

Meta Oversight Board's Nascent Standard on Hate Speech: Towards Plural Standard Setting in International Human Rights Law*

Bilyana Petkova

Table of contents

1. Introduction. – 2. International Human Rights Law's procedural standard on hate speech. – 3. International Human Rights Law's developing new standard on hate speech. – 4. The Pluralist Landscape of Offline and Online Standard Setting in IHRL. – 5. Hate Speech as Censorship Creep or Firm Enforcement of Human Rights? – 6. Concluding Remarks

1. Introduction

The main difference between the offline and online world is that freedom of expression in both the European constitutional parlance and freedom of speech in the American constitutional context¹ is now increasingly controlled by algorithms through what is known worldwide as content moderation practices by social media companies. Content moderation is subject to the terms and conditions laid down by private entities, thus establishing a private technocratic regime that controls the protection of freedom of expression online. «By moderating content (or failing to do so), platforms shape public discourse, affect the protection of vulnerable individuals and groups, engage in the balancing of fundamental rights and values against each other, and generally make decisions with far-reaching consequences

* Peer-reviewed article.

This paper presents results as part of the project No. KII-06 AB/6 — The Online/Offline Divide: Can Pre-digital Law Govern the Online World? Focus on freedom of expression at the Meta Oversight Board, funded by the Bulgarian National Science Fund. I am indebted to Oreste Pollicino, Federica Paolucci and the participants in the conference “Rights, Powers and Remedies” held on October 30-31 at Bocconi University in Milan.

¹ Both terms are used interchangeably without prejudice to the fact that a different scope of protection is afforded to freedom of speech in the United States than to freedom of expression in Europe.

in a variety of societal realms».²

The offline protection of freedom of speech and online content moderation have often been discussed as two worlds apart. Where shall the two meet? As Jack Balkin observes: «[t]he speed and scale of digital speech have transformed how speech is governed, regulated, and protected. Social media companies have developed an algorithmic-administrative system for governing speech that does not view speech in terms of rights».³ However, although content moderation at scale remains probabilistic,⁴ international human rights law (IHRL) is becoming its reference point and an authoritative framework. For example, X enlists international soft law like the United Nations Principles on Business and Human Rights as a subsidiary source for guidance in its content moderation practices,⁵ while Meta's Oversight Board goes further by making IHRL central to the rules the Board applies.⁶ Separate from Meta, the Board sits on top of the content moderation hierarchy at the company and hears the most controversial decisions escalated through algorithmic or human intervention. Its structure, various strengths and weaknesses, and first decisions have been analysed.⁷ However, there is no comprehensive study of the Board's overall doctrine. What are the contours of permissible versus impermissible speech that the Board defends, and how do they differ from standards developed for the offline world? The Board, an experiment of self-regulation of online speech governance, has indeed been compared to an international court,⁸ whose developing logic and standards merit scholarly

² P. Ortolani, *Generative AI and Content Moderation* in P. Hacker (ed.) et al., *The Oxford Handbook of the Foundations and Regulation of Generative AI*, Oxford, 2025, forthcoming.

³ J.M. Balkin, *Free Speech Versus the First Amendment* in *UCLA Law Review*, 70, 2023, 1206 ff, 1213.

⁴ R. O'Kane *Meta's Private Speech Governance and the Role of the Oversight Board: Lessons from the Board's First Decisions* in *Stanford Technology Law Review*, 25, 2022, 167 ff, 182.

⁵ *Defending and Respecting the Rights of People Using Our Service*, X, in *help.x.com*.

⁶ The Board should «...pay particular attention to the impact of removing content in light of human rights norms protecting free expression», *Meta Oversight Board Charter*, Art 2, §2, in *oversightboard.com*. The Board's Charter effectively makes international law a third source of law for the Board to apply on a par with Meta's Community Standards and Values. In some sense, international law has even taken precedence over the other sources for the Board since the Values specify that speech that would otherwise violate Meta's Community Standards would remain online if its newsworthy and in the public interest, «judged by international human rights standards», See M. Bickert, *Updating the Values that Inform Our Community Standards*, META, 2019, emphasis added in *about.fb.com*. See also the *Board's Violence Against Women decision*: «...while the second post violated Meta's hate speech Standard, the strongly expressed views in question posed no risk of likely and imminent harm and thus removing it was inconsistent with international human rights standards», in *oversightboard.com*.

⁷ K. Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression* in *Yale Law Journal*, 129, 2020, 2418 ff., D. Wong – L. Floridi, *Meta's Oversight Board: A Review and Critical Assessment* in *Minds and Machines*, 33, 2023, 261 ff. See also A. Tiedeke - M. Fertmann, *A Love Triangle? Mapping Interactions between International Human Rights Institutions, Meta and Its Oversight Board*, in *European Journal of International Law*, 34, 2023, 907 ff. and R. Gulati, *Meta's Oversight Board and Transnational Hybrid Adjudication—What Consequences for International Law?* in *German Law Journal*, 24, 2023, 473 ff.

⁸ L.R. Helfer - M. K. Land, *The Meta Oversight Board's Human Rights Future* in *Cardozo Law*

attention. One of the thorniest questions remains the treatment of hate speech; this article focuses on that.

From a comparative constitutional standpoint, hate speech in the offline world merits radically different treatment in the United States than in Europe. As known, the US Supreme Court has proclaimed that «the proudest boast of our free speech jurisprudence is that we protect the freedom to express ‘the thought that we hate’».⁹ As Claudia Haupt writes, «[b]ut whereas Nazi parades were deemed a necessary evil of free speech protection in the United States, the European postwar consensus had determined that there are limits to democratic public discourse that require a policing of its boundaries».¹⁰ Thus, when it comes to hate speech in certain cases, the European Court of Human Rights (ECtHR) goes as far as to eschew its traditional balancing approach to freedom of expression, declaring certain (often Nazi) speech out of the scope of protection of Article 10 (on freedom of expression) of the European Convention on Human Rights (ECHR).¹¹ The ECtHR moreover finds positive obligations for the state to prevent hate speech under Article 8 ECHR (right to private life) when the violation to reputation reaches a certain threshold of seriousness.¹² Both approaches of the Strasbourg Court have been criticized as overreaching.¹³ At the same time, there is a growing discontent with the American approach at the other end of the spectrum. Tim Wu, a prominent First Amendment scholar, has written: «[s]ome might argue, based on the sophomoric premise that

‘more speech is always better,’ that the current state of chaos is what the First Amendment intended. But no defensible free-speech tradition accepts harassment and threats as speech, treats foreign propaganda campaigns as legitimate debate or thinks that social-media bots ought to enjoy constitutional protection. A robust and unfiltered debate is one thing; corruption of debate itself is another. We have entered a far more dangerous place for the republic; its defense requires stronger protections for what we once called the public sphere».¹⁴ The presumptively unconstitutional hate speech bills recently passed by the states of New York and California¹⁵ can thus be read in the same spirit of protest against the US constitutional law status quo on freedom of speech. Given the differences between the United States and Europe then, can an interpretation of IHRL

Review, 44(6), 2023.

⁹ *Matal v. Tam*, 137 S. Ct. 1744, 1764 (2017) (quoting *United States v. Schwimmer*, 279 U.S. 644, 655 (1929) (Holmes, J., dissenting)).

¹⁰ C.E. Haupt, *Democratic Self-Defense*, in *Fordham Law Review*, 93, 2024, 1389 ff.

