DOI: HTTP/DOI.ORG/10.370.75/BAL.2024.01.07

Правосъдие в дигиталния свят. стандартите за защита на правата на човека в международното право и решенията на Надзорния съвет на „Мета“

Justice in the Digital World. International Human Rights Protection Standards and The Decisions of Meta's Oversight Board

Антица Генева, Биляна Петкова[1]

SUMMARY

The Oversight Board, an independent body set up by the international company Meta, plays a key role in content moderation by interpreting and applying international human rights norms in the digital environment that the company manages. This article examines the origins, structure and competence of the Oversight Board, analyses the compliance of its activities with

[1] Ас. д-р Антица Генева, ЮФ, УНСС, e-mail: antitsa.geneva@unwe.bg (Asst. Prof. Antitsa Geneva, Law Faculty, UNWE, e-mail: antitsa.geneva@unwe.bg).

Д-р Биляна Петкова, Установен изследовател, УНСС, e-mail: bpetkova@unwe.bg (JD, Bilyana Petkova, Principal Investigator UNWE, e-mail: bpetkova@unwe.bg). The research receives funding from National Scientific Program VIHREN-2024, BNSF of MES under project No. КП 06-ДВ/6 of 17.12.2024 of the UNWE, *The Online/Offline Divide: Can Pre-Digital Law Govern the Online World? Focus on freedom of expression at the Meta Oversight Board*, project manager Dr. Bilyana Petkova. The research receives funding from National Scientific Program VIHREN-2024, BNSF of MES under project No. КП 06-ДВ/6 of 17.12.2024 of the UNWE, *The Online/Offline Divide: Can Pre-Digital Law Govern the Online World? Focus on freedom of expression at the Meta Oversight Board*, project manager Dr. Bilyana Petkova.

the principles of international human rights law – in particular Articles 19 and 20 of the International Covenant on Civil and Political Rights (ICCPR) – and assesses how its decisions reflect and interpret key instruments in international law such as the UN Rabat Plan of Action on the Prohibition of Advocacy for National, Racial or Religious Hatred (Rabat Plan of Action) and the Report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. By analysing the practice of the Oversight Board, the article focuses on the interpretation and application of international human rights norms by private self-regulatory bodies. We discuss the definition of the limits of freedom of speech and expression, as well as the value of such self-regulatory models, based on international standards, to ensure transparent, responsible and principled management of content on digital social platforms.

Keywords

Oversight Board; Freedom of Speech; International Human Rights Standards

1. Как да ограничим неограничените възможности на социалните медии – генезис на модерирането на съдържание от „Мета“

Макар социалните медии да притежават значителен потенциал за положително въздействие върху общественото съзнание във връзка с неограничените си възможности за разпространение на информация, използвана за публични дебати, политически кампании и споделяне на знание, те крият и немалко рискове. Компании като „Мета“ (преди Facebook) регулират съдържанието, споделяно на техните платформи чрез хибридна система, която съчетава алгоритмична модерация и човешки надзор – процес, в който ролята на човешките права е едновременно централна и маргинализирана. За по-уязвимите групи те могат да осигурят видимост и застъпничество, но същевременно спомагат за поддържането на неравенства, резултат от субективни и дискриминационни практики на модериране.

В тази връзка глобалните дигитални платформи за комуникация все по-често се сблъскват с изискването да гарантират, че разпространяването чрез тях на съдържание става по начин, който зачита свободата на изразяване и осигурява на потребителите защита от вреди и потенциални рискове. Това налага правилата за събиране и разпространение на дигитална информация да съответстват на стандартите за информация извън онлайн пространството.

Нарастването на значимостта на социалните платформи създава нови пространства, в които хората могат да документират и споделят съдържание в условия на конфликти, войни и кризи. Публикации, които разкриват нарушения на човешките права, сцени на насилие и разрушения,

се превръщат в ключови източници за разбиране и анализ на социалната и политическата реалност. В същото време обаче тези платформи могат да бъдат използвани и за тяхното потискане, за негативни внушения и за създаване на междуобщностно и междуетническо напрежение чрез реч на омразата. Речта на омразата, която се характеризира се с очерняне и вербално насилие срещу лица или групи въз основа на защитени характеристики, е увеличила негативните психологически ефекти върху жертвите, като допринася за стрес, депресия и създаване на враждебна среда за уязвимите групи от населението[2].

Постигането на баланс в тази сложна среда изисква деликатен и добре преценен подход, който съчетава защитата на свободата на изразяване с необходимостта от опазване на човешките права във виртуалния свят, а това за компаниите, които поддържат такива социални мрежи[3], е сериозно предизвикателство. „Мета“ например модерира съдържание, генерирано дневно от 3,43 милиарда активни потребители[4] на основните продукти на компанията. Необходимостта от онлайн противодействие на речта на омразата стимулира създаването на различни системи за модериране на съдържание, както и на независими механизми за надзор. Те целят да минимизират негативните ефекти и въздействие на дигиталното слово, като гарантират свободата на изразяване във всичките ѝ форми. За изпълнението на тази задача „Мета“ създава система за модериране на съдържание, която предвижда изграждането на Надзорен съвет[5] като форма на отчетност и саморегулация за платформите, управлявани от компанията[6].

В тази връзка трябва да се отчете съдържанието на концепцията за „модерация“ като процес на проверка, оценка, одобряване или на потискане на комуникации от потребители на онлайн платформа[7]. Саморегу-

---

[2] J. M. Pérez *et al.* Assessing the Impact of Contextual Information in Hate Speech Detection, *IEEE*, 11, (2023), pp. 30575 – 30590, DOI: 10.1109/ACCESS.2023.3258973 [access 4.06.2025].

[3] Meta Platforms, Inc. (до 28.10.2021 г.) Facebook, Inc., е северноамериканска многонационална холдингова компания, майка на Facebook, Instagram, WhatsApp и Oculus, [online] https://www.meta. com/about/company-info/?utm_source=about.facebook.com&utm_medium=redirect [access 4.06.2025].

[4] Срвн. за първата четвърт на 2025 г., Statista: [online] https://www.statista.com/statistics/1092227/facebook-product-dau/ [access 4.06.2025].

[5] Oversight Board, [online] https://oversightboard.com/ [access 4.06.2025].

[6] Срвн. D. Endres, L. Hedler, K. Wodajo. Bias in Social Media Content Management: What Do Human Rights Have to Do with It? *AJIL Unbound*, 117 (2023), pp. 139 – 144, DOI:10.1017/aju.2023.23

[7] Срвн. B. Farrand. 'Is This a Hate Speech?' The Difficulty in Combating Radicalisation in Coded Communications on Social media Platforms. *European Journal on Criminal Policy and Research*, 29, 3 (2023), p. 477 – 493, esp. p. 5, https://doi.org/10.1007/s10610-023-09543-z

лирането на частни компании не е ново или необичайно явление. То традиционно се обсъжда от автори като Огус[8]. Той смята, че процесът може да се основава на икономически обосновки за разходите и ефективността, които произтичат от техническите знания и експертиза в тези сектори. Този механизъм е предложен и в контекста на работата на медиите, в Плана за действие от Рабат[9] относно забраната за пропагандиране на национална, расова или религиозна омраза на ООН, като най-подходящия начин да се адресират проблемите, свързани с тази дейност на медиите.

