate.

Consequently, this leads to the question of how to enhance the benefits while preventing and mitigating the potential damage caused by freedom of expression on the Internet without imposing restrictions that endanger the exercise of the right itself. It is necessary to guarantee the safeguards that have been built in terms of freedom of expression so that it continues to make sense in the digital environment.

The delicate balance of content moderation for the protection of freedom of expression must not forget this starting point, nor can this analysis be oblivious to the context in which this public discussion is taking place: to the strength and reach of certain voices,[27] to the volume of content that is distributed through social media, to the speed with which information spreads in digital spaces, to the chilling effect that the coordinated reiteration of certain discourses can have on some voices; even, to its impacts on the democratic process itself.

2.3 Particularities of Content Moderation at Scale[28]

Social media platforms moderate content for very different reasons: to guarantee the free flow of speech online – putting people in direct contact with one another; for economic and financial profit; to comply with government requests; and to avoid harms in specific cases, among others.[29] In this sense, 'content moderation' involves a complex system where many very different values are at stake.[30]

Therefore, to analyse content moderation in the digital era, there are specific particularities that need to be considered. The first is the volume, speed and reach of the content that flows in the digital public sphere, which make it necessary for content moderation to combine *automation*[31] with human review.

24 HRC, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, UN doc. A/74/486 (2019), at para. 41.

25 According to the International Telecommunication Union (ITU), the UN specialised agency for ICTs, 'an estimated 5.3 billion people of the earth's 8 billion [were] using the Internet in 2022, or roughly 66 per cent of the world's population. At the same time, three quarters of the population aged 10 years and over own a mobile phone', www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx (last visited 26 April 2023). Additionally, for example, the number of active users of Facebook – currently the largest social media platform – went from 150 million to 2.963 billion users from 2009 to January 2023, https://datareportal.com/essential-facebook-stats (last visited 26 April 2023).

26 HRC (2016), above n. 5, ('The private sector's role in the digital age appears pervasive and ever-growing, a driving force behind the greatest expansion of access to information in history.').

27 Although everyone can take part in the public conversation these days, not everyone has the same possibility of influencing that public conversation. Not all speech is heard with the same force or reach, nor is it distributed with the same virality.

28 This section is not intended to be a comprehensive review of all the particularities of content moderation at scale, as platforms' systems, designs and rules that can impact content moderation are much broader. References to some of the Oversight Board's decisions have the purpose of illustrating or providing additional examples or information on the issues addressed, although an ampler description of the Board's work can be found in Part III.

29 Gillespie, above n. 6 ('Platforms must, in some form or another, moderate: both to protect one user from another, or one group from its antagonists, and to remove the offensive, vile, or illegal—as well as to present their best face to new users, to their advertisers and partners, and to the public at large.'); and Klonick, above n. 6 ('three main factors influenced the development of these platforms' moderation systems: (1) an underlying belief in free speech norms; (2) a sense of corporate responsibility; and (3) the necessity of meeting users' norms for economic viability…').

30 Douek, above n. 7. ('The scale and pace at which content moderation must operate make the tradeoffs between these individual interests and other goals such as overall speed, accuracy, and consistency …, an assessment of the level of risk in a particular context, and the level of technological capacity for moderating a certain kind of content.').

31 Although there are many varied automated systems used for content moderation, this section only refers to those that classify user-generated content based on matches or predictions, to produce an enforcement decision on specific content.

In social media, posts can go viral and cause damage in a matter of minutes. Therefore, in order for content moderation to minimally meet the requirements of timeliness, as well as to meet growing public demands for greater accountability, safety and security, social media platforms increasingly rely on automated tools.[32] Regulatory measures such as the German NetzDG[33] or the EU Code of Conduct against online hate speech,[34] which oblige platforms to remove content in very short time frames, have increased platforms' use of automated systems to detect illegal material proactively and on a large scale.

Although some infringing content is best identified through automation, as automated systems of enforcement are not sensitive to context[35] and offer little explanation for their decisions,[36] human review is critical where contextual cues are required for enforcement.[37]

Additionally, platforms' investments in both their automated tools and training data vary between regions and languages, with impacts on content moderation prac-tices.[38] Furthermore, concerns have repeatedly been raised about the lack of information surrounding the use and functioning of these systems.[39] Therefore, through different decisions and recommendations, the Oversight Board has sought to make the operation and results of Meta's automated enforcement more transparent and for the necessary safeguards on automation to be put in place to allow enforcement errors to be repaired.[40]

Secondly, *content is not language- or context-agnostic*. Content needs to be analysed in its context and in its language, taking into account the fact that languages change over time and from one place to another. Furthermore, due to the global nature of digital platforms, content transcends borders. Consequently, content published in one part of the world – with a specific meaning – can be accessed in very different regions – where its meaning can differ significantly – as will be addressed in more detail and with examples later. Therefore, distinguishing these contexts at such a large scale is extremely complex, particularly for enforcement through automation on a global level.[41]

Thirdly, moderation at scale *never is, nor can it be, perfect*.[42] In numbers, a recent report noted that, on average, 350 million photographs are uploaded on Facebook every day. Therefore, even if the company's enforcement decisions had a 99.9% accuracy rate, there would still be enforcement errors on 350 thousand photographs per day.[43] Errors in content moderation include both over- and under-enforcement of policies – with *over-enforcement* meaning the removal of non-infringing content and *under-enforcement* the non-removal of violating content. Indeed, part of the design of both platforms' rules and their systems involves deciding in which direction to err, whether in favour of *false positives* – removing non-violating content – or *false negatives* – letting violating content remain online.[44]

For example, in an effort to remove child exploitation and terrorist content online, platforms tend to over-enforce related policies. Although that entails the removal of some non-infringing content, it is a trade-off for ensuring that as much violating content as possible is removed. Nonetheless, the latter has had differential impacts in some regions of the world, particularly in those with a greater number of organizations or individuals designated as 'terrorists', or where people live under the

32 To flag potentially infringing content for human review, or directly remove it, if the likelihood of violation is high.

33 Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act) *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken, Bundesgesetzblatt Jahrgang 2017 Teil I Nr. 61, augegeben zu Bonn am 7. September 2017*.

34 European Commission, EU Code of Conduct on countering illegal hate speech online, https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (last visited 12 May 2023).

35 HRC (2018), above n. 17, at para. 56 ('Company responsibilities to prevent and mitigate human rights impacts should take into account the significant limitations of automation, such as difficulties with addressing context, widespread variation of language cues and meaning and linguistic and cultural particularities.').

36 C. Shenkman et al., 'Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis', *Center for Democracy and Technology* (2021), https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf (last visited 26 April 2023).

37 Oversight Board, 2020-004-IG-UA (*Breast Cancer Symptoms and Nudity*), https://oversightboard.com/decision/IG-7THR3SI1/ (last visited 12 May 2023). ('…enforcement which relies solely on automation, in particular when using technologies that have a limited ability to understand context, leads to over-enforcement that disproportionately interferes with user expression.').

38 *Ibid*. ('Machine learning algorithms rely on enormous amounts of training data…. It is well documented that datasets are susceptible to both intended and unintended biases.') and J. Rowe, 'Marginalised Languages and the Content Moderation Challenge', *Global Partners Digital* (2 March 2022), www.gp-digital.org/marginalised-languages-and-the-content-moderation-challenge/ (last visited 26 April 2023).

39 R. Gorwa, R. Binns and C. Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance', 7 *Big Data & Society* (2020), https://doi.org/10.1177/2053951719897945 (last visited 26 April 2023); HRC (2018), above n. 17, at para. 62.

40 Examples of these can be found in: Oversight Board (2020), above n. 35; Oversight Board, 2021-006-IG-UA (*Öcalan's Isolation*), https://oversightboard.com/decision/IG-I9DP23IB/ (last visited 12 May 2023); Oversight Board, 2022-004-FB-UA (*Colombian Police Cartoon*), https://oversightboard.com/decision/FB-I964KKM6/ (last visited 12 May 2023).

41 R. Caplan, 'Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches', *Data & Society* (2018), https://datasociety.net/wp-content/uploads/2018/11/DS_Content (last visited 12 May 2023).

42 M. Masnick, 'Masnick's Impossibility Theorem: Content Moderation at Scale Is Impossible to Do Well', (20 November 2019), www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well.shtml (last visited 26 April 2023).