¹¹ H. Cannie - D. Voorhoof, *The Abuse Clause and Freedom of Expression in the European Human Rights Convention: An Added Value for Democracy and Human Rights Protection?*, in *Netherlands Quarterly of Human Rights*, 29(1), 2011, 54 ff.

¹² ECtHR, *Minasyan and Others v. Armenia*, app. no. 59180/15 (2025).

¹³ N. Alkiviadou, *Hate Speech, Positive Obligations and Free Speech: The ECtHR’s Expanding Framework in Minasyan and Others v. Armenia (2025)*, in *strasbourghservers.com*, 22 April 2025.

¹⁴ T. Wu, *Opinion, How Twitter Killed the First Amendment* in *New York Times*, 27 October 2017.

¹⁵ *Supra*.

in content moderation provide a bridge that attempts to reconcile these diametrically opposed constitutional law traditions?

A necessary caveat is in order. Social media companies like Meta do not always use IHRL as a framework for their content moderation decisions.¹⁶ This is so since behavioural-based enforcement is increasingly used for a subset of policies like scamming and child safety. Behavioural-based enforcement seeks to avoid abuses of the platform based not on posted content but on the behavioural patterns displayed by users. For instance, users can be sanctioned for their direct and repeated participation in certain communities, the maintenance of several accounts and consumption patterns (for example, when repeatedly searching for violating content). What is moderated in these cases is not content but the overall use of the platform. Although there is a growing interest in this type of enforcement however, it is not used for hate speech cases which are highly contextual and language-specific, and on which this article will focus.

The paper proceeds as follows: Part I provides a snapshot of IHRL on hate speech. I argue that IHRL, influenced by US constitutional law, has developed a very high threshold for proscribing harmful speech – that threshold is heavily connected to incitement to violence, discrimination or hostility understood as imminent danger. Part II examines the Meta Oversight Board full board decisions on hate speech and discerns a novel standard the Board is establishing – let us call it the risk of incitement standard. Part III discusses the legitimacy of novel standards under conditions of pluralist norm setting on the international scene. Part IV is concerned with the treatment of hate speech on a spectrum of censorship and the robust enforcement of human rights. Part V concludes.

2. International Human Rights Law's procedural standard on hate speech

Making IHRL on free speech the norm in content moderation has been both praised¹⁷ and constructively criticized.¹⁸ Among the reasons for adopting IHRL by the platforms that Douek discusses are: giving content moderation legitimacy, providing a single set of global rules, supplying a common vocabulary on online speech for all relevant stakeholders, providing companies with a normative basis for denying authoritarian requests to censor content, constraining platforms' arbitrary exercise of power and – all in all, being the least worst option available in the lack of compelling alternatives.¹⁹ Yet another reason why IHRL might be the way forward is

¹⁶ This section is based on a discussion under Chatham House Rules that took place during the first international conference on the Philosophy of Content Moderation held in California 12-15 April 2025.

¹⁷ D. Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, (2018), U.N. Doc. A/HRC/38/35, 2018.

¹⁸ E. Douek, *The Limits of International Law in Content Moderation*, in *University of California Irvine Journal of International, Transnational, & Comparative Law*, 6, 2021, 37 ff.

¹⁹ *Ibid*, 41-49.

because it could provide a single set of standards for offline and online speech that might avoid a possible divide between how speech is treated online in content moderation and how it is adjudicated offline by the public courts. What is illegal offline would remain illegal online.²⁰ However, while quasi-judicial, quasi-advisory bodies like the Meta Oversight Board accumulate ‘case law’, the more pressing question might become: should what is illegal online be illegal also offline?

There is an international procedural standard on freedom of speech that should guide the content moderation of social media companies that voluntarily subscribe to it. In a series of articles,²¹ academics (some now members of the Meta Oversight Board) and prominent international law practitioners have aimed to clarify that standard.

The interpretation of IHRL given by the authors espouses broad protection for freedom of expression and narrow understanding for hate speech under articles 19 and 20 of the International Covenant on Civil and Political Rights (ICCPR).²² Such an understanding is warranted by authoritative UN interpretative documents.²³ Specifically, first and foremost hate speech, like any other restriction on speech, should be subjected to art. 19(3) tripartite test on legality, legitimacy and necessity. The test applies to hate speech also under art. 4 of the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD).²⁴

²⁰ T. Wischmayer, What is Illegal Offline is Also Illegal Online – The German Network Enforcement Act 2017 in B. Petkova – T. Ojanen (eds.), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries*, Cheltenham-Northampton, 2020, 43 ff.

²¹ E.M. Aswad, The Future of Freedom of Expression Online in *Duke Law & Technology Review*, 17, 2018, 26 ff.; E.M. Aswad, To Protect Freedom of Expression, Why Not Steal Victory from the Jaws of Defeat?, in *Washington & Lee Law Review*, 77, 2020, 609 ff.; E.M. Aswad – D. Kaye, Convergence & Conflict: Reflections on Global and Regional Human Rights Standards on Hate Speech, in *North Western Journal of Human Rights*, 20, 2022, 3 ff.

²² Art. 19 ICCPR reads: «1. Everyone shall have the right to hold opinions without interference. 2. Everyone shall have the right to freedom of expression; this right shall include 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. 3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (ordre public), or of public health or morals». *International Covenant on Civil and Political Rights, Dec. 16, 1966*, in ohchr.org. Art. 20 reads: «1. Any propaganda for war shall be prohibited by law. 2. Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law». Ibid.

²³ U.N. Human Rights Committee, *General Comment No. 34 (2011)*, U.N. Doc. CCPR/C/GC/34 in ohchr.org. The government (or the social media company) restricting hate speech bears the burden of satisfying the cumulative conditions of the test.

²⁴ Art. 4 ICERD reads: «States Parties condemn all propaganda and all organizations which are based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination and, to this end, with due regard to the principles embodied in the Universal Declaration of Human

The three prongs of the test²⁵ are as follows:

1. Legality – laws should be appropriately enacted within their respective legal system; they should give appropriate notice to the ones affected and should not be improperly vague.
2. Legitimacy – the restriction to speech needs to be enumerated in Article 19.3 ICCPR, thus it should be limited to: protecting the rights or reputations of others, national security, public order, public health, or morals (where morals should be read «in the light of the universality of human rights»).
3. Necessity – the restriction must be proportionate and the least restrictive means to achieve the aim.

So far, so good. It is far less clear how or whether on top of the tripartite test there should be relayed additional requirements laid down in the 2012 Rabat Plan of Action which identifies six factors that states should consider when assessing the lawfulness of criminal sanctions and arguably – in content moderation – private companies should apply to speech that is taken down.²⁶ The six factors are: 1) social and political context of the speech 2) the status of the speaker 3) the intent of the speaker (where negligence and recklessness would not suffice) 4) the content and form of the speech 5) the reach of the speech; and 6) the likelihood of harm, including its imminence.²⁷ In addition, a former UN rapporteur on freedom of opinion and expression has stated «three threshold elements for speech to qualify for prohibition under ICCPR Article 20: intent, incitement, and particular harms».²⁸ Whereas the elements of intent and harms arguably match the Rabat Plan factors mentioned above, the element of incitement is in the very text of Article 20 ICCPR; it is said to obtain only when the speech in question «create[s] an imminent risk of discrimination, hostility or

Rights and the rights expressly set forth in article 5 of this Convention, inter alia: (a) Shall declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin, and also the provision of any assistance to racist activities, including the financing thereof; (b) Shall declare illegal and prohibit organizations, and also organized and all other propaganda activities, which promote and incite racial discrimination, and shall recognize participation in such organizations or activities as an offence punishable by law; (c) Shall not permit public authorities or public institutions, national or local, to promote or incite racial discrimination». *International Convention on the Elimination of All Forms of Racial Discrimination Dec. 21, 1965*, in ohchr.org. See also Committee on the Elimination of Racial Discrimination, *General Recommendation No. 35, Combatting Racist Hate Speech (2013)*, U.N. Doc. CERD/C/GC/35, in refworld.org.