Механизмите за саморегулация, макар и концептуално установени, показват значителна сложност при прилагането си в онлайн среда. Тя произтича от динамичната и често непредсказуема природа на дигиталните екосистеми, където взаимодействията между потребители и платформи могат да имат както предвидими, така и непредвидими негативни последици за свободата на словото. Целта е при модерирането на съдържание социалните платформи да осигурят еднакви стандарти за преценката за наличието на нарушение на основните човешки права или при ограничаване на правото на свободно изразяване за предотвратяване от последващи вреди. Тук водещото е прилагането на установените в международното право принципи в областта на правата на човека.

Международното право в областта на правата на човека очертава пътя за осигуряването на единни стандарти за офлайн и онлайн речта, както се отчита от изследователи като Вишмайер и Петкова. Така може да се избегне разделението на режима за онлайн и офлайн модерирането на съдържание[10]. Петкова посочва, че това, което е незаконно офлайн, би останало незаконно онлайн и въпреки това „докато квазисъдебни, квазиконсултативни органи като Надзорния съвет на „Мета“ натрупват „съдебна практика“, по-належащият въпрос може да стане – трябва ли това, което е незаконно онлайн, да бъде незаконно и офлайн[11]“.

---

[8] A. Ogus. *Regulation: Legal form and economic theory*. Oxford and Portland, Hart Publishing 1994, pp. 110 – 111, *id.* Rethinking self-regulation. *Oxford Journal of Legal Studies*, 15, 1 (1995), pp. 97 – 108, DOI: 10.1093/ojls/15.1.97 [access 4.06.2025].

[9] United Nations Office of the High Commissioner for Human Rights (OHCHR), *Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred*, 2012, [online] https://www.ohchr.org/sites/default/files/Rabat_draft_outcome.pdf [access 4.06.2025].

[10] Срвн. T. Wischmayer. What is Illegal Offline is Also Illegal Online – The German Network Enforcement Act 2017. В. *Petkova, T. Ojanen, Fundamental Rights Protection Online: The Future Regulation of Intermediaries*. Elgar, 2020, pp. 43 – 67, http://dx.doi.org/10.2139/ssrn.3256498, подобно B. Petkova, Meta Oversight Board's Nascent Standard on Hate Speech: Towards Plural Standard Setting in International Human Rights Law", forthcoming in La Rivista di diritto dei media *(Media Laws)*, written: Nov. 10, 2025 [online] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5731110 forthcoming in *La Rivista di diritto dei media (Media Laws)* 2026. [access 05.06. 2025].

[11] Petkova (2025).

Поради това е важно да се открои значението на Надзорния съвет (НС) като елемент от саморегулацията на платформата с оглед прилагането на международните стандарти за закрила на правата на човека и в частност на свободата на словото и забраната на езика на омразата. В рамките на този процес централна роля играят няколко ключови въпроса – правните основания на НС, неговата компетентност, връзката на издадените от него решения и препоръки с международноправната рамка и по-конкретно с Международния пакт за граждански и политически права (МПГПП)[12], Плана за действие от Рабат относно забраната за пропагандиране на национална, расова или религиозна омраза на ООН[13] и Доклада на специалния докладчик за насърчаване и защита на правото на свобода на мнение и изразяване[14].

2. НАДЗОРНИЯТ СЪВЕТ – ОГРАНИЧИТЕЛ НА СВОБОДАТА НА СЛОВОТО, СЪД НА БЪДЕЩЕТО ИЛИ ОДИТОР НА ПОЛИТИКИТЕ НА „МЕТА“?

2.1. История и цел

През ноември 2018 г. Марк Зукърбърг публикува в отговор на скандала с „Кеймбридж аналитика“[15] и нарастващата критика срещу модерирането на съдържание пост във Фейсбук, озаглавен „A Blueprint for Content Governance and Enforcement“[16]. Това е план за създаване на независим надзорен орган, т.нар. Надзорен съвет (Oversight Board). Той служи за преразглеждане и обжалване решенията на платформата по

[12] UN General Assembly, *International Covenant on Civil and Political Rights, United Nations, Treaty Series*, vol. 999, p. 171, 16.12.1966, [online] https://www.refworld.org/legal/agreements/unga/1966/en/17703, [access 20.05. 2025].

[13] OHCHR (2012).

[14] D. Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression* UN Doc A/72/350 (18.08.2017) [online] https://digitallibrary.un.org/record/1304394?ln=en& v=pdf [access 4.06.2025].

[15] Това е значим случай на нарушаване цифровата сигурност и защитата на личните данни, разкрит през 2018 г. Неправомерно се събират и използват данни на близо 87 милиона потребители на фейсбук чрез приложение, създадено от Александър Коган. То се използва от около 270 000 души, но получава достъп и до данните на техните приятели, без изричното им съгласие. Тази информация е предоставена на британската фирма за политически консултации „Кеймбридж аналитика“ за създаване на психологически профили на избиратели с цел политическо влияние, включително в кампанията за Брекзит и изборите в САЩ през 2016 г. Случаят предизвика сериозни опасения за ролята на социалните мрежи в манипулацията на общественото мнение и демократичните процеси, срвн. [online] https://www.sciencedirect.com/science/article/abs/pii/S1071581920301002 [access 4.06.2025].

[16] M. Zuckerberg, *A Blueprint for Content Governance and Enforcement* (Facebook, 5.05.2021) [online] https://m.facebook.com/nt/screen/?params=%7B%22note_id%22%3A751449002072082%7D&path=%2Fnotes %2Fnote%2F&_rdr [access 20.05. 2025].

премахването на съдържание. Целта е да се осигури по-голяма прозрачност и отчетност в процесите на модериране въз основа на принципите на върховенството на закона и международните стандарти за правата на човека.

През 2020 г. Надзорният съвет е формално конституиран и представлява ключова иновация в саморегулацията на частни цифрови платформи. За разлика от традиционните механизми за контрол върху словото от държавни органи, обвързани с международноправни задължения, това е частна структура. Тя се установява доброволно съгласно вътрешноинституционалното си правила и по-конкретно „Харта на Надзорния съвет“[17], която възприема да прилага международни норми в областта на защитата на свободата на словото. Това го позиционира като своеобразен „извънсъдебен“ орган, който функционира с делегирана корпоративна компетентност, в рамките на вътрешната структура на компанията. Същевременно той трябва да прилага международното право в областта на човешките права и по-конкретно в областта на свобода на словото.

