C.A.S.C.A

Comments to Meta Oversight Board

New Cases Involve Symbols Adopted by Dangerous Organizations

February, 2025

Comments to Meta Oversight Board

New Cases Involve Symbols Adopted by Dangerous Organizations

Authors: Jui Dharwadkar, Mahek Sangwan, Pranav Gupta, Sayed Kirdar Husain

Research Consultant: Dr. Ivneet Walia, Associate Professor of Law and Officiating Registrar, RGNUL

CENTRE FOR ADVANCED STUDIES IN CYBER LAW AND ARTIFICIAL INTELLIGENCE [CASCA] is a research-driven centre at RGNUL dedicated to advancing scholarly research and discourse in the field of Technology Law and Regulation. As a research centre of a leading institution in India, we are committed to promoting interdisciplinary research, fostering collaboration, and driving innovation in the fields of cyber law, artificial intelligence, and other allied areas.

For more information

Visit **cascargnul.com**

Disclaimer

OVERSIGHT BOARD'S CALL FOR PUBLIC COMMENTS

(Released February 13, 2025)

In three separate Instagram cases, users posted content promoting symbols and messages linked to extremist ideologies. The first case, from April 2016, featured a blonde woman with a scarf and the kolovrat symbol, a type of swastika associated with both neo-Nazis and some pagans. The caption expressed pride in Slavic heritage, associating the symbol with faith, war, peace, hate, and love. In October 2024, another user posted selfies wearing an iron cross necklace and a T-shirt with an AK-47 and the words "Defend Europe." The post, along with its use of neo-Nazi symbols and hashtags, promoted anti-immigrant sentiment. The third case, from February 2024, shared drawings of an Odal rune wrapped around a sword, accompanied by a quote from German author Ernst Jünger, omitting the rune's Nazi connections and presenting it as a symbol of "heritage, homeland, and family." While Meta removed the first two posts after review by subject matter experts in November 2024, the third post was deemed non-violative.

The Oversight Board selected these cases to examine Meta's approach to moderating symbols associated with dangerous organizations while balancing users' freedom of expression. Public comments were invited to provide insights on:

(i) How Meta should treat symbols with different meanings when reviewing at scale, where the review by the company's subject matter experts is limited.

(ii) The significance and prevalence of both the Odal/Othala rune and the kolovrat, particularly on social media.

(iii) To what degree pagan and runic symbols in general have been appropriated by white supremacists and neo-Nazis, and the extent to which they are still used in non-extremist settings.

(iv) Ways in which neo-Nazi and extremist content is disguised to bypass content moderation on social media.

The Comments contributed by CASCA strive to help Meta strengthen its commitment to curb dangerous organisations, respect user's freedom of expression and ensure that its platform is not used to promote violent and hateful content. The original Call for Public Comments can be accessed here.

COMMENTS TO OVERSIGHT BOARD

I. How Meta Should Treat Symbols With Different Meanings When Reviewing At Scale, Where The Review By The Company's Subject Matter Experts Is Limited

When we discuss the limited review of subject matter experts, we understand that while policies and human rights obligations are in place and being followed, there is a discrepancy between when symbols used in a particular context need to be interpreted in a way that causes harm to the larger public interest. We suggest the incorporation of an objective and contextual standard for evaluating symbols with different meanings, when reviewing at scale. International human rights instruments provide an objective criterion to gauge the degree of neutrality of symbols with different meanings. Article 4 of CERD imposes an obligation on the state parties to prevent the dissemination of ideas based on racial superiority or racial hatred.[1] Moreover, Article 19 and 20 of the ICCPR contain provisions regulating freedom of expression and advocacy of hatred in the international sphere. The objective standard is the evaluation of ideas with racist and hateful connotations. These can be operationalized by Meta with the help of the case laws discussed below. In the cases of *Jersild* v. *Denmark,*[2] and *Faurisson* v. *France,*[3] it was observed that the item in question must be viewed from an objective point of view to judge as for its purpose of propagation of racist views and ideas. In the former case, while assessing the racist statement made by a journalist, the Court observed that the item in question must be considered as a whole in the context in which the statement was made, taking into consideration factors like associated history, contextual analysis, etc. Moreover, Meta has recently brought the proactive detection rate of hate speech to 97.1% through its new Artificial Intelligence ("AI") Policy,[4] with the use of the multiple language detection model. However, the policy lacks a detection mechanism which takes into account the multiple latent meanings that can be attached to a particular symbol. A contextual analysis of the post, including evaluating the user's history, will be more effective in determining the likely intent and impact of the post, rather than merely evaluating the symbol and the post in isolation.