²⁵ For a detailed analysis of each of the prongs, see E.M. Aswad – D. Kaye, *Convergence & Conflict*, cit.

²⁶ U.N. High Commissioner for Human Rights, *Report on the Expert Workshop on the Prohibition of Incitement to National, Racial, or Religious Hatred, A/HRC/22/17/Add.4 (2012) (Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that Constitutes Incitement to Discrimination, Hostility, or Violence)*, in documents.un.org, 2013.

²⁷ *Ibid.*, 11.

²⁸ F. La Rue, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2012)* U.N. Doc. A/67/357 quoted in E.M. Aswad – D. Kaye, *Convergence & Conflict*, cit., 177.

violence against persons belonging to those [targeted] groups».²⁹ Another UN report further states that freedom of expression cannot be restricted unless it can be demonstrated that (a) the expression is intended to incite *imminent* violence; (b) it is *likely* to incite such violence; and (c) there is a direct and immediate connection between the expression and the likelihood or occurrence of such violence.³⁰ Yet another instantiation on the UN level of the “incitement to imminent violence” referral is Human Rights Council Resolution 16/18³¹ which attempts to tackle religious intolerance by avoiding blasphemy bans. Much like in the influential US constitutional law case of *Brandenburg v. Ohio*³², incitement in IHRL is thus linked to a «*real and imminent danger of violence* resulting from the expression».³³

The standard of incitement set by the US Supreme Court in *Brandenburg v. Ohio* is the culmination of fifty years of separate opinions. In *United States v. Schenck*, the US Supreme Court introduced the clear and present danger test for the first time. However, the test was best articulated later in Justice Holmes’s dissent in *Abrams v. United States*, where the Justice stated that «only the present danger of immediate evil or an intent to bring it about» should allow Congress to restrict free speech. In his concurring opinion in *Whitney v. California*, Justice Brandeis clarified that «...even advocacy of violation, however reprehensible morally, is not a justification for denying free speech where the advocacy falls short of incitement and there is nothing to indicate that the advocacy would be immediately acted on. The wide difference between advocacy and incitement, between preparation and attempt, between assembling and conspiracy, must be borne in mind. In order to support a finding of clear and present danger, it must be shown either that immediate serious violence was to be expected or was advocated, or that the past conduct furnished reason to believe that such advocacy was then contemplated».

The emphasis is on the *imminence* of lawless action that leads to a distinction between mere advocacy of such action and direct incitement. Decided in 1969 and still good law, this is what *Brandenburg v. Ohio* stands for.³⁴ Thus in *Brandenburg*, the US Supreme Court adopted a very high threshold for speech bans, introducing an “imminent lawless action test”.³⁵ The *Brandenburg* test permits punishment (by the government or if we are to analogize – by taking down speech by a social media company) «for advocating violence only when speakers *intend* to cause *imminent* violence and

²⁹ *Ibid.*

³⁰ Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 2011 (UN Doc. A/HRC/17/27), emphasis mine.

³¹ *Un Resolution 16/18 adopted by the Human Rights Council Combating Intolerance, Negative Stereotyping and Stigmatization of, and Discrimination, Incitement to Violence and Violence Against, Persons Based on Religion or Belief* in *ohchr.org*, 3.

³² *Brandenburg v. Ohio*, 395 U.S. 444 (1969).

³³ D. Kaye, *Promotion and protection of the right to freedom of opinion and expression* (2019), U.N. Doc. A/74/486, emphasis mine.

³⁴ B. Petkova, *The evolution of Incitement Online: from Brandenburg v. Ohio to Depiction of Zwarte Piet*, in *medialaws.eu*, 23 April 2025.

³⁵ M.A. Franks, *Fearless Speech. Breaking Free from the First Amendment*, New York, 2024, 15.

only in situations in which such violence is *likely* to occur».³⁶ Brandenburg was a KKK leader. As US constitutional law scholar Mary Ann Franks writes, by acquitting him the US Supreme Court gave free rein to “reckless speech”. Although the Brandenburg test might be hailed by free speech absolutists, it must be reckoned that «reckless speakers like Brandenburg attack vulnerable groups, creating a risk of harm to others».³⁷ The indirect influence of *Brandenburg* on IHRL is further traceable to the report of a former UN Special Rapporteur on the Right to freedom of opinion and expression where incitement is defined as «*the imminent risk or likelihood* that the conduct desired or intended by the speaker will result from the speech in question».³⁸

3. International Human Rights Law’s developing new standard on hate speech

Given that there is no international human rights court that can shed light on the weighing of different factors, novel international standards on hate speech are being developed also through the decisions of the Meta Oversight Board. Why should we think that the Board justifies so much attention? First off, it has the de facto power to decide on the freedom of speech online of 3.07 billion monthly active users of Facebook, 2 billion active users of Instagram and over 350 million monthly active users of Threads worldwide. As Stefania di Stefano has shown, «strictly speaking, the Oversight Board is not empowered to apply IHRL directly. Rather, [its] Charter empowers it to interpret Meta’s content policies and values, with international human rights norms protecting free expression informing their interpretation».³⁹ However, the Board has chosen to fully embrace IHRL in its deliberations, «relying on a wide range on human rights standards. The Board, in fact, has not only applied human rights treaty provisions directly, but their interpretation is also informed by UN Treaty Bodies’ general comments and individual communications, reports from UN Special Procedures, UN General Assembly, Human Rights Council resolutions and other relevant UN documents».⁴⁰ With that, the Board has merited the qualification that it has been given by Laurence Helfer and Molly Land that compare it to an international human rights court.⁴¹ Further, as Helfer and Land explain, the Board’s legitimacy ultimately depends on its audience, with Meta users agreeing to be governed «by the company’s Community Standards and Values and by the Oversight

³⁶ *Ibid.*

³⁷ *Ibid.*, 16.

³⁸ D. Kaye, *Promotion and protection of the right to freedom of opinion and expression*, U.N. Doc. A/74/48050, emphasis mine.

³⁹ S. Di Stefano, *Translating and Developing International Human Rights Law in the Online Sphere*, forthcoming in I. Couzigou – E. Fromageau (eds.), *International Law and Technological Change: Testing the Adaptability of International Law*, Cheltenham-Northampton, 2025.

⁴⁰ *Ibid.*

⁴¹ *Supra.*

Board's review of those standards—including under international human rights law—when they create Facebook or Instagram accounts. They also agree that [Oversight Board] decisions are binding on the company. The large and growing number of appeals to the [Oversight Board] suggests that many users see value in challenging Meta's content moderation decisions». ⁴² For external parties in turn, the Board's legitimacy rests, among others, primarily on the persuasiveness of its decisions and on its independence from Meta. I now turn to the Board's persuasiveness in it arguably going beyond the interpretation of human rights law in developing a novel standard on hate speech. Independence is discussed later, in Part IV. My survey of the Board's hate speech and hateful conduct decisions finds that within the tripartite test, the most important factors motivating decisions for taking down speech are context and a novel element – that of *the risk of incitement*.