Според Хартата на Надзорния съвет неговата цел е „да защитава свободното изразяване чрез вземане на принципни, независими решения относно важни части от съдържанието“ и да издава консултативни становища по политиките. Това осигурява справедливост, прозрачност и последователност при решенията, а те имат значително влияние върху изразяването на потребителите[18].

Функцията и правната природа на Съвета се интерпретира различно. Евелин Дуек например го разглежда като „смел експеримент в саморегулацията“, но подчертава, че легитимността му е частична, тъй като „Мета“ запазва решаваща роля при определяне на политиките и границите на правомощията му. Тя отбелязва, че въпреки структурната автономия, реалната независимост е ограничена, особено в случаи, свързани с политически чувствителни теми[19]. Кейт Клоник вижда НС като опит да се създаде съдебен модел в рамките на частна технологична платформа. Тя го нарича „частен върховен съд“, който цели да въведе върховенство на правото в модерирането на съдържание. Според нея обаче липсата на реална принуда върху „Мета“ и ограниченото прилагане на

[17] Oversight Board, Charter of the Oversight Board, 2019 [online] https://oversightboard.com/governance/#charter [access 4.06.2025].

[18] Meta Platforms, Inc. *Oversight Board Charter*. Menlo Park, CA: Meta, September 2019. Introduction, [online] https://about.fb.com/wp-content/uploads/2019/09/oversight_ board_ charter.pdf [access 4.06.2025].

[19] E. Douek, *The Meta Oversight Board and the Empty Promise of Legitimacy*, Harvard Journal of Law & Technology, Vol. 37, No. 2, (2024), pp. 373 – 445, DOI: 10.2139/ssrn.4565180, [access 4.06.20 25].

препоръките поставят под въпрос ефективността на този модел[20]. Друг дискутиран въпрос засяга приноса на НС за увеличаване на прозрачността и отчетността в дигиталната среда[21], както и критериите „законност, необходимост, пропорционалност“ при изразяването онлайн. Тази стандартизация цели по-голяма защита на свободата на изразяване[22], като НС е представен като реалистичен механизъм в това отношение и особено полезен за маргинализирани езикови общности[23]. Механизмът е също критикуван на базата на либертарно разбиране за икономическите свободи[24].

Въпреки различните становища безспорно е, че Съветът представлява уникална институционална конструкция в сферата на дигиталното управление на съдържание и надзорен орган без аналог към този момент. Той е безпрецедентен опит за създаване на саморегулация чрез извънсъдебен, независим орган в рамките на частна корпорация, който да осигурява нормативно ръководство относно свободата на изразяване и модерирането на съдържание.

За разлика от държавите, които имат юридически задължения по международните договори за правата на човека, частните компании като „Мета“ не са формални носители на отговорност по международното право. Въпреки това, структурата и дейността на НС са усилие, от една страна, да се обвърже контролната му функция с обективното право, което е свързано с държавно-организираната реалност. От друга, се запълва нормативната празнина във връзка с отговорността на социалните медии и компаниите, които стоят зад тях. Инструмент за това е прила-

[20] K. Klonick, The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression, *Yale Law Journal*, 129(8), (2020), pp. 2418 – 2499, [access 4.06.2025].

[21] D. Wong and L. Floridi. Meta's Oversight Board: A Review and Critical Assessment. *Minds & Machines*, 33 (2023), pp. 261 – 284, https://doi.org/10.1007/s11023-022-09613-x; N. Suzor, T. Van Geelen, S. Myers West. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette*, 80, 4 (2018), pp. 385 – 400, [online] https://doi.org/10.1177/1748048518757142, [access 06.06.2025].

[22] S. Di Stefano. Translating and Developing International Human Rights Law in the Online Sphere: The Role of Meta's Oversight Board, *I. Couzigou and E. Fromageau (eds.), International Law and Technological Change: Testing the Adaptability of International Law*, Cheltenham, Northampton, Edward Elgar Publishing, 2025, [online] SSRN: https://ssrn.com/abstract=4920875 or http://dx.doi.org/10.2139/ssrn.4920875 [access 04.06 2025], esp. pp. 10 – 13 за чл. 19 от МПГПП и прилагането на пропорционалността в решенията на НС.

[23] J. C. York. Meta Oversight Board's Latest Policy Opinion a Step in the Right Direction. *Electronic Frontier Foundation*, 01.2024, [online] https://www.eff.org/deeplinks/2024/03/meta-oversight-boards-latest-policy-opinion-step-right-direction, [access 4.06.2025].

[24] B. Zankova and V. Dimitrov, Social Media Regulation: Models and Proposals, Journalism and Mass Communication, 10, 2 (2020), pp. 75 – 88.

гането от частни субекти на правилата за закрила на свободата на словото, заложени в МПГПП. „Мета“ има традиция в тази посока с приложението в практиката на Надзорният съвет[25] на Ръководните принципи на ООН относно бизнеса и правата на човека („Ръководни принципи“)[26]. Така се гарантира, че практиките на „Мета“ по модериране на съдържание зачитат основните човешки права и особено свободата на изразяване и се осигурява ефективен и прозрачен достъп до правна защита.

Освен това НС извършва „оценки на въздействието на решенията си върху правата на човека“ (human rights impact assessment), за да идентифицира и смекчи потенциални рискове, в съответствие с Принцип 31 от Ръководните принципи[27]. След публикуването на съответното решение се наблюдава неговото въздействие, за да се установи реалното му влияние и се правят изводи. Тези механизми имат допълваща роля за гарантирането на зачитането на човешките права и усилията на „Мета“ и НС в тази посока. Съветът не само осигурява съответствие, но и подобрява защитата правата на човека и минимизирането на рисковете[28], контролира корпоративната власт и внедрява стандартите за свободата на изразяването в дигиталната сфера. Така се очертават границите на корпоративната отговорност за онлайн вреди и се изяснява приложението в социалните медийни платформи на нормите за правата на човека.

С оглед на комплексната правна природа на НС е важно да се изследват неговата структура, функции, цели, като и тълкуването и приложението на международното право и стандартите в областта на правата на човека и по-конкретно в свобода на изразяване, как това е отразено в платформената политика.

---

[25] The Oversight Board. *Operationalizing the UN Guiding Principles on Business and Human Rights (Submission to the Office of the High Commissioner for Human Rights*, 02.2022, [online] https://oversightboard.com/decision/ [access 31.05.2025].

[26] OHCHR, *Guiding Principles on Business and Human Rights: Implementing the United Nations „Protect, Respect and Remedy“ Framework* (UN Doc A/HRC/17/31, 21 March 2011) [online] https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights, [accessed 31.05.2025]. Ръководните принципи имат три глави, защита, зачитане и средства за защита. Всяка от тях определя конкретни, приложими стъпки за правителствата и компаниите, за да изпълнят съответните си задължения и отговорности за предотвратяване на нарушения на правата на човека в дейността на компаниите и да осигурят правна защита, ако такива нарушения се случат. Вж. допълнително OHCHR, [online] https://www.ohchr.org/sites/default/files/Documents/Issues/Business/Intro_Guiding_PrinciplesBusinessHR.pdf [accessed 31.05.2025].