43 Branka, *Facebook Statistics – 2023* (7 January 2023), https://truelist.co/blog/facebook-statistics/#:~:text=6.-,350%20million%20photos%20are%20uploaded%20on%20Facebook%20every%20day.,over%204%2C000%20photos%20every%20second (last visited 26 April 2023); Masnick, above n. 42.

44 E. Douek, 'Governing Online Speech: From 'Posts-as-Trumps' to Proportionality and Probability', 121 *Columbia Law Review* 759 (2021), https://columbialawreview.org/wp-content/uploads/2021/04/Douek-Governing_Online_Speech-from_Posts_As-Trumps_To_Proportionality_And_Probability.pdf (The choice to err in some cases 'is the price of getting it right, within a reasonable time frame (or at all), in the vast majority of cases'.).

doi: 10.5553/ELR.000253

de facto authority of designated entities.[45] On the other hand, platforms have been strongly criticized for the under-enforcement on hate speech and content that incites violence, due to the offline consequences it has produced – for example in Myanmar.[46]

However, error rates vary depending on multiple factors, ranging from platform design decisions to the resources that companies invest in the development and review of different mechanisms and tools for content moderation, which, as mentioned before, differ between languages and regions.

An example of this is the report on the human rights due diligence assessment conducted on the impact of Meta's policies and activities during the crisis in Israel and Palestine in May 2021, which 'identified both over-enforcement … and under-enforcement … of Meta content policies…, especially the Dangerous Individuals and Organizations … and Violence and Incitement….' It further stated that 'Arabic content had greater over-enforcement (e.g., erroneously removing Palestinian voice) on a per user basis (i.e., adjusting for the population size difference between Arabic and Hebrew speakers in Israel and Palestine)', which 'can be likely attributed in large part to Meta's policies which incorporate certain legal obligations relating to designated foreign terrorist organizations…, the fact that there was an Arabic hostile speech classifier but not a Hebrew hostile speech classifier … and the loss of Hebrew-speaking FTEs and outsourced content moderators in the weeks leading up to May 2021'.[47]

Therefore, although error choices are not alien to platform design, their assessment should include the need for timely action, the existing risk level in a particular context and the enforcement tools available, among other relevant factors.[48] Regardless, an adequate and robust human rights impact assessment is needed on the impacts of certain decisions – both on the design and enforcement of policies and on the platform's tools, products and investments – pursuant to the platform's responsibilities under the UNGPs. This includes the permanent assessment of the contexts of greatest risk or conflict, where the possibility of severe human rights violations could be higher.

Nonetheless, it is crucial to distinguish between human errors or reasonable disagreements and systemic substantial underlying problems. To this end, transparency on system design and operation, as well as on erroneous decisions and their reasons, is fundamental. The Oversight Board has therefore drawn Meta's attention several times to the fact that its transparency reports are not sufficient to meaningfully assess whether the types of errors detected in specific cases reflect a systemic problem and has recommended that it publicise more information and metrics that allow for an evaluation.[49]

Finally, content moderation does *not occur in a vacuum*. There are legal and extra-legal pressures from governments to remove content, and platform decisions on how to respond to government requests can have consequences on human rights.[50] Consequently, the Oversight Board has insisted that Meta be more transparent about governments' requests and the way they are addressed.[51]

2.3.1 Analysis of Context in Content Moderation

Context plays an essential role when it comes to enforcing rules based on concrete facts, considering how the same content can differ significantly in meaning and effect and how this impacts different groups in different ways – particularly marginalised communities. To understand speech, its context must be taken into account. While certain speech may be permitted, or could even be particularly relevant in a specific context, it may be dangerous or forbidden in a diverse context, where it may cause harm to people's lives offline or even become a challenge to democracy.

In this regard, digital platforms usually enforce rules globally, as those rules apply to a global public sphere – although exceptionally some have a regional scope.[52] Despite the analysis of context being essential for content moderation, it presents enormous challenges at a global level, particularly due to the volume and speed of content on online platforms.[53]

Thus, the *social and political context* in which a message is published can be decisive in determining if it is likely to cause harm. For example, in a recent decision of the Oversight Board in the context of the ongoing protests

45 Oversight Board 2022, 2022-005-FB-UA (*Mention of the Taliban in News Reporting*), https://oversightboard.com/decision/FB-U2HHA647/ (last visited 12 May 2023).

46 A. Warofska, 'An Independent Assessment of the Human Rights Impact of Facebook in Myanmar' (5 November 2018), https://newsroom.fb.com/news/2018/11/myanmar-hria/ (Meta 'commissioned an independent human rights impact assessment on the role of [their] services in Myanmar … The report concludes that, prior to this year, we weren't doing enough to help prevent our platform from being used to foment division and incite offline violence.') and T. Miles, 'U.N. Investigators Cite Facebook Role in Myanmar Crisis', (12 March 2018), www.reuters.com/article/us-myanmar-rohingya-facebook-idUSKCN1GO2PN ('Marzuki Darusman, chairman of the U.N. Independent International Fact-Finding Mission on Myanmar, told reporters that social media had played a "determining role" in Myanmar').

47 It was commissioned by Meta following a recommendation by the Oversight Board in the *Shared Al Jazeera post* decision. BSR, 'Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021. Insights and Recommendations', (2022), www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf (last visited 12 May 2023).

48 Douek (2022), above n. 7.

49 Examples of these can be found in: Oversight Board 2021, 2021-003-FB-UA (*Punjabi Concern Over the RSS in India*). https://oversightboard.com/decision/FB-H6OZKDS3/ (last visited 12 May 2023); Oversight Board (2021), above n. 40; Oversight Board 2022, 2022-007-IG-MR (*UK Drill Music*), https://oversightboard.com/decision/IG-PT5WRTLW/ (last visited 12 May 2023); Oversight Board (2022), above n. 45.

50 HRC (2016), above n. 5, at para. 5.

51 Examples of these can be found in: Oversight Board (2021), above n. 40; Oversight Board 2021, 2021-009-FB-UA (*Shared Al Jazeera Post*), https://oversightboard.com/decision/FB-P93JPX02/ (last visited 12 May 2023); Oversight Board (2022), above n. 49.

52 An example of these can be found in Meta's lists of prohibited slurs, which are market-specific. Oversight Board 2021, 2021-011-FB-UA (*South Africa Slurs*), https://oversightboard.com/decision/FB-TYE2766G/ (last visited 12 May 2023); In response to the Board's questions, Meta noted that while its prohibition against slurs is global, the designation of slurs on its internal slurs list is market oriented.

53 Caplan, above n. 41.

in Iran against the laws on mandatory hijab, the Board analysed the use of the slogan *marg bar Khamenei* – which literally translates into 'death to [Iran's supreme leader] Khamenei', but is often used as rhetorical political speech meaning 'down with Khamenei'. The Board stressed that in the context of the protests in Iran, such content posed very little risk of inciting violence[54] and that it should therefore remain on the platform. However, it recognised that in other contexts, 'death to' statements directed at public figures and government officials might not convey the same rhetorical meaning and should be treated differently.[55]

Similarly, the *linguistic context* of an expression matters as well. On many occasions, the harm caused by an expression derives precisely from the meaning it has in a specific context. Thus, expressed in a global network, words that may be inconsequential in one context may circulate and generate enormous harm in another. For example, the use of the phrase 'kill the cockroaches' would have no impact in most of the world. In fact, most would agree that they dislike cockroaches. However, in the context of the Rwandan genocide this same phrase was a call to kill the country's minority Tutsi population, the target group of the genocide, who were called 'cockroaches'.[56] Hence, a message in Rwanda calling to 'kill the cockroaches' was a call for genocide.

Furthermore, when analysing the *South Africa Slurs* case, the Oversight Board decided to uphold Meta's decision to remove a post in which the term 'kaffir' was used – a term that was on Facebook's list of prohibited slurs for the sub-Saharan market – as it is 'widely understood as South Africa's most charged racial epithet, closely linked to discrimination and the history of apartheid in that country'. However, in the same decision, the Board pointed out that previously, in the *Protest in India Against France* case,[57] it had ordered Meta to restore content that also used the term 'kafir'. The difference in the outcome was due to the fact that the term with one 'f', used in that case in India, was not a slur in that context but rather referred to 'non-believers'. The Board concluded that this situation demonstrates the difficulty for Meta to enforce 'a blanket prohibition on certain words globally, where similar or identical terms in the same or different languages can hold different meanings and pose different risks depending on their contextual use'.[58]

However, the need to analyse the context when moderating content is not limited to the social, political or linguistic context; there is also a *digital context* that must be taken into account. There are certain coordinated, organic or artificial online behaviours that have potential offline effects. These are designed through broader campaigns that can be aimed at harassing a specific person or group, for instance, or even at inciting violence. However, their assessment cannot be made solely from the analysis of an individual piece of content that is part of that campaign but requires the analysis of the digital context in which it was spread as well.