Meta's current approach to moderating symbols with multiple meanings during at-scale reviews relies heavily on automated systems and limited human oversight, which often fails to account for critical contextual nuances. While the company uses tools like image recognition and keyword detection to flag content, these systems struggle to distinguish between harmful uses of symbols (e.g., neo-Nazi propaganda) and benign or culturally significant ones (e.g., pagan religious imagery). For example, posts containing symbols like the kolovrat or Odal rune—which have both historical and extremist connotations—are frequently flagged or left online for

---

[1] Toby Mendel, 'Hate Speech: Can the International Rules be Reconciled?' (*UN OHCHR,* 13 October 2011) <https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/ICCPR/Santiago/TobyMendel.pdf> accessed 24 February 2025.

[2] Jersild v Denmark (1995) 19 EHRR 1.

[3] Faurisson v France (1996) 4 IHRR 444.

[4] Officer MSCT and Meta, 'Update on Our Progress on AI and Hate Speech Detection' (*Meta,* 22 December 2021) <https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/> accessed 24 February 2025.

years until subject-matter experts manually intervene, as seen in the cases referred to the Board. This reactive, binary approach leads to inconsistent outcomes: harmful content persists undetected, while legitimate expressions are unnecessarily removed. The over-reliance on explicit policy violations (e.g., direct links to banned groups) further exacerbates the problem, as it overlooks subtler forms of hate promotion, such as pairing symbols with extremist slogans like #DefendEurope or weapon imagery.

To address these gaps, Meta should adopt a multi-layered, context-aware moderation system. ***First***, AI models must be trained to analyse combinations of signals—such as symbols paired with hashtags, user history, or adjacent text—to better infer intent. For instance, an Odal rune alongside violent rhetoric or accounts linked to hate groups could trigger priority escalation for human review, while the same symbol in a museum's educational post might be whitelisted. ***Second***, Meta should collaborate with historians, cultural experts, and marginalized communities to build a dynamic, publicly accessible database documenting the dual uses of contested symbols, informing both automated systems and moderator guidelines. ***Third***, a tiered review process could streamline decision-making: high-risk content (e.g., symbols + hate slogans) would require urgent expert assessment, while ambiguous cases could be temporarily flagged with user-facing warnings, allowing creators to provide context before removal. Additionally, user reporting tools should be refined to capture contextual details (e.g., a "cultural/educational use" option), and transparency reports should clarify how symbol-related decisions align with policy goals.

Meeta's policies must balance proactive harm prevention with freedom of expression. A "presumptive removal + appeal" model for high-risk symbols in non-historical contexts, paired with clear avenues for contextual appeals, would reduce the circulation of hateful content while preserving legitimate discourse. Without integrating these steps, Meta risks perpetuating a cycle of over-censorship and under-enforcement, undermining trust in its platforms and failing to protect marginalized communities targeted by evolving extremist symbolism.

II. The Significance And Prevalence Of Both The Odal/Othala Rune And The Kolovrat, Particularly On Social Media

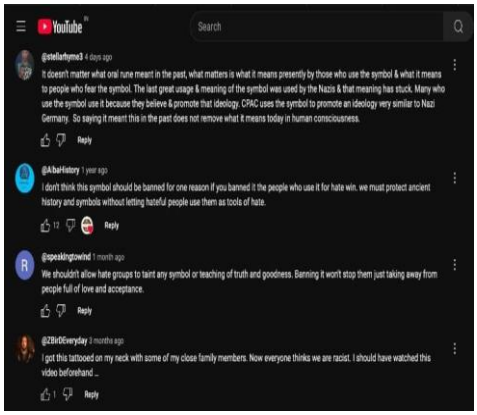
The *Odal/Othal Rune* emerged in the Medieval times, particularly in Norway and Sweden.[5] The *Rune* recognized the individual's right in the landed family property. Old Norse *odal* is generally understood as inherited landed property, family estate, and allodial property.[6] It was an emblem of the "Prinz Eugen" division of the Secret Service. Its recruits carried out massacres against the Slavs and civilians,[7] as they believed that the