For example, when deciding to uphold Facebook's decision to remove hate speech based on the protected characteristic of national origin in the *Armenians in Azerbaijan* decision, ⁴³ the Board relied on the danger that the slur in question can escalate into physical violence. Facebook's Hate Speech Community Standard at the time prohibited some discriminatory expression, including slurs, «absent any requirement that the expression incite violent or discriminatory acts». ⁴⁴ The Board found that in accordance with IHRL whereas offline such an approach «...would raise concerns if imposed by a Government...», ⁴⁵ it is still acceptable by a social media company. The Board ultimately based its application of the necessity prong of the tripartite test on the *potential* for adverse outcomes. Although for the Board the speech did not rise to the level of incitement, in its examination of Facebook values it considered that: «Left up, an accumulation of such content may create an environment in which *acts of discrimination and violence are more likely*». ⁴⁶ Therefore, it seems that the *ratio decidendi* of the decision was rooted in the assessment of the *risk of incitement*. That conclusion is buttressed by the assessment of context by the Board: an ongoing armed conflict between Armenia and Azerbaijan. The Board made its risk assessment of possible incitement based on that context. Specifically, it held that: «The content in question was posted to Facebook shortly before a ceasefire went into effect. This context was especially relevant for the Board. While pointed language may be a part of human interactions, particularly in conflict situations, the danger of dehumanizing slurs proliferating in a way that escalates into acts of violence is one that Facebook should take seriously». ⁴⁷

Similarly, in the *Depiction of Zwarte Piet's* decision ⁴⁸, the Board upheld Facebook's decision to take down speech which violated the protected charac-

⁴² *Ibid*, 2296-2297.

⁴³ *Armenians in Azerbaijan*, in *oversightboard.com*, 28 January 2021.

⁴⁴ *Ibid.*, para. 8.3c.

⁴⁵ *Ibid.*

⁴⁶ *Ibid*, para. 8.2., emphasis mine.

⁴⁷ *Ibid.*

⁴⁸ *Depiction of Zwarte Piet*, in *oversightboard.com*, 13 April 2021.

teristic of race and ethnicity.

When discussing the necessity and proportionality of the ban on blackface, the Board was concerned with «the accumulation of degrading caricatures of Black people on Facebook creat[ing] an environment where *acts of violence are more likely* to be tolerated and reproduce discrimination in a society».⁴⁹ In its necessity and proportionality analysis moreover, the Board quoted the documented experience of discrimination and violence of Black people in the Netherlands that were connected to the practice of Zwarte Piet. It «noted reported episodes of intimidation and violence against people peacefully protesting Zwarte Piet».⁵⁰ All this to say that although for the Board the depiction of Zwarte Piet did not rise to the level of incitement, it did contain *risk of incitement*. In *Depiction of Zwarte Piet* moreover, the Board relied on a broader notion of harm compared to the one espoused in the Rabat Plan of Action and unlike in the Rabat Plan, didn't require intent to obtain for the content to be removed. For example, it stated that: «[a] majority found that allowing such posts to accumulate on Facebook would help create a discriminatory environment for Black people that would be degrading and harassing». With that, the Board pointed out to the long-term societal consequences of the speech, not its real or potential possibility for imminent harm. This interpretation differs from the definition of direct incitement that leads to imminent harm, as described in paragraph 21 of the Rabat Plan of Action.

Among others, in both *Armenians in Azerbaijan* and *Depiction of Zwarte Piet*, the Board relies on a report by the former UN Special Rapporteur on freedom of expression, Professor David Kaye.⁵¹ In the report, David Kaye states that: «The question arises as to whether States may restrict advocacy of hatred that does not constitute incitement to discrimination, hostility or violence».⁵² For Professor Kaye, such speech can be restricted if it meets the article 19.3 tripartite test, with blasphemy bans and laws that ban the denial of the Holocaust falling short of the test.⁵³ The Board has interpreted the report to mean that there are different obligations for social media companies and states. By way of an “obiter dicta”, in *Zwarte Piet* the Board reasoned that: «international human rights law would not allow a state to impose a general prohibition on blackface through criminal or civil sanctions, except under the conditions foreseen in ICCPR Article 20, para. 2 and Article 19, para. 3 (e.g., advocacy of hatred constituting incitement to violence). Expression that does not reach this threshold may still raise concern in terms of tolerance, civility and respect for others, but would not be necessary or proportionate *for a state* to restrict...In the Board's view, the individual post in this case would fall within this category of protection from state restriction».⁵⁴

Perhaps it is normal for the Board to want to protect a newly formed

⁴⁹ *Ibid.*, para. 8.3.III., emphasis mine.

⁵⁰ *Ibid.*

⁵¹ D. Kaye, *Promotion and protection of the right to freedom of opinion and expression*, cit.

⁵² *Ibid.*, para. 19.

⁵³ *Ibid.*, paras. 21-22.

⁵⁴ *Supra* n 48, para. 8.3.III., emphasis mine.

standard on freedom of expression by narrowing down said standard (of a risk to incitement) to the context of social media. However, an external perspective may well be that to preserve the consistency and uniformity of IHRL there should not be diverging interpretations on hate speech in the online and offline environment. There seems to be good reasons why in the context of hate speech, the US inspired test of “real and imminent danger” might be reconsidered and relaxed to a risk of incitement test that takes context and empirical facts as substantiations for the risk assessment. The Board further reaffirmed its risk of incitement standard in the *Alleged Crimes in Raya Kobo* case.⁵⁵ There the Board found that «Rumours alleging that an ethnic group is complicit in mass atrocities, as found in this post, are dangerous and *significantly increase the risk* of imminent violence».⁵⁶ Moreover, the Board relied on and cited to its previous decision in *Armenians in Azerbaijan*: «in situations of armed conflict in particular, the risk of hateful, dehumanizing expressions accumulating and spreading on a platform, leading to offline action impacting the right to security of person and potentially life, is especially pronounced».⁵⁷ Thus, the Board assessed the risk of incitement as follows: «Unverifiable rumours in a heated and ongoing conflict could lead to grave atrocities» and «...unsubstantiated claims regarding civilian perpetrators are likely to *heighten risks* of near-term violence».⁵⁸ In its necessity and proportionality assessment the Board again admitted that «A direct call for violence is absent from the post in this case, although there is a reference to ‘our struggle’».⁵⁹ Although the Board deemed it appropriate to assess the post from the perspective of its Violence and Incitement Community Standard rather than its Hate Speech standard⁶⁰, it still suggested that the offline standard of incitement to violence might not be fulfilled in this case. However, there existed *risk to incitement* that went beyond the imminence of the danger. Finally, among others, the Board recommended that all of Meta’s Community Standards should reflect the higher risk to the rights of life and security of persons that unverified rumours pose during war and violent conflict. That controversial speech is uttered during an ongoing war and violent conflict however would not on its own suffice to meet the new standard of risk of incitement with which the Board is grappling. Moreover, risk of incitement is balanced against the context of the speech and evaluated in the light of the Rabat Plan. The Rabat Plan factors are thus folded into the necessity and proportionality prong of the tripartite test in the Board’s

⁵⁵ *Alleged Crimes in Raya Kobo*, in *oversightboard.com*, 14 December 2021.

⁵⁶ *Ibid.*, case summary, emphasis mine.

⁵⁷ *Ibid.*

⁵⁸ *Ibid.*

⁵⁹ *Ibid.*, para. 8.3.III.