For example, since the departure of the International Commission against Impunity, CICIG,[59] from Guatemala, some of the judges who were involved in anti-corruption measures in the country have faced threats and harassment.[60] This has included coordinated campaigns to spread criticism on and disqualifications of these judges, both in social media and in different media outlets that support the government, creating an environment that could legitimise their persecution. Now, individually considered, posts aimed against them would probably be allowed, even more so as criticism of public figures is a particularly protected speech.[61] However, these posts should not be analysed individually or in a vacuum, but as a whole in a digital context so that they are addressed as part of that coordinated campaign.

The analysis of the digital context has become increasingly relevant in content moderation, due to the large disinformation and electoral interference campaigns[62] that have been part of critical questions raised against companies. These questions concern the arising damages on their platforms, which have real impacts on the lives and rights of people, and even on democracy.

Similarly, the analysis of the context cannot be oblivious to the harms that some content can generate in a specific group – particularly in the case of groups that are tar-

54 To reach this decision, the Board considered the six-factor test described in the Rabat Plan of Action. HRC, *Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence*, UN doc. A/HRC/22/17/Add.4 (2013).

55 As an example, the case decision states that 'during events similar to the January 6 riots in Washington D.C., 'death to' statements against politicians would need to be swiftly removed…. In such a situation, politicians were clearly at risk, and 'death to' statements are less likely to be understood as rhetorical or non-threatening in English'. Oversight Board 2022, 2022-013-FB-UA (*Iran Protest Slogan*), https://oversightboard.com/decision/FB-ZT6AJS4X/ (last visited 12 May 2023).

56 UN, 'Outreach Programme on the 1994 Genocide Against the Tutsi in Rwanda and the United Nations', www.un.org/en/preventgenocide/rwanda/historical-background.shtml (last visited 12 May 2023).

57 Oversight Board 2020, 2020-007-FB-UA (*Protest in India Against France*), https://oversightboard.com/decision/FB-R9K87402/ (last visited 12 May 2023).

58 Oversight Board (2021), above n. 52.

59 An independent international anti-corruption body, charged with investigating and complementary prosecuting serious crimes in the country, that derived from a request for assistance by the Government of Guatemala to the UN, which was terminated unilaterally in 2019 by then President Jimmy Morales, who was under investigation by that Commission for campaign financing.

60 OHCHR, 'Press Release: *Guatemala: UN Expert Condemns Targeting of Prosecutor and Judge*' (25 November 2022), www.ohchr.org/en/press-releases/2022/11/guatemala-un-expert-condemns-targeting-prosecutor-and-judge and the Inter-American Human Rights Commission (IACHR) Res. 55/2019, 23 October 2019; IACHR Res. 56/2019, 25 October 2019 and IACHR, 'IACHR Grants Precautionary Measures to Protect Justice Operators in Guatemala', Press release no. 276/19 (28 October 2019), https://www.oas.org/en/iachr/media_center/PReleases/2019/276.asp (last visited 26 April 2023).

61 HRC (2011), above n. 14.

62 For example, research from the Center for Countering Digital Hate found that just 12 people, 'The Disinformation Dozen', are responsible for up to 73% of anti-vaccine content in Facebook. Center for Countering Digital Hate, *The Disinformation Dozen: Why platforms Must Act on Twelve Leading Online Anti-Vaxxers*, https://252f2edd-1c8b-49f5-9bb2-cb57bb47e4ba.filesusr.com/ugd/f4d9b9_b7cedc0553604720b7137f8663366ee5.pdf (last visited 26 April 2023).

doi: 10.5553/ELR.000253

gets of violence and discrimination or in conflict situations – due to the *cumulative effect* of repeated exposure to certain hateful, discriminatory and/or dehumanising expressions accumulating and spreading on a platform. As the Oversight Board noted in the *Alleged Crimes in Raya Kobo* case, while, , they may not cause direct and immediate harm individually, 'when such content appears on an important, influential and popular social media platform' in specific contexts, such as during an ongoing conflict, 'the risk and likelihood of harm become more pronounced'. Moreover, 'cumulative impact can amount to causation through a 'gradual build-up effect', as occurred in Rwanda, where calls for genocide were repeated'.[63]

Likewise, in the *Depiction of Zwarte Piet* case in the Netherlands, the Oversight Board upheld Meta's decision to remove specific content that violated the express prohibition on posting caricatures of Black people in the form of blackface, as stated in its Hate Speech Community Standard. The Board argued that allowing the accumulation of such posts on Facebook 'creates an environment in which acts of violence are more likely to be tolerated and reproduce discrimination in a society'.[64]

Another relevant issue to consider regarding the context analysis is the *difference that the enforcement of the same rules* may have in different groups or in different regions. This is the case with some rules that appear to be neutral, positive or at least not particularly problematic but that can have a disproportionate impact on certain groups or regions. For example, Facebook has a policy that prohibits female nudity. Hence, as a general rule, any post that shows female nipples will be removed. Regardless of the fact that certain societies may or may not agree with the importance and necessity of this rule, the impact it generates in some contexts is entirely disproportionate.

As addressed by the UN Special Rapporteur on freedom of expression, this policy 'may have significant "hyper-local" impacts on [certain] communities'.[65] For example, on some indigenous groups where women live with their trunks uncovered. In those groups, if such women try to denounce abuses or human rights violations they are exposed to through Facebook – which is usually the only means they have to make their voice heard – their posts will be removed, not because of the content of their messages but because they will appear with their trunks – and therefore their nipples –uncovered, as that is part of their culture and how they live.

Along the same lines, in the *Gender identity and nudity* cases, the Oversight Board raised concerns about how Meta's nudity policies disproportionately affect the speech rights of both women and LGBTQI+ users of its platforms. In this regard, the Board stressed that these 'impacts are reflected in both policy and enforcement and limit the ways in which groups can express themselves, resist prejudice and increase their visibility in society', as its policies rely 'on subjective and speculative perceptions of sex and gender that are not practicable when engaging in content moderation at scale'.[66]

Finally, when analysing the context, it is also essential to consider the *differences in access to information in different regions* of the world. In some regions, particularly in the Global North, access to information is generally not limited to a specific platform but rather is accessible through diverse media sources, including digital media. However, this is different in other parts of the world, such as the Global South or Global Majority. This is because many countries provide access to specific social media platforms through a practice called 'zero-rating'. This practice entails an agreement between platforms and mobile operators, ensuring that customers are not charged for the use of data through specific platforms, but have to pay additional fees if they access the Internet outside of those platforms. Consequently, given the economic divide and limited access to platforms free of charge, this makes up a disproportionate part of the customers' access to the Internet. As a result, these platforms become the most important means for people to continue to communicate and inform themselves.[67]

Furthermore, despite the digital media platforms being particularly relevant for the exercise of human rights in these regions, they present a greater risk of harm owing to the lack of invested resources. Therefore, as highlighted by the Oversight Board in various decisions, it is concerning that Community Standards and internal implementation standards are not translated into the different languages of platform users and that there is an insufficient number of reviewers in different markets with a large number of users and higher levels of risk.

For example, in the *Mention of the Taliban in news reporting* decision, the Board expressed its concern 'that the Urdu language queue only had less than 50 reviewers in mid-2022' and noted that 'Meta allocates Urdu reviewers to different workflows based on need. These

63 Oversight Board 2021, 2021-014-FB-UA (*Alleged Crimes in Raya Kobo*), https://oversightboard.com/decision/FB-MP4ZC4CC/ (last visited 12 May 2023), taking up on the Nahimana, Case nº ICTR-99-52-T, at paras. 436, 478 and 484-485.

64 Oversight Board 2021, 2021-002-FB-UA (*Depiction of Zwarte Piet*), https://oversightboard.com/decision/FB-S6NRTDAJ/ (last visited 12 May 2023). Drawing on the UN Special Rapporteur's guidance. HRC (2018), above n. 17, at para. 54. ('The scale and complexity of addressing hateful expression presents long-term challenges and may lead companies to restrict such expression even if it is not clearly linked to adverse outcomes (as hateful advocacy is connected to incitement in Article 20(2) of the ICCPR). Companies should articulate the bases for such restrictions, however, and demonstrate the necessity and proportionality of any content actions.').