[27] The Oversight Board. *Operationalizing the UN Guiding Principles on Business and Human Rights. Submission to the Office of the High Commissioner for Human Rights, United Nations on the practical application of the UNGPs to the activities of technology companies*, 02.2022, p. 5, [online] https://www.ohchr.org/sites/default/files/2022-03/Oversight-Board.pdf [access 04.06.2024].

[28] Пак там.

2.2. Структура и управление на Надзорния съвет

НС функционира като независим орган, основан чрез неотменим тръст от „Мета“ от 130 милиона щатски долара, които да гарантират институционална автономия[29]. В състава му могат да бъдат включени до 40 довереници от различни професионални и географски среди, избирани чрез многоетапен процес, включващ съпредседатели, „Мета“ и независими[30].

Управленската рамка на Съвета е изградена върху три основни акта, Хартата[31], която определя неговата мисия, функции и връзката му с „Мета“; Правилникът[32] – регламентира процедурите и отговорностите на членовете; и Тръстовият договор[33] – осигурява финансовата и административната автономия на този външно корпоративен орган.

Основната цел на НС е да съчетае защитата на свободата на изразяване със спазването на международните стандарти за правата на човека, прилагани към процесите по модериране на съдържание. Компетентността обхваща преди всичко разглеждането на индивидуални казуси по жалби директно от потребители или препратени от „Мета“ след изчерпване на общите вътрешни платформени механизми за апелация. Приоритет се дава на тези със значителен обществен интерес и потенциал да формират бъдеща платформена политика[34]. Освен конкретни казуси Съветът има мандат да предоставя консултативни становища по политики, които, макар и незадължителни, могат да окажат влияние върху платформеното регулиране на съдържанието[35]. Тези становища се оформят на основата на принципи като правна яснота, пропорционалност и съразмерност на ограниченията на речта, съгласно международните стандарти по чл. 19, ал. 3 от Международния пакт за граждански и политически права.

Важен аспект на работата на НС е осигуряването на прозрачност при вземането на окончателните решения и отчетност. В тази връзка всички решения се публикуват публично, с цел повишаване на доверието в модерирането и улесняване на академичен и обществен контрол върху

[29] Oversight Board, Oversight Board Charter (2019) https://oversightboard.com/governance/ [access 04.06.2025].

[30] Пак там, с. 7 – 8.

[31] Oversight Board Charter, [online] https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf, [достъпен на 4 юни 2025].

[32] Oversight Board. (2023, November). *Oversight Board Bylaws*. https://www.oversightboard.com/wp-content/uploads/2023/11/926998147956593.pdf, [достъпен на 4 юни 2025].

[33] Oversight Board Trust Agreement, [online] https://about.fb.com/wp-content/uploads/2019/12/Trust-Agreement pdf [достъпен на 4 юни 2025].

[34] Пак там, с. 10 – 11.

[35] Пак там, с. 12.

платформената политика[36]. Дадена е възможност за участие и на широката общественост чрез подаването на становища от самостоятелни лица, организации и други заинтересовани страни. Решенията относно конкретни случаи са задължителни за „Мета“, освен ако тяхното прилагане би противоречало на местното законодателство. За разлика от това, политическите препоръки са със незадължителен характер и „Мета“ има право да прецени дали да ги изпълни.

Посредством структурата и делегираната му компетентност НС се позиционира като извънсъдебен механизъм в рамките на дигиталната сфера създавана от Фейсбук платформата. Това е опит за институционализиране на отговорността на социалните медии, като същевременно утвърждава нормативна рамка, приложима за дейността на онлайн платформите. Тя се основава на международното право, без да дублира ролята на държавата и наднационалните институции като ООН. Какъв обаче е процесът на утвърждаването на тази нормативна рамка и каква е нейната правно обвързваща сила?

Параметрите на приемливите форми на онлайн изразяване остават оспорвани и неясни, а платформите на социалните медии са изправени пред постоянни предизвикателства при тълкуването и прилагането на принципите на правата на човека към своите практики за модериране на съдържание. Това допълнително се усложнява от културните различия и езиковите нюанси, които изискват локализиран подход на модерирането на съдържание в съответствие със специфичния контекст, културната и историческа чувствителност на различните региони. Сложността на противодействието на речта на омразата се утежнява от глобалния обхват на платформите на социалните медии, които надхвърлят националните граници и правните юрисдикции. Тълкуването и прилагането на тези принципи в онлайн контекст са предмет на текущ дебат особено за обхвата на допустимите ограничения на свободата на изразяването и отговорностите на социалните медии при модерирането на съдържание. Те често разчитат на наказателни подходи, като забрани и премахване на съдържание, за да се справят с вредното съдържание.

Някои учени предлагат отдалечаване от чисто наказателния (репресивен) подход и насочване към т.нар. възстановително правосъдие, допълнителен механизъм за гарантирането и защитата на човешките права[37]. Този модел се отстоява от Сарита Шонебек и Линдзи Блекуел и поставя в центъра нуждите на пострадалите. При него чрез посредничество се търси признаване на причинената вреда и се насърчава поемането на отговорност[38]. За разлика от настоящите практики на санкции

[36] Oversight Board Trust Agreement, [online] https://about.fb.com/wp-content/uploads/2019/12/Trust-Agreement pdf [достъпен на 4 юни 2025], с. 13.

[37] Вж. S. Schoenebeck and L. Blackwell. Reimagining Social Media Governance: Harm, Accountability, Repair, *Yale Journal of Law and Technology (Special Issue)*, 23 (2021), p. 113.

[38] Schoenebeck and Blackwell (2021), p. 113, 138.

като блокиране или премахване на съдържание често без прозрачност и без отчитане преживяванията на жертвите, възстановителното правосъдие предлага всеобхватен подход. Той включва посредничество между извършителя и потърпевшия, публични признания, извинения и дори структурни промени в дизайна на платформите за предотвратяване на бъдещи вреди[39].

На този етап НС противодейства на вредите, свързани със съдържанието в социалните медии, като разглежда конкретни случаи и предоставя препоръки за политики, а не чрез методи на възстановително правосъдие. Това включва основно задължителни решения за премахване или възстановяване на съдържание, както и незадължителни препоръки към „Мета“ за подобряване на практиките по модериране. Действията на НС са в съответствие с Ръководните принципи на ООН за бизнеса и правата на човека, особено за отчетността и правната защита (Принципи 25 – 31). Те не са форми на възстановително правосъдие и вместо да насърчават диалог, признаване на вредата от извършителите или възстановяване на щети в рамките на общността, попадат в рамките на правно-формалните и процедурните средства. Фокусът е върху политиките на платформата и прилагането на международните стандарти в областта на правата на човека.