⁶⁰ The importance which the Board attaches to assessing incitement when restricting speech has been discussed by Wong and Floridi who predicted that the Board might assess removal of posts more often under its Violence and Incitement Community Standard rather than its Hate Speech Standard, see D. Wong – L. Floridi, *Meta’s Oversight Board*, cit., 274 ff.

assessment. In its *Russian poem* decision⁶¹, the Board found that the existing risk of incitement fell below the threshold for taking speech down. The risk assessment of the Board went as follows: «To the extent that the post, with its quotations from Simonov’s poem, could be considered to refer to soldiers now committing atrocities against civilians, there is a risk that readers will read this as a call to violence against Russian soldiers today. The Board nonetheless agrees with Meta’s conclusion that the post does not violate the Violence and Incitement Standard because its primary meaning, in context, is a warning against a cycle of violence».⁶²

To be sure, the novel standard of risk of incitement is not used by the Board only when incitement to violence is implicated. Another basis on which the Meta Oversight Board has decided to take down hate speech has been risk to incitement of discrimination. Short of satisfying the offline standard of incitement to discrimination, the Board has found it necessary and proportionate to restrict racial slurs in the context of post-apartheid South Africa. It held that: «In this case, the historical and social context was crucial, as the Board notes the use of the word...is closely linked with discrimination and the history of apartheid in South Africa».⁶³

By analogy to the *South Africa slurs* decision, the Board found risk of incitement to discrimination also in its *Knin Cartoon* decision.⁶⁴ The Board stated: «Regarding the context, this relates to a region that has recently experienced ethnic conflict and the backdrop of online hate speech and incidents of discrimination against ethnic minorities in Croatia... [The post] intends to incite ethnic hatred, and this may contribute to individuals taking discriminatory action».⁶⁵ In *Knin Cartoon*, since the post fell short of imminent danger and therefore incitement, the Board was again adamant to limit the remit of its decision to the online world. It stated that: «... Meta can legitimately remove posts from Facebook that encourage violence in a less immediate way. This is justified, as 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».⁶⁶

Meta’s reversal of policy regarding the Holocaust can be read against the same backdrop of online-offline tension.⁶⁷ In its *Holocaust denial* decision⁶⁸, the Board reversed Meta’s original decision to leave up an Instagram post containing false and distorted claims about the Holocaust. In its assessment of the legitimate aim prong of the tripartite test, the Board mentioned the many public comments Meta received when it changed its policy to introduce the prohibition on Holocaust denial. Comments such as

⁶¹ *Russian poem*, in *oversightboard.com*, 16 November 2022.

⁶² *Ibid.*, para. 8.1. II.

⁶³ *South Africa slurs*, in *oversightboard.com*, 28 September 2021, para. 8.3. II.

⁶⁴ *Knin Cartoon*, *oversightboard.com*, 17 June 2022.

⁶⁵ *Ibid.*, para. 8.3.III.

⁶⁶ *Ibid.*

⁶⁷ M. Bickert, *Removing Holocaust Denial Content*, in *about.fb.com*, 12 October 2020.

⁶⁸ *Holocaust denial (2024)*, in *oversightboard.com*, 23 January 2024.

those by the American Jewish Committee and its Jacob Blaustein Institute for the Advancement of Human Rights pointed out that Meta's reversal of policy was necessary to safeguard against increased incitement to violence and hostility.⁶⁹ At least some members of the Board's majority, moreover, did not interpret its obligations to differ to those of governments in the offline world by stating that: «[t]he ban on Holocaust denial is therefore in conformity with the ICCPR and the obligations expressed in Article 4 (a) of the International Convention on the Elimination of All Forms of Racial Discrimination».⁷⁰ For other members of the majority that assessed necessity and proportionality in the light of previous Board's decisions like *South Africa slurs* and *Zwarte Piet*, however, a variation of online and offline standards was justified given that online prohibitions of speech would be periodically reviewed.⁷¹ The opinion was split further, with the minority highlighting incompliance with IHRL that declaring a speech ban on a Holocaust denial posed. Specifically, for the minority of Board members «reliance on Article 4 of the ICERD is misplaced» as the «Committee on the Elimination of Racial Discrimination (CERD, which is charged with monitoring implementation of the ICERD) specifically addressed the topic of genocide denial, stating it should only be banned when the statements “clearly constitute incitement to racial violence or hatred».⁷² For the minority, therefore, since the high threshold of imminence and incitement was not reached, the Board could not impose a ban on Holocaust denial. Interestingly, the minority insisted that should there be a departure from IHRL, the Board needs to be explicit about it, triggering public dialogue.

4. The Pluralist Landscape of Offline and Online Standard Setting in IHRL

IHRL has not provided a conclusive answer on the possible bifurcation of proscribed speech under Article 20 ICCPR into two broad classifications: (1) hate speech that does not rise to the level of incitement; and (2) speech that incites violence.⁷³ Both can be understood to constitute sub-catego-

⁶⁹ *Ibid.*, para. 8.2. II.

⁷⁰ *Ibid.*, para. 8.2. III.

⁷¹ *Ibid.*

⁷² *Ibid.*

⁷³ In the *Ruggiu* case, a Trial Chamber of the International Criminal Tribunal for Rwanda (ICTR) found that a crime against humanity – persecution – can be committed via speech «without the speaker explicitly calling for violence». A later ICTR Trial Chamber similarly found that «...there need not be a call to action in communications that constitute persecution». However, in another ad hoc tribunal judgement – that of a Trial Chamber of the International Criminal Tribunal for the former Yugoslavia (ICTY) – the Chamber found that in the lack of direct incitement to violence the defendant's «political speeches were never racially inflammatory nor were they intended to foment hatred of Bosnian Muslims by Bosnian Croats». Thus, the speech in question did not constitute persecution. For a detailed analysis, see G. Gordon, *Hate Speech and Persecution*, in *Vanderbilt Journal of Transnational Law*, 46, 2013, 2 ff.

ries of hate speech, but should they merit the same or different treatment? Should, say, the latter be criminalized and if we are searching for analogies, should it be taken down by social media companies with the possible suspension of accounts in more severe cases, while the former penalized by civil law or demoted? Of relevance to this debate a few years back, David Kaye poignantly asked whether «[g]iven the speed, reach, and capacity for disinformation online, ...our definition of “online hate speech” [should] be different from offline hate speech». ⁷⁴ Should posting on social media that a fire is taking place in a theatre when it, in fact, it is not, be treated differently than standing up and yelling, “fire” in that same crowded theatre? ⁷⁵ In the context of misinformation, Robert Post has held that: «[Online] speech that is simultaneously distributed to millions of persons may produce harm in ways that cannot meaningfully be conceptualized through the discrete pathways of simple causality. We must instead think in terms of the statistical probabilities of harm. At present, however, we lack any legal framework capable of assessing stochastic harms in ways that will not drastically overregulate speech. If we were to suppress every form of communication that might cause harm if distributed in sufficiently large numbers, we would have precious little communication left unregulated». ⁷⁶

The Meta Oversight Board promotes the view that certain online speech, even in the absence of direct incitement can be taken down for risk of incitement. The risk of incitement analysis is well-substantiated by the Board and based, among others, on the notion of accumulative harm. Such approach seems to fit the tripartite test and is not restrictive of speech beyond what is necessary. Introducing a more stringent standard for online hate speech can however potentially go against the grain of IHRL. With the rise in digital crimes, it became clear that the digital space, characterized by an un-territoriality of data as it might be, ⁷⁷ cannot remain disentangled from the physical world and left unregulated. The United Nations, followed by regional organizations, adopted what has been dubbed as the ‘normative equivalency paradigm’, ⁷⁸ basically proclaiming, albeit in a series of non-binding resolutions, that the same human rights that people have offline must be protected also online. In this understanding, the internet is one important medium among several in which human rights can be exercised. Equally though soft law, the European Parliament asserted that it «[c]ondemns all kinds of hate crime, hate speech and accusations devoid of foundation or formulated in bad faith, *both offline and online*, motivated by discrimination based on any grounds such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any

⁷⁴ D. Kaye, *Four Questions about Online Hate Speech*, in *medium.com*, 12 August 2019.