65 HRC (2018), above n. 17, at para. 54.

66 Oversight Board 2022, 2022-009-IG-UA and 2022-010-IG-UA (*Gender Identity and Nudity*), https://oversightboard.com/decision/BUN-IH313ZHJ/ (last visited 12 May 2023).

67 Access, 'Policy Brief: Access' Position on Zero Rating Schemes' (11 October 2016), www.accessnow.org/cms/assets/uploads/archive/Access-Position-Zero-Rating.pdf ('These schemes limit user access to those services and applications chosen by dominant tech and telecom companies.... Free expression and access to information depend on access to the full, unfettered internet; anything less harms users' rights.... Zero rating programmes do not provide access to the internet but only to select internet-connected services and applications. These programmes therefore create second-tier users, who can only access a part of the whole internet.').

reviewers are shared across multiple review types, meaning they are not solely dedicated to a single workflow'. Accordingly, the Board considered that 'the size of the Indian market, the number of groups Meta has designated as dangerous in that region, and therefore the heightened importance of independent voices, warrant greater investment from the company'.[68]

Similarly, in the *Reclaiming Arabic words* decision, the Board observed that although Meta has insisted that its reviewers are fluent in English, 'providing [them] with guidance in English on how to moderate content in non-English languages is innately challenging. The [internal guidelines provided to reviewers] are often based in US-English language structures that may not apply in other languages, such as Arabic'.[69]

Therefore, despite the necessity of a global perspective, content moderation in social media must also take the flow of information in different regions and contexts into account. This must be done with special attention to countries or regions most at risk of harm, whether this harm derives from excessive silencing of voices; the proliferation of discriminatory, degrading or inciteful content; or other context-specific reasons.[70]

3 How Should Social Media Platforms be Regulated and Held Accountable?

In recent years, the rise of harmful speech on platforms without sufficient remedies and the growing over-moderation of content by platforms have made the discussion of what content should remain or be removed from those, and who should decide it, increasingly urgent for society.

A general point of international consensus on the subject is that the minimum standards to be enforced are precisely those contained in IHRL. In this regard, in his 2018 report, the UN Special Rapporteur on freedom of expression called on states and ICT companies to apply IHRL, instead of domestic laws or company policies that represent private interests.[71]

However, this raises a number of questions worth reflecting on. First, it raises the question of who should design content norms: a state or a private entity? Although an obvious answer would be that states have the authority to determine this, as the UN Special Rapporteur himself pointed out in the same report:

> National laws are inappropriate for companies that seek common norms for their geographically and culturally diverse user base. But human rights standards, if implemented transparently and consistently with meaningful user and civil society input, provide a framework for holding both States and companies accountable to users across national borders.[72]

While the global nature of platforms and the fact that they transcend borders is a key structural element in determining that it should not be up to states to set the content rules for platforms, another argument also supports this position: the reinforced guarantees of freedom of expression that have been built up over more than 200 years stem precisely from the need to avoid or confront censorship by states.

Related to the foregoing, it is important to bear in mind that according to the latest Freedom House report, world freedom faces a serious threat. This is the 'product of 16 consecutive years of decline in global freedom. A total of 60 countries suffered declines over the past year, while only 25 improved. As of today, some 38 percent of the global population live in Not Free countries, the highest proportion since 1997. Only about 20 percent now live in Free countries'.[73]

Bearing these figures in mind, it is important to consider that throughout history, limiting freedom of expression is usually one of the first measures taken by autocratic or weak democratic governments, particularly the freedom of expression of those who criticise or oppose them. Although political speech is one of the most protected expressions under IHRL,[74] it is the first to be censored by authoritarian regimes. It is therefore very difficult to determine that states should control speech norms when the risk of censorship is so high. Also, the risk is even greater in countries with closed information environments or countries that already face extreme restrictions on freedom of expression and assembly.

Furthermore, the experience of regulatory measures that order the removal of certain content – such as the German NetzDG or the EU Code of Conduct against online hate speech – shows that 'when platforms face legal

68 Oversight Board (2022), above n. 45.

69 Oversight Board (2022), 2022-003-IG-UA (*Reclaiming Arabic Words*), https://oversightboard.com/decision/IG-2PJ00L4T/ (last visited 12 May 2023).

70 This has been addressed by the Oversight Board in different decisions. For examples, Oversight Board (2021), above n. 51; Oversight Board (2022), above n. 45; Oversight Board (2021), above n. 63; Oversight Board (2022), above n. 55; Oversight Board 2021, 2021-010-FB-UA (*Colombia Protests*), https://oversightboard.com/decision/FB-E5M6QZGA/ (last visited 12 May 2023); Oversight Board 2022, 2022-002-FB-UA (*Sudan Graphic Video*), https://oversightboard.com/decision/FB-AP0NSBVC/ (last visited 12 May 2023); Oversight Board 2022, 2022-006-FB-MR (*Tigray Communications Affairs Bureau*), https://oversightboard.com/decision/FB-E1154YLY/ (last visited 12 May 2023); Oversight Board 2022, 2022-008-FB-UA (*Russian Poem*), https://oversightboard.com/decision/FB-MBGOTVN8/ (last visited 12 May 2023); Oversight Board 2022, 2022-011-IG-UA (*Video After Nigeria Church Attack*), https://oversightboard.com/decision/IG-OZNR5J1Z/ (last visited 12 May 2023).

71 That is, not only in the creation of content rules but also in due diligence assessments of how platforms' designs affect human rights, and the establishment of remedies for those harmed by their decisions. HRC (2018), above n. 17.

72 *Ibid.*, at para. 41.

73 Freedom House, 'Freedom in the World 2022. A Global Expansion of Authoritarian Rule' (February 2022), https://freedomhouse.org/sites/default/files/2022-02/FIW_2022_PDF_Booklet_Digital_Final_Web.pdf (last visited 12 May 2023).

74 HRC (2011), above n. 14.

doi: 10.5553/ELR.000253

risk for user speech, they routinely err on the side of caution and take it down'.[75] In other words, the way platforms have coped with these types of laws is by over-enforcing the relevant rules, to avoid false negatives that could generate liability for them, regardless of the impacts it creates on freedom of expression.

Nonetheless, as suggested by the UN Special Rapporteur on freedom of expression, when platforms create norms, they should apply IHRL, with a meaningful multi-stakeholder approach to their development. They should also consider the UNGPs, which 'establish principles of due diligence, transparency, accountability and remediation that limit platform interference with human rights through product and policy development'.[76]

The second question that stems from the application of IHRL to content moderation on platforms would be, should these rules be enforced by states or private entities? Here, although there may be those who would advocate that it should be the state in order to have the guarantees and due process resources offered by democracy, the material capacity of states to effectively control all the content on the platforms should be considered. Both the volume of content and the speed at which it moves on platforms, and the possibilities that state institutions have of exercising effective control over it, need to be taken into account.

However, does this mean the absence of state regulation as a desiderata? Perhaps the question is not whether or not there should be state regulation but rather what it should be about: content rules or the way in which the platforms design and enforce them? Considering both the need to establish external controls on platforms, as well as the unintended impacts and censorship risks implicit in state regulation of content rules, I find the answer to be that states should regulate transparency, accountability and perhaps the minimum procedural guarantees on the platforms but never content rules. Transparency regulations should include design decisions, metrics and results of the enforcement of their rules, data protection, products and investments and financial revenues received through different means. They should also include human rights due diligence assessments and measures to be taken or that have already been taken to mitigate adverse human rights impacts of their operations, as stated by the UNGPs.

This approach is similar to much of what has been proposed in the EU's Digital Services Act.[77] However, it is important to bear in mind that its implementation is still pending, which is of enormous complexity – both for the platforms and the authorities – and that many of its main elements and scope have not yet been defined - among others, the certification and functioning of the independent redress mechanisms provided for in Article 21.

In this context, it would seem that in a complementary manner, a decision such as the creation of the Oversight Board, i.e., an independent mechanism for self-regulation of private entities, may be useful. Beyond the establishment of specific obligations to platforms in certain jurisdictions, the existence of such an oversight body, which is not fragmented territorially, but analyses the global operation of the platforms,[78] can contribute to a better understanding of the way these platforms operate, to achieve greater accountability and to better guarantee users' rights. Undoubtedly, self-regulatory mechanisms will differ according to the characteristics of each platform, but the fundamental characteristic that should be replicated is that of independence.