3. ИМА ЛИ ПРИЛОЖЕНИЕ И РОЛЯ МПГПП В МОДЕРИРАНЕТО НА СЪДЪРЖАНИЕ?

3.1. Правната рамка

Свободата на изразяване е гарантирана от чл. 19 от МПГПП, който гарантира всеобщата свобода на мнение и изразяване. Това включва и правото на тяхното свободно отстояване, както и получаването и разпространяването информация и идеи независимо от медиите и границите. Това е основен стълб на демократичните общества[40], но не представлява абсолютно право[41]. В тази връзка както международното право, така и повечето национални конституции признават, че свободата на изразяване може да бъде ограничена. Всички ограничения обаче трябва да останат в строго определени граници и при строго определени условия. Член 19, ал. 3 от МПГПП допуска ограничения на свободата на изразя-

[39] Schoenebeck and Blackwell (2021), p. 38.

[40] D. Adjei. Freedom of Expression and its Legal Consequences in the Era of Social Media. *Amicus Curiae*, Vol. 5, No. 3 (2023), [online] https://journals.sas.ac.uk/amicus/article/view/5712, [access 05.06.2025]

[41] Art. 19, *ICCPR: Links between Articles 19 and 20* (Article 19, 2008) [online] https://www.article19.org/ data/files/pdfs/ conferences/iccpr-links-between-articles-19-and-20.pdf [access 05.06.2025].

ване само като изключение, когато тя е свързана със специални задължения и отговорности. Те са допустими единствено, ако кумулативно са изпълнени три условия: да са предвидени от закон, да преследват легитимна цел (защита на правата и репутацията на другите, националната сигурност или обществения ред, общественото здраве или морала) и да бъдат необходими и пропорционални[42] в едно демократично общество. Липсата на което и да е от тези условия води до нарушение на чл. 19 от МПГПП. Задължение държавата е да докаже, че ограничението е наложено при наличие и на трите условия[43]. Критериите „законност", „необходимост" и „пропорционалност" са известни като „тристъпковия тест". Той се прилага към всяко ограничение на свободата на словото, за да се валидира неговата законност.[44] Тези критерии са кумулативни условия за законността на ограниченията и са задължителни стъпки в анализа на пределите на свободата на словото.[45]

3.2. Взаимовръзка между международното право и практиката на НС

3.2.1. Приложение на тристъпковия тест на чл. 19 (3) от МПГПП

Надзорният съвет изрично прилага тристъпковия тест на чл. 19 (3) от МПГПП, за да определи дали премахването, или възстановяването на съдържание е в съответствие с международните стандарти. Този подход е установен в решенията по делата му и отразява ангажимента му към Ръководните принципи на ООН за бизнеса и правата на човека.

Тази рамка се прилага в решенията, свързани с речта на омразата и подбуждане към насилие, като се уповава на своеобразен списък за проверка за съответствие със стандартите на международното право в областта на човека. Придържането към него е очевидно от повтарящите се стъпки в юридическия анализ на всяко едно решение. Така например в случая **„Арменци в Азербайджан"**, Съветът потвърждава премахването на съдържание, съдържащо етническа обида, като в съответствие с чл. 20, ал. 2 от МПГПП[46] се позовава на потенциала за вреда в контекст на етническо напрежение. В решението по случая **„Изображение на Зварте Пит"** Съветът приема, че „черното лице" (blackface) представ-

---

[42] Вж. *Romanchik and Shchukina v. Belarus, (CCPR/C/135/D/2917/2016),* [online] https://juris.ohchr.org/casedetails/3679/en-US [access 20.05.2025].

[43] Срвн. Communications No. 1830/2008, *Pivonos v. Belarus,* Views, 29.10.2012, п. 9.3 и № 1785/2008, *Olechkevitch v. Belarus, Views,* 18.03.2013, п. 8.5.

[44] UN Human Rights Committee (HRC), General comment no. 34, Art. 19, Freedoms of opinion and expression, CCPR/C/GC/34, 12 September 2011 [online] https://www.refworld.org/legal/general/hrc/2011/en/83764 [access 20.05.2025].

[45] Вж. *Androsenko v. Belarus (CCPR/C/116/D/2092/2011), n. 7.3,* [online] https://juris.ohchr.org/casedetails/2103/en-US [access 05.06.2025].

[46] Oversight Board, Armenians in Azerbaijan, 2021 [online] https://oversightboard.com/decision/fb-qbjdascv/ [access 20.05.2025].

лява расов стереотип и подкрепя премахването на съдържанието съгласно стандартите на международното право по правата на човека[47]. В решението при **„Рая Кобо“** се премахва публикация, която обвинява етническа група в престъпления, без доказателства – акт с потенциал да подбуди насилие по време на въоръжен конфликт[48]. В **„Карикатурата от Книн“** първоначално „Мета“ оставя публикацията, но Съветът постановява премахване на изображение, което приравнява сърби на плъхове – смятано за класически пример за дехуманизираща реч.[49] В случая **„Отричане на Холокоста“** Съветът отхвърля решението на „Мета“ да не премахва публикацията, като приема, че отричането на геноцид представлява реч на омраза, която трябва да бъде забранена.[50]

Във всички тези случаи Съветът прилага последователно тристъпковия тест от чл. 19, ал. 3 на МПГПП. Изискването за „законност“ се установява чрез препратка към ясно формулирани стандарти на общността. Легитимността се анализира чрез идентифициране на заплаха за защитените права – като достойнство, недискриминация и обществения ред. Най-важна е оценката на обстоятелствата налагащи необходимост от вземане на мерки и тяхната пропорционалност, основана на контекстуалните рискове от съответната реч и особено в региони с история на насилие или напрежение. Този подход е в съответствие с тълкуванията на Специалния докладчик на ООН по свободата на изразяване, който подчертава, че ограниченията трябва да са специфични за контекста и най-малко да засягат негативно правата[51].

Съществуват обаче някои различия в тълкуването на тези стандарти в зависимост от политическия и културен контекст. В конфликтни зони (като Етиопия например), Съветът приема по-консервативен подход към потенциално подбуждаща реч, като отчита по-високия риск от реално насилие. За разлика от това, в решенията по случаи като **„Зварте Пит“** и **„Отричане на Холокоста“** се отдава особено значение на историческата памет и опасността от нормализиране на расизъм или отричане на геноцид. Тези различия подчертават развитието в практика на Съвета, който цели да хармонизира модерирането на съдържание с международните правозащитни норми, като същевременно отчита културния и исторически контекст във всеки конкретен случай.

---

[47] Oversight Board, Depiction of Zwarte Piet, 2021, [online] https://oversightboard.com/decision/fb-s6nrtdaj/ [access 20.05.2025].