⁷⁵ C. Logfren, *Regulating Free Speech: Online vs. Offline*, in *linkedin.com*, 20 February 2020.

⁷⁶ R. Post, *The Internet, Democracy and Misinformation* in R. Krotoszynski – A. Koltay – C. Garden (eds), *Disinformation, Misinformation, and Democracy*, Cambridge, 2024.

⁷⁷ J. Dascal, *The Un-territoriality of Data*, in *Yale Law Journal*, 125, 2015, 2 ff.

⁷⁸ D. Dror-Shpoliansky – Y. Shany *It’s the End of the (Offline) World as We Know It: From Human Rights to Digital Human Rights – A Proposed Typology* in *European Journal of International Law*, 32, 2021, 1249 ff.

other opinion, membership of a national minority, property, birth, disability, age or sexual orientation».⁷⁹ Moreover, states do not have carve outs for online speech and their ordinary criminal laws apply. This advocates in favour for the adoption of one unified standard on hate speech for both the online and the offline world.

This is without prejudice however to the fact that the internet environment enables more flexibility, for example, when it comes to the application of the necessity and proportionality prong of the tripartite test. Clearly, proportionality online can entail the demotion of harmful speech instead of taking it down.⁸⁰ Moreover, proportionality for social media companies can entail affixing warning labels online, a measure not available in the offline world. Thus, the human rights obligations of states and social media companies – although only voluntary in the latter case – will vary in substance regarding the severity of the penalty. It can be contested if the magnitude and scope of power that states exert according to criminal law procedures matches the magnitude and power that a social media company may exert over its users, with deplatforming arguably being the harshest penalty. However, the more capaciously social media plays a role in our life – for example, when users rely on platforms for their basic income or for satisfying a vital community interest – the more the distinction between states and social media (and their concomitant human rights law obligations) starts to get blurred.

Granted, IHRL norms are dynamic; they are not set in stone. Moreover, traditional international law⁸¹ has an overwhelmingly state-centred focus but the realities of non-state actors' prominence on the international scene effectively challenge that view. New interpretations of IHRL are being developed offline: by UN bodies including but also going beyond the International Law Commission, regional organizations, public courts and academics. However, also newly formed private institutions that have a final say on online speech on a platform such as the Meta Oversight Board can contribute to an overall understanding of updated IHRL standards.

⁷⁹ European Parliament, *Resolution of 15 September 2022 on the Situation of Fundamental Rights in the European Union in 2020 and 2021 (2021/2186(INI))*, in [europarl.europa.eu](https://eur-parl.europa.eu), 15 September 2022, emphasis added.

⁸⁰ J. Howard - B. Kira, *Remove or Reduce: Demotion, Content Moderation, and Human Rights* in *ssrn.com*, 17 July 2024. See also S. Benesch, *But Facebook is not a Country: How to Interpret Human Rights Law for Social Media Companies*, in *Yale Journal on Regulation Bulletin*, 2020, 38 ff. Although it goes out of the scope of this work, further research on the similarities and differences of states and social media companies can shed light on the possible need (or the lack thereof) for diverging IHRL interpretations online and offline. It can be argued that the similarities between states and social media companies are compelling. To avoid tyranny, both states and social media companies need to be held to a procedural understanding of the rule of law, understood as the condition of exerting power only according to rules that are public, fairly applied, and meaningfully contestable. S. Donahue, *Platform Tyranny, Rule of Law, and Virtual Community*, presentation given at conference on the Philosophy of Content Moderation held in California 12-15 April 2025. This claim is sustained also by digital constitutionalists who observe the transfer of constitutional principles to the digital sphere as both an empirical fact and one that is normatively desirable. See, generally, G. De Gregorio, *The Rise of Digital Constitutionalism in the European Union*, in *International Journal of Constitutional Law*, 19(1), 2021.

⁸¹ A. Bianchi, *International Law Theories: An Inquiry into Different Ways of Thinking*, Oxford, 2016, 21 ff.

What is more, arguably the new standard on risk of incitement developed by the Oversight Board is also supported by academic interpretations that have sustained the view that a relaxed interpretation of incitement is necessary to protect the rights of others.⁸²

In this pluralistic landscape, regional institutions and law have a say too. Thus, X's content moderation practices are said to be primarily rooted in two seemingly contradictory regional legal instruments – that of the United States Bill of Rights and the European Convention on Human Rights.⁸³ Likewise, Meta's newsworthiness and public figure standard has been inspired by the First Amendment,⁸⁴ whereas its hate speech community standard can be traced back to European norms.⁸⁵ Perhaps by bridging the differences between regional norms, the content moderation practices of major social media companies create novel standards that can inform an integrated understanding of IHRL.

Similarly, by diverging from authoritative interpretations of IHRL, regional courts like the European Court of Human Rights can be contributing to the bottom-up development of a new IHRL standard. In this regard, the ECtHR's holdings that non-violent advocacy of National Socialism should be declared incompatible with the Convention⁸⁶ can be read in the light of the risk of incitement standard that prompted the Meta Oversight Board to prohibit denial of the Holocaust. Not unlike the Meta Oversight Board but without a detailed substantiation of risk of incitement, the Court has often found that even in the absence of direct incitement, speech can be proscribed. In a case about distributing leaflets with homophobic statements about gay people - *Vejdeland and Others. v Sweden* – the ECtHR «highlighted that incitement to hatred need not necessarily advocate for violence or call for criminal acts».⁸⁷ As one of the concurring opinions in the case stated: «Statistics on hate crimes show that hate propaganda always inflicts harm, be it immediate or potential. It is not necessary to wait until hate speech becomes a real and imminent danger for

⁸² S. Soriall, *Hate Speech and Distorted Communication: Rethinking the Limits of Incitement*, in 34 *Law and Philosophy*, 34, 2014, 3 ff; J. Howard, *Dangerous Speech in Philosophy & Public Affairs*, 47(2), 2019, 208 ff.

⁸³ The scope of protection of the First Amendment and that of Article 10 of the European Convention on Human Rights significantly vary specifically regarding hate speech. See *infra* in Introduction and E.M. Aswad, *The Future of Freedom of Expression Online*, cit.

⁸⁴ T.E. Kadri – K. Klonick, *Facebook v. Sullivan: Public Figures and Newsworthiness in Online Speech*, in *Southern California Law Review*, 93, 2019, 37 ff.

⁸⁵ «In the wake of recent terror attacks and the resurgence of hate groups, European regulators have pressured tech companies to change their speech rules and practices to remove extremist material appearing on their platforms. Some of the resulting changes may do some good». D. Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, in *Notre Dame Law Review*, 93, 2018, 3 ff. See also European Commission Press Release IP/16/1937, *European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech* (2016) in *europa.eu*, 31 May 2016.

⁸⁶ J. Mchangama – N. Alkiviadou, *Hate Speech and the European Court of Human Rights: Whatever Happened to the Right to Offend, Shock or Disturb?*, in *Human Rights Law Review*, 21, 2021, 1008 ff., 1034.