4 Meta's Oversight Board

The Oversight Board is a self-regulatory mechanism created by Meta but with guarantees of independence. Its purpose is to protect freedom of expression by making independent and principled decisions on important content on Facebook and Instagram and issuing recommendations on Meta's content policies.[79] Hence, its ultimate goal is to contribute to increasing Meta's levels of transparency and congruence, thus making the company more accountable for how it moderates content and how it decides what content is available through its platforms.

Currently, the Board is composed of 23 members[80] from 18 different countries from all regions of the world, speaking more than 27 languages, with diverse professional, cultural, political and religious backgrounds and points of view. Due to the fact that Meta is a global company, regional diversity in Board members is fundamental.

The members of the Board are independent of Meta and have institutional, functional and financial guarantees of independence.[81] Members are hired directly by the Oversight Board for a fixed period,[82] are not employed by Meta and cannot be removed by Meta. The payment of their remuneration is based on the fulfilment of their duties and 'will not be conditioned or withheld based on the outcome of board decisions'.[83] Financial independence is guaranteed by the establishment of an irrevoca-

75 D. Keller, 'Internet Platforms. Observations on Speech, Danger, and Money', *Hoover Institution Aegis Series Paper* (2018), www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf (last visited 12 May 2023). A similar analysis can be found in Douek (2021), above n. 44.

76 HRC (2018), above n. 17, at para. 41.

77 EP and Council Regulation 2022/2065, OJ 2022 L 277/1.

78 Recognising that the Oversight Board's scope of action is limited, as it is not charged with the oversight of Meta's platforms' functioning as a whole.

79 Oversight Board Charter, https://oversightboard.com/governance/ (last visited 12 May 2023).

80 Though the Board may grow up to 40 members. Oversight Board Charter, Art. 1, Section 1.

81 Similar to those of the institutions of the judiciary, although created by a private company and not a state. UN, 'Basic Principles on the Independence of the Judiciary', www.ohchr.org/en/instruments-mechanisms/instruments/basic-principles-independence-judiciary (last visited 12 May 2023); GA res. 40/32 29 November 1985; GA res. 40/146 (13 December 1985).

82 Oversight Board Charter, above n. 79, Art. 1, Section 3.

83 *Ibid.*, Art. 1, Section 5.

doi: 10.5553/ELR.000253

ble trust fund, independent of Meta. All of this is designed to protect Board members' independent judgment and to allow decisions that are free from influence or interference and that are without regard to the economic, political or reputational interests of the company.

In terms of its scope of action, the Board will review individual cases referred by both users and Meta.[84] Regardless, Meta may also request policy advisory opinions from the Board, which may relate to clarification of a previous Board decision or guidance on Meta's content policies. The Board has the sole authority to accept or reject cases and requests referred through these processes.[85]

At the start of its operations in October 2020, users could only appeal to the Board in cases where Meta had removed their content from Facebook and Instagram, after exhausting the company's internal appeal mechan-isms. As of April 2021, its scope of action has been extended to user appeals to remove content. In October 2022 it acquired the power to make binding decisions to enforce a warning screen on content. Recently, in February 2023, the Board announced that in addition to its standard decisions and policy advisory opinions, it will also review expedited[86] and summary[87] decisions.

In addition to the binding decisions it issues on specific cases – to remove, restore and now to enforce warning screens on content – which, according to the Charter Meta must implement unless their enforcement may violate the law, the Board may also make policy recommendations to Meta based on its decisions on specific cases or policy advisory opinions. When such recommendations are made, Meta is not obliged to comply with them but does have the obligation to respond to them publicly.

Since the Board began accepting appeals in October 2020, it has issued 35 case decisions,[88] as well as two policy advisory opinions. As part of this work, it has made 186 recommendations to Meta. The cases decided by the Board have addressed the following policies: Hate Speech (13 cases), Dangerous Individuals and Organisations (7), Violence and Incitement (6), Adult Nudity and Sexual Activity (2), Sexual Solicitation and Adult Nudity and Sexual Activity (2), Regulated Goods (2), Violent and Graphic Content (2), Bullying and Harassment (1) and Sexual Exploitation of Adults (1). In 26 of the cases, the Board decided to overturn Meta's decision. Nonetheless, in 12 of them the company itself changed its outcome after the Board selected the case, and in 9 it upheld Meta's decision. In one case it upheld the original decision, which was later changed by Meta. Its policy advisory opinions have dealt with sharing private residential information on the platform when it is considered 'publicly available',[89] and with Meta's cross-check programme.[90] As for the recommendations issued, they have addressed very diverse topics, such as the use of automation in enforcement; expanding transparency reporting;[91] and the implementation of internal audit procedures, accuracy assessments on specific policies and human rights due diligence assessments – both independent and internal. Other examples are the additional information that should be provided to users whose content has been removed or who have reported content, the translation of public and internal rules into different languages, Meta's processes for assessing context, including 'at escalation',[92] the development of certain policies, the treatment of violating messages from political leaders and other influential users and the need to provide users with the opportunity to appeal to the Board any decisions made through Meta's internal escalation process.[93]

84 The latter can include many types of significant and difficult cases, including accounts, advertising or Groups, among others. Cases are considered *significant* when the content in question has real-world implications and raises serious, large-scale or important issues for public discourse and *difficult* when the content raises questions about current policies or their enforcement, with compelling arguments for removing or keeping the content under review.

85 Oversight Board Charter, above n. 79, Art. 2, Section 1.

86 Expedited decisions will review Meta's decision on content within days in urgent cases. However, it is important to note that the Oversight Board was not created to prevent or respond quickly to content issues in real time and that the Board's ability to hear expedited cases does not eliminate Meta's responsibility to act first and quickly in these situations.

87 Summary decisions will review Meta's original decision in cases where it has subsequently changed its mind. As in the course of the almost 3 years that the Board has been operating, it has been frequent that in reviewing pre-selected cases, Meta has changed its initial decision – either to remove or keep up content. The *Breast Cancer Symptoms and Nudity* case was the first occasion in which the Board decided to hear one of these *enforcement error* cases. Although Meta argued that the Board should recuse itself from hearing the case because the issue was already moot, the Board disagreed and argued that the Charter only requires 'disagreement between the user and [Meta] at the moment the user exhausts [Meta's] internal process. This requirement has been met. The Board's review process is separate from, and not an extension of [Meta's] internal appeals process. For [Meta] to correct errors the Board brings to its attention *and thereby exclude cases from review* would integrate the Board inappropriately to [Meta's] internal process and undermine the Board's independence.' Oversight Board (2020), above n. 35.

88 The Board cannot make a decision on the entirety of user appeals it receives, and, therefore, it prioritises the most significant and relevant cases that may be emblematic of structural problems, may affect many users, are of vital importance to public discourse, raise questions about Meta's policies, or transcend issues that are occurring offline. To select them, the Board created a Case Selection Committee, and once selected, the cases are assigned to a five-member panel, which will always include at least one member from the region involved in the content and a mixed gender representation.

89 Oversight Board 2021, PAO-2021-01 (*Policy Advisory Opinion on Sharing Private Residential Information*), https://oversightboard.com/decision/PAO-2021-01/ (last visited 12 May 2023).

90 Oversight Board (2021), above n. 12.

91 Transparency on: automation; Facebook's Community Standards and Instagram's Community Guidelines, the exceptions to its policies and the allowances it applies; the enforcement of specific policies; removal and error rates per language, country and policy; how Meta collects, preserves and shares information to assist in investigation of grave violations of international law; the strikes and penalties process; account restrictions; government requests and their outcome, distinguishing those based on violations of platform rules, local law and requests that led to no action; fact-checking; escalation procedures, among others.

92 Decisions 'at escalation' are those made by Meta's internal, specialist teams rather than through the 'at scale' content review process. Oversight Board (2022), above n. 49.

93 As stated by the Board in the *UK Drill music* decision, when 'Meta takes a content decision "at escalation", users are unable to appeal the decision to the company or to the Board.... Decisions made at escalation are likely to be among the most significant and difficult, where independent over-

doi: 10.5553/ELR.000253

More specifically, in the *Policy advisory opinion on Meta's cross-check programme* the Board's recommendations[94] focused on prioritising the protection of expression important for human rights rather than business interests, radically increasing transparency around cross-check and how it operates and reducing and mitigating harm caused by content left up during enhanced review.