[48] Oversight Board, Alleged Crimes in Raya Kobo, 2021, [online] https://oversightboard.com/decision/fb-mp4zc4cc/ [access 20.05.2025].

[49] Oversight Board, Knin Cartoon, 2022 [online] https://oversightboard.com/decision/fb-jrq1xp2m/ [access 20.05.2025].

[50] Oversight Board, Holocaust Denial, 2024, [online] https://www.oversightboard.com/decision/ig-zj7j6d28/ [access 20.05.2025].

[51] UN Special Rapporteur on Freedom of Expression, A/HRC/44/49, 2020.

Чрез прилагането на международните правозащитни норми работата на Съвета допринася към нарастващия корпус от т.нар. „дигитален конституционализъм“[52]. Обвързването с международното право е причината НС да се разглежда от някои учени като въплъщение на дигитален конституционализъм. Той действа като механизъм за надзор, базиран на правата на човека и на независимостта, като използва юридическа аргументация и уподобява съдебна структура в рамките на платформената екосистема. Така се различава от т.нар. „трансформативен (обществен) конституционализъм“, свързван с активното участие на потребителите на всеки етап от създаването на нормативната база и политики на платформите. Въпреки че НС допуска публични коментари към казусите предложени за разглеждане, в основната му структура и модел на работа липсва реално включване на потребителите в създаването на корпоративните политики. Така НС може да се разглежда и като експериментален орган създаващ частен дигитален конституционализъм. НС изпълнява функции, аналогични на конституционен контрол – интерпретация на норми, процедурна легитимация и защита на основни права – но без класическите характеристики на публична власт и демократична отчетност. Това поставя под въпрос доколко подобни механизми могат устойчиво да допринасят за запазването на либералните конституционни ценности в глобалната дигитална среда[53].

3.2.2. Критерият „риск от подбуждане“ в международното право по правата на човека и в практиката на НС

Държавите са задължени да забранят не само проповядването на омраза, но и подбудата към дискриминация, нападение или маргинализиране на хора. Член 20, ал. 2 от МПГПП изисква законова забрана за проповядването на национална, расова или религиозна омраза, което подбужда към дискриминация, враждебност или насилие. Терминът „подбуждане“ включва форми на изразяване, които активно подтикват или насърчават дискриминацията, враждебността или насилието срещу лица или групи въз основа на тяхната национална, расова или религиозна идентичност. Разпоредбата налага позитивно задължение на държавите да ограничат такива форми на изразяване, които надхвърлят защитената

[52] В този подход частни технологични платформи се подчиняват на принципите на международното право.

[53] Вж. Pollicino, Oreste and Paolucci, Federica, Digital Constitutionalism (December 31, 2024). L. Floridi, M. Ziosi, M. Taddeo, *Companion to Digital Ethics*, Oxford University Press, forthcoming, SSRN: https://ssrn.com/abstract=5098492 or http://dx.doi.org/10.2139/ssrn.5098492 [access 4.06.2025], където авторът разглежда трансформациите на конституционализма в дигиталната среда с особен акцент върху либералните режими, хибридизацията на публичната и частната власт и предизвикателствата пред защитата на основните права извън рамките на държавата.

„свобода на словото“ и подбуждат към увреждане на други свободи, особено изрично защитени, раса, националност или религия**.**

НС въвежда нов критерий в практика си по чл. 20, ал. 2 и чл. 19, ал. 3 от МПГПП, „потенциалният или реален риск от подбуждане“. Биляна Петкова го определя като „риск от подбудителство“ и го приема за „новаторски елемент“ в анализа на „езика на омразата“. НС го използва в анализа на отделните случаи предвид културния, социалния или политическия контекст[54]. Тя подчертава, че: „докато елементите намерение и вреда в общи линии съответстват на факторите от Рабатския план, елементът подбуждане се съдържа изрично в самия текст на чл. 20 от Пакта“, където се смята, че той е налице само когато дадено изказване „създава непосредствен риск от дискриминация, враждебност или насилие спрямо лица, принадлежащи към съответната (засегната) група“[55]. Така рискът от подбуждане не само играе съществена роля при решенията, свързани с реч на омразата, но и изпълнява главна функция в прилагането на тристепенния тест по чл. 19, ал. 3 от МПГПП. Друго проявление на „подбуждането към непосредствено насилие“ на международно равнище се свързва с Резолюция 16/18[56] на Съвета по правата на човека, насочена срещу религиозната нетърпимост, като избягва законодателни забрани срещу богохулство и затвърждава значимостта на този нов критерий и систематичното му тълкуване на международно ниво[57].

Тази мярка има особено значение при анализа на критериите за необходимост и пропорционалност. Според Общия коментар № 34 на КПЧ на ООН, „ограниченията не само трябва да служат на някоя от изброените легитимни цели; те също така трябва да са необходими за тяхната защита“[58]. Необходимостта предполага оценка, която да отчита дали изявлението представлява „належаща обществена нужда“[59] и дали са използвани „най-малко инвазивните средства“ за постигане на целта[60]. В този контекст НС последователно разглежда риска от подбуждане

[54] Petkova (2025), p. 3.

[55] *Cnfr*. Petkova (2025), p. 4.

[56] Resolution 16/18: Combating Intolerance, Negative Stereotyping and Stigmatization of, and Discrimination Incitement to Violence and Violence Against, Persons Based on Religion or Belief. A/HRC/RES/16/18, [online] https://www2.ohchr.org/english/bodies/hrcouncil/docs/16session/a.hrc.res.16.18_en.pdf [access 05.06.2025].

[57] Пак там, с. 5.

[58] UN Human Rights Committee, General Comment No. 34: Art. 19 – Freedoms of Opinion and Expression (12 September 2011) UN Doc CCPR/C/GC/34, [online] https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf [access 04.06.2025], § 22.

[59] UN Human Rights Committee, General Comment No. 34: Art. 19 – Freedoms of Opinion and Expression (12 September 2011) UN Doc CCPR/C/GC/34, [online] https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf [access 04.06.2025], § 33.

[60] Пак там, § 34.