⁸⁷ *Ibid.*, 1032.

democratic society».⁸⁸ Moreover, in an earlier case about anti-immigration rhetoric – *Féret v Belgium* – the ECtHR similarly found that there is no need for the proscribed speech to present «the calling of a specific act of violence or another criminal act».⁸⁹ However, instead of extensively outlining the potential risks connected to the incitement to discrimination that the far-right party's speech could have in this case,⁹⁰ for the most part in *Féret* the Court focused on the offensive elements of the speech, thereby triggering criticism by commentators who saw the ban as incompatible with political expression protected under art. 10 of the European Convention on Human Rights (ECHR).⁹¹ In *Soulas and Others v France*, the Court was again criticized for deficient reasoning when it proclaimed as proscribed hate speech the phrase «it is only if an ethnic civil war breaks out that the solution can be found» for potentially inciting aggression against a particular group.⁹² In *Soulas*, the potential for incitement could have been substantiated better by a more detailed risk assessment. In another case – *Baldassi and others v. France*, the ECtHR itself overruled a domestic court for failing to present enough support on why a ban on a boycott of Israeli products could risk «a call for violence, hatred or intolerance».⁹³

This is by no means to suggest that the ECtHR has presented a coherent case law on hate speech. In *Perinçek v Switzerland*, «the applicant's prosecution for publicly describing the Armenian genocide as 'an international lie' was found to be in violation of Article 10».⁹⁴ The Court therefore did not find the case to present incitement to hatred or intolerance, neither it discussed possible risks that the denial of genocide, different than the Holocaust, could pose. The denial of the Holocaust has therefore been isolated to a special category, leaving hate speech with potentially high risk of incitement to discrimination, hostility or violence that denies genocide in other circumstances to remain protected under art. 10 of the European Convention on Human Rights.

To summarize, it is submitted that IHRL on hate speech is developed in a legal pluralist framework that allows points of entry to multiple actors – UN institutions, regional courts and global private self-regulatory mechanisms. The Meta Oversight Board is one such mechanism but the

⁸⁸ ECtHR, *Vejdeland and Others v. Sweden*, app. no. 1813/07 (2012), Concurring Opinion of Judge Yudkivska joined by Judge Villiger, § 11.

⁸⁹ J. Mchangama – N. Alkiviadou, *Hate Speech and the European Court of Human Rights*, cit., 1030.

⁹⁰ Possibly with the exception of § 77: «[Political parties must avoid advocating solutions to immigration] by inciting racial discrimination and resorting to vexatious or humiliating comments or attitudes, as such behaviour risks arousing reactions among the public that are incompatible with a peaceful social climate and undermining confidence in democratic institutions». Own translation. *Féret v Belgium*, cit.

⁹¹ J. Mchangama – N. Alkiviadou, *Hate Speech and the European Court of Human Rights*, cit., 1031.

⁹² *Ibid.*

⁹³ *Baldassi and others v. France*, app. 15271/16 (2020), §. 79.

⁹⁴ J. Mchangama – N. Alkiviadou, *Hate Speech and the European Court of Human Rights*, cit., 1022.

advent of the European Digital Services Act⁹⁵ might bring others. The DSA mandates that online platforms «provide users access to an internal complaint-handling system of the platform whereby users can file complaints about content removal and visibility suspension/termination of user accounts; suspension/ termination of provision of service to the users; and suspension/termination of users' monetization of content».⁹⁶ Further, Article 21 of the DSA empowers users to select newly-formed “out-of-court dispute settlement bodies” that can oversee complaints that could not be resolved through the internal mechanisms.⁹⁷ The first such non-binding bodies have seen the light of day;⁹⁸ it remains to be seen whether they will also construct their own standards on hate speech anchored in IHRL.

5. Hate Speech as Censorship Creep or Firm Enforcement of Human Rights?

On one side of the spectrum stand those that would not like to see any speech banned, including what might fall under the net of hate speech. In this view, hate speech codes of any kind «no matter how deliberate and carefully crafted, are inherently vague and overbroad, inevitably harming expression and equality».⁹⁹ Another view, inspired by American constitutional law that has made forays in IHRL, is that hate speech should be banned only when it reaches the threshold of real and imminent danger of incitement to discrimination, hostility or violence. Such a high threshold would possibly offer protection against terrorist speech but would leave outside of the remit of proscribed speech otherwise harmful expression. Arguably, a third solution is to discern a developing novel standard – that of proscribing speech that reaches a certain risk of incitement, not imminence. Although not coinciding on their respective justifications but on outcomes, the Meta Oversight Board and the Strasbourg Court can be seen to pedal such a middle way.

As a matter of theory, Andrew Kenyon writes: «[i]f some form of public speech counters or undermines the possibility of self-government, it could warrant limitation. That might well apply to hate speech. Limits on public discourse would be based on the adjudged effect of the speech on the possibility of democratic self-government. Limits would remain

⁹⁵ Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM(2020) 825 final (Feb. 15, 2020).

⁹⁶ D. Wong – L. Floridi, *Meta's Oversight Board*, cit., 265.

⁹⁷ *Ibid.*

⁹⁸ T. Bernard, *Oversight Board Trust Launches EU Out-of-Court Dispute Settlement Service*, in *techpolicy.press*, 10 October 2024.

⁹⁹ D. Inserra, *The Return of Blasphemy and Sacrilege Laws in the Most Unlikely Places*, in *cato.org*, 23 January 2024. The author refers to the influential work of N. Strossen, *Hate: Why We Should Resist it with Free Speech, Not Censorship*, Oxford, 2018.

an issue of legal and academic debate, but limits would be possible».¹⁰⁰ Self-government would be arguably severely damaged if the free speech ideal is reduced «to a simplistic and reductionist promotion of reckless speech» that Professor Franks defines as speech in the service of «powerful antidemocratic interests: misogyny, racism, religious fundamentalism, and corporate self-interest».¹⁰¹ Further, as Katharine Gelber reveals, hate speech «(in both vituperative and more modest forms) is capable of harming political participation»,¹⁰² that is in turn vital to self-government. In IHRL overtones, Gelber's argument can be particularly well suited to the risk of incitement to discrimination that hate speech might pose. She primarily conceives of hate speech as speech «that is directed at historically identifiable minorities, that targets them with speech that is harmful to their involvement in processes of democratic legitimation».¹⁰³ As both Franks and Gilber demonstrate, such speech can silence vulnerable individuals and groups of individuals.

It is a common trope that there is no universally accepted legal definition on hate speech. However, the UN offers a prompt: «The term hate speech is understood as any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor».¹⁰⁴ Meta's community standard has elaborated on other identity factors by defining hateful conduct as follows: «direct attacks against people – rather than concepts or institutions – on the basis of what [Meta] call[s] protected characteristics (PCs): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. Additionally, [Meta] consider[s] age a protected characteristic when referenced along with another protected characteristic. [Meta] also protect[s] refugees, migrants, immigrants and asylum seekers from the most severe attacks..., though [it] do[es] allow commentary on and criticism of immigration policies. Similarly, [Meta] provide[s] some protections for non-protected characteristics, such as occupation, when they are referenced along with a protected characteristic. Sometimes, based on local nuance, [Meta] consider[s] certain words or phrases as frequently used proxies for protected characteristics».¹⁰⁵ Although Meta's definition might appear to offer a broad scope for hate speech (e.g. a wide range of PCs), arguably a requirement for a well substantiated risk of incitement can warrant the narrow application of the definition.¹⁰⁶

¹⁰⁰ A.T. Kenyon, *Democracy of Expression. Positive Free Speech and Law*, Cambridge, 2021, 64 ff.