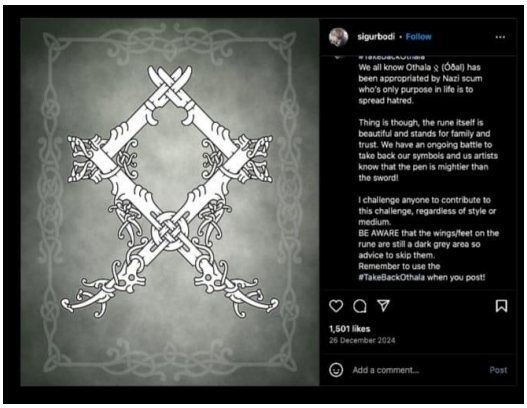
[5] Anders Andrén, 'Places, Monuments, and Objects: The Past in Ancient Scandinavia' (2013) 85(3) Scandinavian Studies 267, 274 < https://doi.org/10.5406/scanstud.85.3.0267> accessed 24 February 2025.

[6] Torun Zachrisson, 'The Background Of The Odal Rights: An Archaeological Discussion' (2017) 6(2) DJA <https://www.diva-portal.org/smash/get/diva2:1172167/FULLTEXT01.pdf> accessed 24 February 2025.

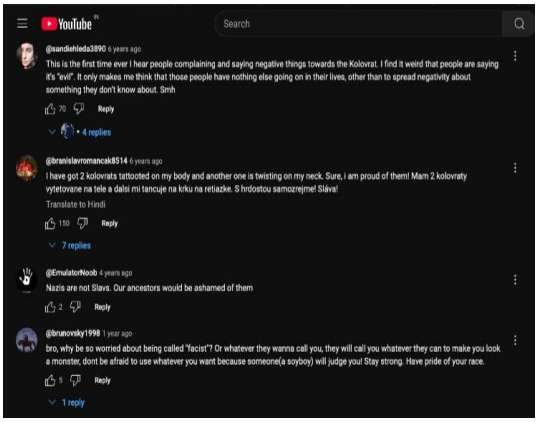
[7] Mirna Zakic, '"MY LIFE FOR PRINCE EUGENE": History and Nazi Ideology in Banat German Propaganda in World War II' [2020] pp. 79-94.

Slavs who refused to submit to German rule deserved no mercy. Neo-Nazi propaganda has increased with the use of social media.[8]





The *Kolovrat* is an ancient Slavic symbol. '*Kolo*' is Slavic for wheel and '*Vrat*' means turn or rotation. It emphasizes the circular nature of existence and the continuous motion of life and the universe. The *Kolovrat* represents cosmic balance and the eternal cycle that governs everything in the natural world.[9] The Slavs used the symbol on everyday objects ranging from jewellery, utensils and agricultural cycles for good luck and fortune. Cultural appropriation has turned the symbol into a tool for Slavic supremacy and far-right racism in Russia and Ukraine.[10] An analysis of the comment sections of certain YouTube and Instagram posts reveals that the general public believes in bringing back the usage of the symbols to reclaim their original meaning of truth and goodness.



III. To What Degree Pagan And Runic Symbols In General Have Been Appropriated By White Supremacists And Neo-Nazis, And The Extent To Which They Are Still Used In Non-Extremist Settings

[8] Panos Kompatsiaris and Yiannis Mylonas, 'The Rise of Nazism and the Web: Social Media as Platforms of Racist Discourses in the Context of the Greek Economic Crisis' in Christian Fuchs and Daniel Trottier (eds), *Social Media, Politics and the State: Protests, Revolutions, Riots, Crime, and Policing in the Age of Facebook, Twitter and YouTube* (Routledge 2015).

[9] 'The Kolovrat: An Ancient Symbol of the Sun and the Cycles of Life' (*Perun*, 8 October 2024) <https://www.perun.watch/blog/kolovrat> accessed 24 February 2025.

[10] 'Kolovrat' (*Reporting Radicalism*) <https://reportingradicalism.org/en/hate-symbols/movements/modern-racist-symbols/kolovrat> accessed 24 February 2025.