In response to the Board's recommendations,[95] Meta has adopted various measures, including the following: created a new Community Standard on misinformation, which includes health misinformation,[96] adopted a Crisis Policy Protocol to govern its responses to crises,[97] translated its rules into 15 Asian and African languages, including Farsi, Hausa and Punjabi,[98] released the findings of an independent due diligence report on the impact of the company's policies in Israel and Palestine during the May 2021 conflict,[99] reformed its penalties and strikes system,[100] added text about its satire exceptions across several Community Standards and provided further information on how users can make the intent behind their posts clear,[101] launched new notifications globally that detail specific policy violations on different Community Standards,[102] rolled out new messaging in certain locations telling people whether automation or human review resulted in their content being removed, updated its automatic nudity detection models to account for health contexts,[103] engaged in a policy development process of its Dangerous Individuals and Organizations policy,[104] and initiated an in-depth policy review of its approach to preventing the identification of victims of sexual violence.[105]

Although many of the Board's recommendations have not been accepted by Meta, or have not yet been implemented, the company has committed to removing the exception that allows the sharing of private residential information when it is considered 'publicly available', ensuring the protection of users' privacy in Facebook and Instagram,[106] providing new information both on government requests and its newsworthiness allowance in its transparency reporting, and recently, implementing some of the Board's recommendations on its cross-check programme.[107]

Moreover, there are six points worth highlighting that are a result of the Board's work, as they are relevant to the issues that have been addressed in this article.[108]

sight is at its most important. The Board's governing documents provide that all content moderation decisions that are within scope and not excluded by the Bylaws (Bylaws Article 2, Sections 1.2, 1.2.1) and that have exhausted Meta's internal appeal process (Charter Article 2, Section 1) be eligible for people to appeal to the Board'. *Ibid*.

94 After highlighting flaws in key areas of the programme, which should be addressed by the company: unequal treatment of users, delayed removal of violating content, failure to track core metrics, and lack of transparency around how cross-check works. Oversight Board (2021), above n. 12.

95 To follow up on these responses, the Board established an Implementation Committee, which has led efforts on sharpening the Board's recommendations and ensuring they are focused on specific, measurable impacts. Oversight Board, '2021 Annual Report' (2021), https://oversightboard.com/news/322324590080612-oversight-board-publishes-first-annual-report/ (last visited 12 May 2023).

96 In response to the Board's recommendation that Meta 'set out a clear and accessible Community Standard on health misinformation, consolidating and clarifying existing rules in one place (including defining key terms such as misinformation)'. Oversight Board 2020, 2020-06-FB-FBR (*Claimed COVID Cure*), https://oversightboard.com/decision/FB-XWJQBU9A/ (last visited 12 May 2023).

97 Following the Board's recommendation that Meta 'develop and publish a policy that governs [its] response to crises or novel situations where its regular processes would not prevent or avoid imminent harm'. Oversight Board 2021, 2021-001-FB-FBR (*Former President Trump's Suspension*), https://oversightboard.com/decision/FB-691QAMHJ/ (last visited 12 May 2023).

98 As in different case decisions, the Board has urged Meta to translate its Community Standards into all languages widely spoken by its users. The increase in the languages to which its rules have been translated has resulted in around 800 million people in Global Majority countries can now read Meta's rules in their native language. Oversight Board (2021), above n. 49; Oversight Board (2022), above n. 69; Oversight Board, 'Oversight Board Q2 2022 Transparency Report' (2022), https://oversightboard.com/news/784035775991380-oversight-board-publishes-transparency-report-for-second-quarter-of-2022-and-gains-ability-to-apply-warning-screens/ (last visited 12 May 2023).

99 As recommended in the *Shared Al Jazeera Post* decision. Oversight Board (2021), above n. 51.

100 Which focuses on explaining why content has been removed, provides greater transparency about the system and its penalties and is fairer to users who have been disproportionately impacted in the past. Oversight Board, 'Oversight Board Response to Meta's Announcement on Reforming Its Penalty System' (February 2023), https://oversightboard.com/news/507876928181835-oversight-board-response-to-meta-s-announcement-on-reforming-its-penalty-system/ (last visited 12 May 2023).

101 As recommended by the Board in various case decisions. Oversight Board 2021, 2021-005-FB-UA (*'Two Buttons' Meme*), https://oversightboard.com/decision/FB-RZL57QHJ/ (last visited 12 May 2023); Oversight Board (2021), above n. 40; Oversight Board 2020, 2020-005-FB-UA (*Nazi Quote*), https://oversightboard.com/decision/FB-2RDRCAVQ/ (last visited 12 May 2023); Oversight Board (2020), above n. 57.

102 Initially for the Hate Speech, Dangerous Individuals and Organizations, and Bullying and Harassment policies, though Meta informed it was working to expand the messaging to all Community Standards and to multiple languages. Oversight Board, 'Oversight Board Q3 2022 Transparency Report' (2022), https://oversightboard.com/news/1147590722555454-oversight-board-publishes-transparency-report-for-third-quarter-of-2022/ (last visited 12 May 2023).

103 Both in response to the different recommendations in the *Breast Cancer Symptoms and Nudity* decision. Oversight Board (2020), above n. 35.

104 Specifically, on how it assesses whether content amounts to 'praise', 'substantive support' or 'representation' of a designated individual or organisation, in response to recommendations made in various case decisions. For example, Oversight Board (2020), above n. 101; Oversight Board (2021), above n. 40; Oversight Board (2021), above n. 51; Oversight Board (2022), above n. 45.

105 In response to the Board's recommendation that Meta 'undergo a policy development process, including as a discussion in the Policy Forum, to determine whether and how to incorporate a prohibition on functional identification of child victims of sexual violence in its Community Standards'. Oversight Board 2021, 2021-016-FB-FBR (*Swedish Journalist Reporting Sexual Violence Against Minors*), https://oversightboard.com/decision/FB-P9PR9RSA/ (last visited 12 May 2023).

106 Oversight Board (2021), above n. 89.

107 Oversight Board (2021), above n. 12; Meta, 'Oversight Board Selects a PAO on Meta's Cross-Check Policies' (24 April 2023), https://transparency.fb.com/es-la/pao-cross-check-policy/ (last visited 12 May 2023).

108 Although these do not offer a comprehensive overview of all of the Board's work since 2020. Nonetheless, different academic articles have analysed both the creation and the functioning of the Oversight Board. These offer diverse perspectives of analysis. Among others, J. Barata, 'The Decisions of the Oversight Board from the Perspective of International Human Rights Law' (2022), https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2022/10/The-Decisions-of-the-OSB-from-the-Perspective-of-Intl-Human-Rights-Law-Joan-Barata-.pdf; K. Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression', 129 *Yale Law Journal* 2418 (2020), www.

doi: 10.5553/ELR.000253

The first point concerns the standards by which the Board reviews Meta's decisions. While the Charter expressly states that the Board 'will review content enforcement decisions and determine whether they were consistent with Meta's content policies and values', as well as in reviewing decisions, the Board 'will pay particular attention to the impact of removing content in light of human rights norms protecting free expression',[109] since it issued its first decision in January 2021, the Board decided to base its analysis on a broader set of rights under IHRL. This has been the case even before Meta announced its commitment to respecting human rights standards in line with the UNGPs – which encompasses internationally recognised human rights as defined, among other instruments, by the ICCPR – embodied in a new corporate policy launched in March 2021.[110] Thus, in all the Board's decisions, it has reviewed whether Meta's decisions are consistent with Facebook's and Instagram's policies and values, as well as with the company's commitment to upholding the right to freedom of expression within the framework of international human rights standards. In doing so, it has used the three-part test in Article 19(3) of the ICCPR, as well as other instruments of both treaty and soft law,[111] to interpret Meta's voluntary human rights commitments, both for the individual content decision and for what this says about Meta's broader approach to content governance.[112]

The second point to highlight is that although the Board's scope of action relates to decisions adopted in specific cases, the investigation of those cases has not been limited to determining the reasons why a given decision made by Meta was right or wrong – in light of its policies and values, as well as IHRL. It has also inquired into the design and functioning of the system and factors external to the decision itself – such as the use of automation, the company's processes and the involvement of governments – that led to such errors. It has also made recommendations for greater transparency in this regard.