като доказателство за необходимост. В случаите Арменци в Азербайджан[61] и Рая Кобо[62] се заключава, че изявлението, насочено към етнически групи, дори и да не е пряко насилствено, може значително да допринесе за „висок риск от изостряне на напрежението и причиняване на вреда в нестабилни региони“. В решението по случая Карикатурата от Книн[63] се отбелязва, че целта на видео, което изобразява хората като плъхове, е „да подбуди етническа омраза, което може да допринесе за извършването на дискриминационни действия от страна на отделни лица“. Подчертава се не само обидното съдържание, но и потенциалът да се възроди враждебност в постконфликтна среда. За разлика от това, в случаи със значима историческа чувствителност като Зварте Пит[64] и Отричане на Холокоста[65] фокусът се измества към превантивно разбиране за подбуждането, основано на кумулативните ефекти от расови стереотипи и исторически реваншизъм. Това е в съзвучие и с Плана за действие от Рабат, който призовава за ограничения не само въз основа на съдържанието, но и според контекстуални фактори като „положението или статуса на говорещия“, „степента на разпространение“ и „вероятността, включително нейната близост (непосредственост), за вреда“[66]. Както отбелязва Сюзън Бенеш в своята концепция за „опасна реч“, масовата вреда не винаги е незабавна, но може да възникне чрез „постоянно нормализиране на дехуманизиращи наративи“[67]. Следователно рисковият фактор при подбуждане се утвърждава като решаващ елемент при установяване, че ограничението е не само легитимно, но и пропорционално спрямо предотвратимата вреда. Това укрепва хармонизираното приложение на тристъпковия тест за законност, легитимност и пропорционалност в модерирането на съдържание в онлайн среда.

Решаващи елементи са критериите, заложени в Плана за действие от Рабат. Той предлага шест критерия, които трябва да се използват при оценката дали дадено изразяване представлява подбуждане към дискриминация, враждебност или насилие: „(1) контекст, (2) статус на говорещия, (3) намерение, (4) съдържание и форма на изразяване, (5) обхват на

[61] Oversight Board, Armenians in Azerbaijan, case number 2020-003-FB-UA, 2021, [online] https://oversightboard.com/decision/fb-qbjdascv/ [access 04.06.2025].

[62] Oversight Board, Alleged crimes in Raya Kobo, case number 2021-014-FB-UA, 2021, [online] https://oversightboard.com/decision/fb-mp4zc4cc/ [access 04.06.2025].

[63] Oversight Board, Knin cartoon case number 2022-001-FB-UA, 2022, [online] https://oversightboard.com/decision/fb-jrq1xp2m/ [access 04.06.2025].

[64] Oversight Board, Depiction of Zwarte Piet, case number 2021-002-FB-UA 2021, [online] https://oversightboard.com/decision/fb-s6nrtdaj/ [access 04.06.2025].

[65] Oversight Board, Holocaust Denial, 2024, case number 2023-022-IG-UA, [online] https://www.oversightboard.com/decision/ig-zj7j6d28/ [access 04.06.2025].

[66] OHCHR, Rabat Plan of Action, 2012, § 22.

[67] S. Benesch. Dangerous Speech: A Proposal to Prevent Group Violence, World Policy Institute (2012), p. 5.

разпространение и (6) вероятност, включително нейната близост (непосредственост), за причиняване на вреда“[68]. НС не се позовава изрично на тези фактори във всяко свое решение, но прилага техните принципи в практиката си. В случая Рая Кобо в контекста на въоръжен етнически конфликт, анонимността на говорещия и внушението за колективна вина спрямо населението на Тиграй сочат към висок риск от подбуждане. Той съответства на критериите за контекст и вероятност от вреда. Подчертава се, че съдържанието е имало потенциал да „разпали напрежение и да подкопае усилията за мир“[69]. В Карикатурата от Книн визуалната форма на изразяване (изобразяване на сърби като плъхове) в съчетание с историческия контекст на Балканските войни повишава риска от нормализиране на етническа омраза, като попада под критериите за съдържание и контекст[70]. В случая Отричане на Холокоста факторът „намерение“ се извежда от съзнателното отричане на установени исторически факти за Холокоста, което според Съвета „може да допринесе за антисемитизъм и подкопаване на паметта на жертвите“[71].

Тези примери показват, че дори при различна тематика и културен контекст Съветът прилага качествена и контекстуална оценка, вдъхновена от модела на Плана за действие от Рабат, за да прецени дали дадена публикация преминава границата от свобода на изразяване към подбуждане към омраза. Така се потвърждава, че допълнителните 6 критерия от Плана за действие от Рабат, както и тристъпковия тест на чл. 19, ал. 3 са интегрирани в анализа на съдържание в съответствие с принципите на международното право в областта на правата на човека в практиката на Съвета.

Заключение

Надзорният съвет на „Мета“ представлява новаторски механизъм в управлението на социалните платформи, функциониращ като извънсъдебен и външен регулаторен орган. Той прилага международното право в областта на правата на човека при разрешаването на конфликтни въпроси, свързани със свободата на словото и използване на език на омразата. Като частна, но независима структура, Съветът черпи легитимност

[68] OHCHR, Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, 2012, § 22. Официален текст, [online] https://www.ohchr.org/en/documents/publications/rabatplan-action [access 04.06.2025].

[69] Oversight Board, Case Decision: Alleged Crimes in Raya Kobo, 2021, [online] https://oversightboard.com/decision/fb-mp4zc4cc/ [access 04.06.2025].

[70] Oversight Board, Case Decision: Knin Cartoon, 2022, [online] https://oversightboard.com/decision/fb-jrq1xp2m/ [access 04.06.2025].

[71] Oversight Board, Case Decision: Holocaust Denial, 2024, [online] https://oversightboard.com/news/33289 9856283341 [access 04.06.2025].

от ангажимента си към глобални стандарти и най-вече МПГПП, видно от изричното му позоваване в решенията, свързани с оценка на съдържание и съответствието му с международните стандарти. Неговата дейност запълва празнината между политиките на технологичните компании и международно право в областта на правата на човека. Това укрепва идеята, че правилата за големите платформи трябва да са идентични както онлайн, така и офлайн. Както отбелязва С. ди Стефано: „Съветът показа как международните права на човека могат да бъдат тълкувани и прилагани спрямо предизвикателствата, поставени от новите технологии“[72].

Системно в аргументацията на Съвета заляга прилагането на тристъпковия тест по чл. 19, ал. 3 от МПГПП, който изисква всяко ограничение на изразяването да бъде (1) установено със закон, (2) да преследва легитимна цел и (3) да бъде необходимо и пропорционално на целите, които преследва. Тази рамка се използва последователно, като се преценява дали премахването на дадено съдържание е в съответствие с ясно формулирани политики („законността“), цели защита на потребителите от дискриминация и вреда – реална или потенциална („легитимност“), както и дали санкцията е ограничена само до необходимото („пропорционалност“). Тази правна структура гарантира, че модерирането е подчинено на утвърдени правила и базирано на аргументи, свързани със защита на достойнството и обществения морал. Както подчертават Хелфър и Ланд:

> „Съветът вече използва стандартите на правата на човека, за да разшири своя авторитет, да развива правни норми и да влияе върху регулацията на платформите“[73].