¹⁰¹ M.A. Franks, *Fearless Speech*, cit., 16.

¹⁰² K. Gelber, *Hate Speech – Definitions and Empirical Evidence*, in *Constitutional Commentary*, 2017, 559. In the same symposium, see also A. Stone, *Viewpoint Discrimination, Hate Speech Laws, and the Double-Sided Nature of Freedom of Speech*, in *Constitutional Commentary*, 40, 2017.

¹⁰³ K. Gelber, *ibid.*, 625.

¹⁰⁴ *United Nations Strategy and Plan of Action on Hate Speech*, in [un.org](https://www.un.org/en/development/desa/policy/hate-speech/).

¹⁰⁵ *Hateful Conduct, Meta Community Standards*, in transparency.meta.com.

¹⁰⁶ Out of the so far thirty-one Board's decisions on hate speech and hateful conduct, the

Amidst purported concerns of censorship, Meta has recently revised its policy on hate speech. In its previous iterations, the policy prohibited «discriminatory expression including slurs, absent any requirement that the expression incite violence or discriminatory acts».¹⁰⁷ It therefore remains to be seen if the Meta Oversight Board will continue to enforce consistently its nascent standard on risk of incitement, as discussed above. Since Zuckerberg's change of heart has been accused of partisanship,¹⁰⁸ it is even more imperative that the Board perseveres to preserve its neutrality. Instead of censorship, the consistent application of IHRL would ensure that human rights are not providing a mere veneer of legitimacy for the Board.¹⁰⁹ Moreover, the convergence on outcomes by the Meta Oversight Board and the ECtHR dictates that the latter might benefit from considering the Board's methodology in its reasoning. This is even more so since the Strasbourg Court's approach on hate speech regarding varying standards on incitement has been criticized as adhocery.¹¹⁰

Hateful expression, both online and offline, presents long-term challenges even if it is not clearly linked to adverse outcomes. Restricting such expression however should be properly justified, subject to the UN tripartite test and helped by a risk assessment of incitement in the spirit of the Rabat Plan. Professor Richard Moon echoes a common concern: that social media platforms, «in responding to popular or legal demands to remove speech that is hateful or harassing, will be overinclusive in the enforcement of any ban, resulting in the exclusion of speech that is lawful».¹¹¹ However, he also admits that «[d]eceptive and harassing speech that previously was not viewed as sufficiently harmful to justify its restriction, is now, in the online world, seen as more dangerous to individuals and damaging to public discourse». The question remains of course whether such speech can be equally damaging to public discourse in the offline world and whether it is simply identified as more damaging online because it is first detected there. Our daily lives have irrevocably changed, causing a lot of our usual routine to occur in the digital space. Therefore, international court-like institutions such as the Meta Oversight Board become harbingers of what might be expected to come in the physical space.

Although the Board's pronouncements are only on a handful of highly emblematic cases, its potential impact is much bigger given the precedential value of its decisions. However, Meta determines when «past decisions can serve as precedent for future content moderation decisions that resemble the past decision in terms of the facts or issues of the case».¹¹² The

Board found that speech should be taken down in 16.

¹⁰⁷ S. Di Stefano, *Translating and Developing International Human Rights Law in the Online Sphere*, cit.

¹⁰⁸ J. Bayer, *Zuckerberg's Strategy: Leveraging Trump to Defy European Regulation?*, in *verfassungsblog.de*, 21 January 2025.

¹⁰⁹ E. Douek, *The Meta Oversight Board and the Empty Promise of Legitimacy*, in *Harvard Journal of Law & Technology*, 37, 2024, 373 ff.

¹¹⁰ J. Mchangama – N. Alkiviadou, *Hate Speech and the European Court of Human Rights*, cit., 1032.

¹¹¹ R. Moon, *The Life and Death of Freedom of Expression*, University of Toronto, 2024, 309.

¹¹² D. Wong – L. Floridi, *Meta's Oversight Board*, cit., 264.

Board has so far been building a doctrine on hate speech based on a consistent referral to its past decisions.¹¹³ A potential reversal of said doctrine by the Board or refusal to follow precedent by Meta¹¹⁴ because of the new climate at the company can only mean chaos and legal uncertainty, and the underenforcement of international human rights.

6. Concluding Remarks

«According to the available statistics, in the European Union (EU) alone, the twenty-two largest platforms and search engines make hundreds of millions of content moderation decisions per month».¹¹⁵ If consistently applied, IHRL has the potential of straddling the chasm that can open between the probabilistic, algorithmic online and the individualistic, rights-based offline speech protection. Moreover, IHRL can serve as a bridge between different constitutional law understandings such as the US and European traditions on hate speech.

Human rights standards and interpretations evolve. This has been recognized in different constitutional law systems. For instance, theories on the ‘living constitution’ method of interpretation find supporters in the United States¹¹⁶ and Europe¹¹⁷ alike. It is not unfathomable to imagine that in order to ensure the evolution of human rights, international private bodies that are publicly minded should also be allowed a voice in shaping IHRL. Meta Oversight Board’s nascent standard on hate speech includes application of the widely recognized UN tripartite test on legality, legitimate aim and necessity. It also considers the UN Rabat Plan of Action on hate speech, often in the necessity prong of the test. The Board has however started shaping a relaxed interpretation of incitement to discrimination, hostility or violence that is not inconsistent with an evolving understanding of Article 20 of the ICCPR. In this interpretation, what plays decisive role for speech to be taken down is the well-substantiated risk to incitement rather than the imminence of danger. However, that the Board provides more elaborate reasons for its choices,¹¹⁸ especially when advocating for divergent protection of offline and online speech remains an ongoing demand.

¹¹³ See *infra*. Part II.

¹¹⁴ Meta has committed to applying past OB decisions to «identical content in parallel context» where doing so is «technically and operationally feasible...[u]sing these criteria, Meta—who interprets them—could effectively nullify the use of any precedent by citing trivial differences between two cases or by overestimating the difficulty of applying precedent to a current case». D. Wong – L. Floridi, *Meta’s Oversight Board*, cit., 270.

¹¹⁵ *Supra*.

¹¹⁶ S. Breyer, *Reading the Constitution. Why I choose Pragmatism, not Textualism*, New York, 2024.

¹¹⁷ G. Letsas, *The ECHR as a Living Instrument: Its Meaning and its Legitimacy*, in A. Follesdal – B. Peters – G. Ulfstein (eds.), *Constituting Europe. The European Court of Human Rights in a National, European and Global Context*, Cambridge, 2013.

¹¹⁸ E. Douek, *The Meta Oversight Board and the Empty Promise of Legitimacy*, cit., 414-425. See also D. Kaye, *Promotion and protection of the right to freedom of opinion and expression*, cit.

Abstract

In a legal pluralist landscape, newly formed private bodies like the Meta Oversight Board contribute to the establishment of a novel international human rights law (IHRL) standard on hate speech. If expression is to be restricted on the ground of being hateful, the question becomes how to justify said restrictions. I argue that IHRL on freedom of expression, influenced by US constitutional law, has developed a very high threshold for proscribing harmful speech – that threshold is heavily connected to incitement to violence, discrimination or hostility being understood as imminent danger. The nascent standard that the Meta Oversight Board is advancing offers a broader understanding of harm resulting in content being taken down even in the absence of direct incitement and intent. The Board is thus often concerned with the long-term social consequences of hate speech rather than with the imminence of harm.

Keywords

International Human Rights Law – freedom of expression – free speech
– hate speech – Meta Oversight Board