For example, even though Meta asked the Board to 'focus on the outcome of enforcement, and not the method'[113] in the *Breast Cancer Symptoms and Nudity* case, both the investigation and recommendations in that case focused on the system – which was automated and potentially without human review or appeal – that led to the adverse outcome, and not just the specific decision in the particular case, which Meta had already acknowledged was incorrect.[114]

Likewise, when addressing the issue of cross-check system[115] and the newsworthiness allowance[116] in the analysis of *Former President Trump's Suspension* case, the Board noted that there was limited public information available regarding the system and the allowance and that this was relevant because 'different processes may lead to different substantive outcomes'.[117]

Furthermore, in the *Öcalan's isolation*, *Shared Al Jazeera post* and *UK drill music* case decisions, the Board made different recommendations to Meta to provide greater transparency on the governmental requests it receives, distinguishing those based on infringements of community rules, local legislation and requests that did not lead to any action.[118]

In this regard, a third point to highlight is that through its decisions, and in the investigation undertaken in connection with them, the Board has requested more information from Meta than what was publicly available on the operation of its systems, designs, processes, pol-

yalelawjournal.org/feature/the-facebook-oversight-board (last visited 12 May 2023); E. Douek, 'Facebook's 'Oversight Board:' Move Fast with Stable Infrastructure and Humility', 21 *North Carolina Journal of Law & Technology* 1 (2019), https://scholarship.law.unc.edu/ncjolt/vol21/iss1/2/ (last visited 12 May 2023).

109 Oversight Board Charter, above n. 79, Art. 2, Section 2.

110 As stated in Meta's Corporate Human Rights Policy, its 'commitment encompasses internationally recognized human rights as defined by the International Bill of Human Rights —which consists of the Universal Declaration of Human Rights; the International Covenant on Civil and Political Rights; and the International Covenant on Economic, Social and Cultural Rights— as well as the International Labour Organization Declaration on Fundamental Principles and Rights at Work'. MetaCorporate Human Rights Policy, https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf (last visited 12 May 2023).

111 Such as the Rabat Plan of Action, UN Committees' General Comments and Recommendations, and reports of the UN Special Rapporteur on freedom of expression.

112 Oversight Board, 'Case Decisions and Policy Advisory Opinions', https://oversightboard.com/decision/ (last visited 12 May 2023).

113 Oversight Board (2020), above n. 35.

114 E. Douek, 'The Facebook Oversight Board's First Decisions: Ambitious, and Perhaps Impractical' (28 January 2021), www.lawfareblog.com/facebook-oversight-boards-first-decisions-ambitious-and-perhaps-impractical ('This decision [Breast Cancer Symptoms and Nudity] sends a strong shot across the bow to Facebook. The board is establishing that it will not limit its view to just the outcomes in the cases before it, but will interrogate the systems that led to them.').

115 The 'cross-check program' is a system that provides 'additional layers of human review for certain posts initially identified as breaking [the platforms'] rules. When users on Meta's cross-check lists post such content, it is not immediately removed as it would be for most people, but is left up, pending further human review'. Meta describes it 'as a mistake-prevention strategy that allows it to balance protecting users' voice from false positives with the need to quickly remove violating content'. Oversight Board (2021), above n. 12.

116 Through its 'newsworthiness allowance' Meta allows content that violates its policies to remain on the platform if it determines that it is newsworthy and 'keeping it visible is in the public interest [and] after conducting a balancing test that weighs the public interest against the risk of harm'. Oversight Board (2022), above n. 70.

117 Oversight Board (2021), above n. 97. After this decision, the Board has insisted on the need for more transparency, both in the cross-check program and the newsworthiness allowance. Oversight Board (2021), above n. 12; Oversight Board (2021), above n. 70; Oversight Board (2022), above n. 70 and Oversight Board (2022), above n. 55.

118 Oversight Board (2021), above n. 40; Oversight Board (2021), above n. 51 and Oversight Board (2022), above n. 49. Although in the *UK Drill Music* decision the Board acknowledged Meta 'has made progress in relation to transparency reporting since the Board's first decisions addressing this topic' – which 'includes conducting a scoping exercise on measuring content removed under the Community Standards following government requests, and contributing to Lumen, a Berkman Klein Center for Internet & Society research project on government removal requests', it further recommended that 'Meta should create a section in its Transparency Center, alongside its "Community Standards Enforcement Report" and "Legal Requests for Content Restrictions Report", to report on state actor requests to review content for Community Standard violations. It should include details on the number of review and removal requests by country and government agency, and the numbers of rejections by Meta'.

doi: 10.5553/ELR.000253

icies and enforcement decisions. In most cases the company provided the requested information, but even when it did not, the Board publicised its request and the lack of compliance by the company.

By doing so, the Board has publicised information that previously did not exist in the public domain, so that users, researchers, civil society organisations, academics and any other interested person can get a better understanding of how the company works, its rules, exceptions and allowances, and its different processes and systems. Likewise, in different recommendations it has made, the Board has asked Meta to incorporate more and clearer information both in its Community Standards and in its Transparency Reports.

For example, during the investigation in the *Breast Cancer Symptoms and Nudity* case the Board identified that, although not communicated to Instagram users, Instagram's Community Guidelines are interpreted in line with Facebook's Community Standards and that, in case of inconsistency, the latter prevails. It recommended Meta to clarify this in its public policies.[119]

Moreover, although Meta did not mention cross-checking in its initial referral or in materials sent to the Board in the *Former President Trump's Suspension* case, it described this programme in response to a question from the Board about any different treatment the account may have received.[120] However, following documentation on this system disclosed by the *Wall Street Journal* based on revelations by former employee and company critic Frances Haugen, and after the Board called on Meta to commit to making this system transparent, the company sent the Board a request for a policy advisory opinion. As a result, the Board made public the way this system works and made 32 recommendations to Meta regarding the structuring of the system, both to meet Meta's human rights commitments and to address the problems identified by the Board.[121]

Also related to the *Former President Trump's Suspension* case, the decision expressly notes that although the Board sought clarification from Meta on 'the platform's design decisions, including algorithms, policies, procedures and technical features, amplified Mr. Trump's posts after the election and whether [Meta] had conducted any internal analysis of whether such design decisions may have contributed to the events of January 6', Meta declined to answer these questions.[122]

Similarly, in the *Punjabi concern over the RSS in India* case, the Board noted that Meta refused to provide specific answers to its questions 'regarding possible communications from Indian authorities to restrict content around the farmer's protests, content critical of the government over its treatment of farmers, or content concerning the protests', as it determined that the requested information was not reasonably required for decision-making.[123]

Fourth, the UN Special Rapporteur on freedom of expression, although 'companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression'.[124] Following on that, the Board has considered that 'clarifying the nature of those questions and adjudicating whether [Meta's] answers fall within the zone of what the UN Guiding Principles require, is the principal task facing this Board.'[125]

Specifically, in the *Depiction of Zwarte Piet*, *Armenians in Azerbaijan*, *South Africa Slurs* and *Knin cartoon* cases, the Board addressed the *cumulative effect* of repeated exposure to certain hateful, discriminatory and/or dehumanising expressions accumulating and spreading on a platform. In particular, the Board noted that moderating content or prohibiting 'some discriminatory expression',[126] 'to address the cumulative harms of hate speech, even where the expression does not directly incite violence or discrimination, can be consistent with [Meta`s] human rights responsibilities in certain circumstances',[127] when 'left up, an accumulation of such content may create an environment in which acts of discrimination and violence are more likely'.[128] The Board also concluded that 'the human rights responsibilities of Meta as a company differ from the human rights obligations of states. Meta can apply less strict standards for removing content from its platform than those which apply to states imposing criminal or civil penalties'.[129] In these cases, the analysis of the context played an essential role to demonstrate the necessity and proportionality of those actions.[130]

The fifth highlight stems from Meta's commitment to respecting human rights as set out in the UNGPs, which state that companies should conduct human rights due diligence to assess the impact of their activities (Principle 17). In different cases, the Board has recommended Meta to conduct human rights due diligence assessments in specific regions and situations to improve its policies and platform design. It has also recommended that Meta's development of its policies should include a comprehensive human rights impact assessment, with broad and inclusive stakeholder engagement.

More specifically, in the 'Former President Trump's Suspension' decision, the Board concluded that '[w]hen [Meta's] platform has been abused by influential users in a way that results in serious adverse human rights im-

119 Oversight Board (2020), above n. 35. This recommendation was reiterated in the *Öcalan's Isolation* and *Ayahuasca Brew* decisions. Oversight Board (2021), above n. 40; Oversight Board 2021, 2021-013-IG-UA (*Ayahuasca Brew*), https://oversightboard.com/decision/IG-0U6FLA5B/ (last visited 12 May 2023).