Един от най-съществените приноси на Съвета е усъвършенстваното прилагане на критерия „риск от подбуждане“ като решаващ фактор при определянето на необходимостта от ограничение и сваляне/изтриване на съдържание онлайн. Съветът анализира контекста, съдържанието, намерението и вероятността от вреда, като изхожда от разпоредбата на чл. 20, ал. 2 на МПГПП и принципите на Плана за действие от Рабат, когато решава дали се достига прага на подбуждане към омраза, дискриминация, насилие, етническа нетърпимост и пр. МПГПП осигурява правна рамка, а Планът за действие от Рабат дава допълнително тълкуване на обхвата на тази правна рамка чрез шест допълнителни критерия, който позволяват вземане на решение въз основа на структурен риск, а не само на непосредствена заплаха. Решенията на Съвета показват как терминът „подбуждане“ се тълкува от Съвета не само в смисъл на пряк призив към

[72] Di Stefano (2025).

[73] L. R. Helfer, M. K. Land. The Meta Oversight Board’s Human Rights Future. *Cardozo Law Review*, 44, 6, (2023), pp. 2233 – 2300, DOI: 10.2139/ssrn.4197107, [online] SSRN: https://ssrn.com/abstract=4197107, [access 4.06.2025].

насилие, а и като изявление, което приема омразата за нормална, разрушава социалните принципи за поведение или поддържа дискриминация (вж. случаите „Карикатурата Книн“, „Рая Кобо“ и „Отричане на Холокоста“). Това е контролен механизъм, който не само проверява съответствието на съдържанието със стандартите на „Мета“, но и оценява дали тези стандарти и тяхното прилагане са съвместими с международноправните задължения за защита на свободата на словото и другите човешки права.

В перспектива НС зависи от способността да съхранява независимостта си, да изгражда последователна съдебна практика и да отчита разнообразието от правни и културни контексти. В епоха на дигитализирани конфликти и транснационално слово, тълкувателната роля на Съвета ще бъде решаваща за формирането на норми за реакцията на платформите към речта на омраза, дезинформацията и отричането на исторически престъпления. Продължаващата му ангажираност с международните стандарти за права на човека, в съчетание с инструменти като Плана за действие от Рабат, го позиционира като модел за правозащитно управление в технологичния сектор. Както обобщават Хелфър и Ланд:

> „Съветът има потенциал да бъде реален контролен механизъм спрямо „Мета“ и да допринесе съществено за утвърждаването на правата на човека онлайн“[74].

Използвани източници

Adjei, D. Freedom of Expression and its Legal Consequences in the Era of Social Media, *Amicus Curiae*, 5(3), (2023), [online] https://journals.sas.ac.uk/amicus/article/view/5712 [access 5.06.2025].

Benesch, S. Dangerous Speech: A Proposal to Prevent Group Violence, World Policy Institute, (2012).

Di Stefano, S. Translating and Developing International Human Rights Law in the Online Sphere: The Role of Meta's Oversight Board, forthcoming in: I. Couzigou, E. Fromageau (eds.), *International Law and Technological Change: Testing the Adaptability of International Law*, Edward Elgar, (2025), DOI: 10.2139/ssrn.4920875, [access 4.06.2025].

Douek, E. *The Meta Oversight Board and the Empty Promise of Legitimacy*, Harvard Journal of Law & Technology, Vol. 37, No. 2, (2024), pp. 373 – 445, DOI: 10.2139/ssrn.4565180.

Endres, D., Hedler, L., and Wodajo, K. Bias in Social Media Content Management: What Do Human Rights Have to Do with It?, AJIL Unbound, 117, (2023), pp. 139 – 144, DOI: 10.1017/aju.2023.23, [access 4.06.2025].

Farrand, B. ‘Is This a Hate Speech?’ The Difficulty in Combating Radicalisation in Coded Communications on Social Media Platforms, European

[74] Helfer, Land (2023), pp. 2233 – 2298.

Journal on Criminal Policy and Research, 29(3), (2023), pp. 477 – 495, DOI: 10.1007/s10610-023-09543-z, [access 4.06.2025].

Helfer, L. R., Land, M. K. *The Meta Oversight Board's Human Rights Future*, *Cardozo Law Review*, 44(6), (2023), DOI: 10.2139/ssrn.4197107, Available at SSRN: https://ssrn.com/ abstract=4197107, [access 4.06.2025].

Kaye, D. Report of the Special Rapporteur on the Right to Freedom of Opinion and Expression, UN Doc A/72/350, (2017), https://digitallibrary.un.org/record/1304394, [access 4.06.2025].

Klonick, K. *The Facebook Oversight Board: Creating a Supreme Court for Content Moderation* (2020), *Yale Law Journal*, 129(6), p. 1598 – 1671.

Ogus, A. *Regulation: Legal Form and Economic Theory*, Hart Publishing, (1994), pp. 110 – 111, [access 4.06.2025].

Ogus, A. Rethinking Self-Regulation, *Oxford Journal of Legal Studies*, 15(1), (1995), pp. 97 – 108, DOI: 10.1093/ojls/15.1.97, [access 4.06.2025].

Pérez, J. M. Assessing the Impact of Contextual Information in Hate Speech Detection, *IEEE*, 11, (2023), pp. 30575 – 30590, DOI: 10.1109/ACCESS.2023.3258973, [access 4.06.2025].

Petkova, B. Meta Oversight Board's Nascent Standard on Hate Speech: Towards Plural Standard Setting in International Human Rights Law, forthcoming, *La Rivista di diritto dei media (Media Laws)*, (2026), on file with the author.

Pollicino, O., Paolucci, F. Digital Constitutionalism, forthcoming in: L. Floridi, M. Ziosi, M. Taddeo (eds.), *Companion to Digital Ethics*, Oxford University Press, (2025), DOI: 10.2139/ssrn.5098492, [access 4.06.2025].

Schoenebeck, S., Blackwell, L. Reimagining Social Media Governance, *Yale Journal of Law & Technology*, 23, (2021), p. 113.

Suzor, N., Van Geelen, T., Myers West, S. Evaluating the Legitimacy of Platform Governance: A Review of Research and a Shared Research Agenda, *International Communication Gazette*, 80(4), (2018), p. 385 – 400, DOI: 10.1177/1748048518757142, [access 6.06.2025].

Wischmayer, T. What is Illegal Offline is Also Illegal Online. – In: B. Petkova, T. Ojanen (eds.), *Fundamental Rights Protection Online*, Elgar, (2020).

Wong, D., and Floridi, L. Meta's Oversight Board: A Review and Critical Assessment, *Minds & Machines*, 33, (2023), pp. 261 – 284, DOI: 10.1007/s11023-022-09613-x

York, J. C. Meta Oversight Board's Latest Policy Opinion, EFF Deeplinks Blog, (2024), https://www.eff.org/deeplinks/2024/03/meta-oversight-boards-latest-policy-opinion-step-right-direction, [access 20.05.2025].

Zankova, B., Dimitrov, V. Social Media Regulation: Models and Proposals, *Journalism and Mass Communication*, 10, 2, (2020), pp. 75 – 88.

Zuckerberg, M. A Blueprint for Content Governance and Enforcement, Facebook, (2021), [access 20.05.2025].