120 Oversight Board (2021), above n. 97.

121 Oversight Board (2021), above n. 12.

122 Oversight Board (2021), above n. 97.

123 Oversight Board (2021), above n. 49.

124 HRC (2019), above n. 24.

125 Oversight Board 2020, 2020-003-FB-UA (*Armenians in Azerbaijan*), https://oversightboard.com/decision/FB-QBJDASCV/ (last visited 12 May 2023).

126 Oversight Board (2021), above n. 52.

127 Oversight Board (2021), above n. 64.

128 Oversight Board (2020), above n. 125.

129 Oversight Board 2022, 2022-001-FB-UA (*Knin Cartoon*), https://oversightboard.com/decision/FB-JRQ1XP2M/ (last visited 12 May 2023).

130 Drawing up on the UN Special Rapporteur's guidance. HRC (2019), above n. 24, at para. 48.

pacts, it should conduct a thorough investigation into the incident. [Meta] should assess what influence it had and assess what changes it could enact to identify, prevent, mitigate, and account for adverse impacts in future. In relation to this case, [Meta] should undertake a comprehensive review of its potential contribution to the narrative of electoral fraud and the exacerbated tensions that culminated in the violence in the United States on January 6, 2021. This should be an open reflection on the design and policy choices that [Meta] has made that may enable its platform to be abused. [Meta] should carry out this due diligence, implement a plan to act upon its findings, and communicate openly about how it addresses adverse human rights impacts it was involved with'.[131]

Similarly, in the *Shared Al Jazeera post* decision, the Board recommended that Meta should '[e]ngage an independent entity not associated with either side of the Israeli-Palestinian conflict to conduct a thorough examination to determine whether [Meta's] content moderation in Arabic and Hebrew, including its use of automation, have been applied without bias. This examination should review not only the treatment of Palestinian or pro-Palestinian content, but also content that incites violence against any potential targets, no matter their nationality, ethnicity, religion or belief, or political opinion. The review should look at content posted by [Meta] users located in and outside of Israel and the Palestinian Occupied Territories'.[132]

Furthermore, in the *Alleged Crimes in Raya Kobo* decision, the Board recommended that Meta 'commission an independent human rights due diligence assessment on how Facebook and Instagram have been used to spread hate speech and unverified rumours that heighten the risk of violence in Ethiopia. The assessment should review the success of measures Meta took to prevent the misuse of its products and services in Ethiopia. The assessment should also review the success of measures Meta took to allow for corroborated and public interest reporting on human rights atrocities in Ethiopia. The assessment should review Meta's language capabilities in Ethiopia and if they are adequate to protect the rights of its users. The assessment should cover a period from June 1, 2020, to the present'.[133]

Additionally, in the *Gender identity and nudity* decision, the Board recommended that

> [i]n order to treat all users fairly and provide moderators and the public with a workable standard on nudity, Meta should define clear, objective, rights-respecting criteria to govern the entirety of its Adult Nudity and Sexual Activity policy, ensuring treatment of all people that is consistent with international human rights standards, including without discrimination on the basis of sex or gender identity.

For this purpose, the Board considered

> Meta should first conduct a comprehensive human rights impact assessment to review the implications of the adoption of such criteria, which includes broadly inclusive stakeholder engagement across diverse ideological, geographic and cultural contexts. To the degree that this assessment should identify any potential harms, implementation of the new policy should include a mitigation plan for addressing them.[134]

Finally, being charged with the oversight of a global platform, a fundamental part of the Board's work has to do with stakeholder engagement, both for the identification of the most significant cases and for the acquirement of contextual information that is necessary for the resolution of specific cases and policy advisory opinions.[135] In this regard, the Board permanently holds meetings and roundtables with stakeholders in different regions of the world. This includes not only organisations specialised in digital rights and freedom of expression but also human rights organisations, with the purpose of identifying online problems that have offline effects that should be addressed by the Board.

In addition, every time the Board selects a case for review, it opens a public comment period to allow third parties to share their ideas and perspectives with the Board.[136] This allows for expert perspectives on the cases, as well as the perceptions of different users regarding the impact that certain Meta actions or policies have in specific contexts and regions.

The input provided through public comments often shape some of the recommendations the Board has issued and their impact. For example, in the *Iran protest slogan* case, public comments the Board received confirmed that the 'marg bar Khamanei' slogan 'was being widely used in [the ongoing] protests and online' and 'often included perceptions that Meta over-enforces its policies against Farsi language content during protests'. They also emphasised that '[s]ocial media plays a crucial role in ensuring people in Iran can exercise their rights, particularly in times of protest'. Considering the aforementioned, Meta recently announced that, in response to the Board's recommendation, 'it would allow use of

131 Oversight Board (2021), above n. 97.

132 Oversight Board (2021), above n. 51. In September 2022, Meta published the findings in the independent due diligence report it commissioned following this recommendation, as well as its response to it. BSR, above n. 47; Meta, 'An Independent Due Diligence Exercise into Meta's Human Rights Impact in Israel and Palestine During the May 2021 Escalation' (22 September 2022), https://about.fb.com/news/2022/09/human-rights-impact-meta-israel-palestine/ (last visited 12 May 2023).

133 Oversight Board (2021), *Alleged Crimes in Raya Kobo*, above n. 63.

134 Oversight Board (2021), above note 97; Oversight Board (2021), above n. 51; Oversight Board (2021), above n. 63; Oversight Board (2022), above n. 66.

135 To this end, in October 2022, the Board announced seven strategic priorities that it wants to work on with stakeholders to reform Meta's content moderation approach. This seeks to increase the Board's impact in the areas where it can make the biggest difference to people's experience on Facebook and Instagram.

136 Once the Board selects a case for review, it posts a summary of the case on its website and social media and sets a time frame within which third parties may share their ideas and perspectives with the Board. These are published alongside the Board's decision.

doi: 10.5553/ELR.000253

the slogan in the context of ongoing protests in Iran'. This is likely 'to significantly impact on the ability of protesters in Iran to have their voices heard on Facebook and Instagram', as the public comments provided suggested.[137]

Engaging with external stakeholders has been a permanent activity that the Board has carried out since its foundation. However, it is undoubtedly still necessary to strengthen it, particularly in those countries where Meta invests fewer resources in tools, products and reviewers, considering that it is precisely in those places where there may be greater systemic failures in the enforcement of Meta policies, generating problems of differential treatment of users.

5 Conclusion

There are enormous challenges for content moderation in the digital sphere, both because of the dominant role that private entities now play in the exercise of freedom of expression and because of the global nature of platforms. The speed, reach and large volume of content circulating on the platforms entail different trade-offs that impact people's rights both on- and offline. Although the discussion remains open as to who should regulate social media and how it must be done, there are specific elements that can provide opportunities for platforms, states and regulatory bodies.

In particular, through regulatory, self-regulatory – with independent oversight – and co-regulatory measures, it is important to continue to push platforms to ensure that *i)* their policies, rules and exceptions are clear – in different languages – and respectful of human rights; *ii)* they make their content moderation practices transparent – including the design decisions of their platforms and the tools and products they develop – as well as the relevant metrics; *iii)* they provide effective appeal mechanisms; *iv)* they conduct permanent evaluations – both internal and independent – on the enforcement of their policies and the impact of their activities on human rights with broad stakeholder participation – a multi-stakeholder approach – in accordance with the UNGPs, and *v)* they invest more resources into improving their content moderation – both human and automated.

As for the Oversight Board, as a self-regulatory mechanism, its lasting impact depends not only on the decisions and recommendations it issues but also on its ability to influence Meta's policies and practices. As seen previously, many of its decisions and recommendations have led to important changes and actions by the company, although others have not been accepted by Meta or have yet to be implemented. The Board needs to continue with insisting on their full adoption. Moving forward, there are many different issues and topics that still must be addressed by the Board to push Meta to be more accountable and transparent and for users around the globe to be treated equally and fairly.

Nonetheless, rigorous and independent decisions can have another important effect: to assist regulators and even the courts – not as a precedent, since it lacks judicial authority – but as a relevant doctrine that, if consistent, can also be impactful on the evolution of public law.

137 Oversight Board (2022), above n. 55; Oversight Board, 'Q4 2022 Transparency Report' (2023), https://oversightboard.com/news/943702317007222-oversight-board-announces-plans-to-review-more-cases-and-appoints-a-new-board-member/ (last visited 12 May 2023).

doi: 10.5553/ELR.000253