Online communities serve as platforms for creation, appropriation and circulation of far-right memes and messages.[11] A study conducted revealed that between 2012 and 2016, Twitter saw an increase of 600% followers on groups associated with white nationalist movements.[12] Online outlets like encrypted chat apps, social networking sites, and unmoderated message boards,[13] are used to provide online safe havens for extremists.[14] The use of recent technology has facilitated international cooperation between these extremist groups. A recent research by the Institute for Strategic Dialogue ("ISD") has revealed that far-right groups in Germany are using Generative Artificial Intelligence ("AI") to create images and narratives to disseminate their beliefs.[15] These include hateful content showing immigrants as criminals, idealizing Germany as a strong country under threat and creating fake influencers to create a parasocial connection with users and build a bigger following. In an investigation conducted into the killing of 11 members of the Black community,[16] it was found that the weapons recovered from the suspect had engravings of neo-Nazi iconography like the *Othala Rune*. Moreover, a study of transnational Identarian groups across Europe showed that 84% of the cooperation between such groups takes place through the social media platform, Twitter.[17] Hence, the larger extent of appropriation takes place through non-extremist settings like social media. While symbols and runes in Paganism have historically been developed to represent growth, life and divinity of the Norse religion,[18] contemporary usage suggests that these symbols are appropriated and co-opted to spread right wing extremism, racism and neo-Nazism. Pepe the Frog, a cartoon, was co-opted to be portrayed as a Nazi mouthing off anti-semitic and racist remarks on online spaces in the form of memes.[19] Another illustration is when the trademark

[11] Cynthia Miller-Idriss, 'What Makes a Symbol Far Right? Co-opted and Missed Meanings in Far-Right Iconography' (2019) Post-Digital Cultures of the Far Right <https://doi.org/10.1515/9783839446706-009> accessed 24 February 2025.

[12] JM Berger, 'Nazis vs ISIS On Twitter: A Comparative Study Of White Nationalist And ISIS Online Social Media Networks' (*Analysis and Policy Observatory*, 1September 2016) <https://apo.org.au/node/67247> accessed 24 February 2025.

[13] Annelies Pauwels, 'Contemporary Manifestations of Violent Right-Wing Extremism In The EU: An Overview Of P/CVE Practices' (*European Commission*, 2021) <https://home-affairs.ec.europa.eu/system/files/2021-04/ran_adhoc_cont_manif_vrwe_eu_overv_pcve_pract_2021_en.pdf> accessed 24 February 2025.

[14] Maura Conway, Ryan Scrivens and Logan Macnair, 'Right-Wing Extremists' Persistent Online Presence: History and Contemporary Trends' (*International Centre for Counter-Terrorism*, 25 November 2019) 3 <https://icct.nl/publication/right-wing-extremists-persistent-online-presence-history-and-contemporary-trends> accessed 24 February 2025.

[15] Anna Hiller & Pablo Maristany de las Casas, 'Generative AI and the German Far Right: Narratives, Tactics and Digital Strategies' (*Institute for Strategic Dialogue*, 18 February 2025) <https://www.isdglobal.org/isd-publications/generative-ai-and-the-german-far-right-narratives-tactics-and-digital-strategies/> accessed 24 February 2025.

[16] Eric Levenson, Sarah Jorgensen, Polo Sandoval and Samantha Beech, 'Mass Shooting At Buffalo Supermarket Was A Racist Hate Crime, Police Say' (*CNN*, 16 May 2022) <https://edition.cnn.com/2022/05/15/us/buffalo-supermarket-shooting-sunday/index.html> accessed 24 February 2025.

[17] Sting Daniëls and Yannick Veilleux-Lepage, 'The Dutch Identitair Verzet and the European Identitarian Movement' in Katherine Kondor and Mark Littler (eds), *The Routledge Handbook of Far-Right Extremism in Europe* (Routledge 2023).

[18] Gründer and René, 'Rune Secrets: On the Reception of Esoteric Runic Lore in German Neopaganism' (2009) Aries 9(2) 137 <10.1163/156798909X444806> accessed 24 February 2025; Birgit Sawyer, *The Viking-age Rune-Stones: Custom and Commemoration in Early Medieval Scandinavia* (OUP 2000).

[19] Jeffrey Demsky, 'That Is Really Meme: Nazi Pepe the Frog and the Subversion of Anglo-American Holocaust Remembrance' (2021) Nazi and Holocaust Representations in Anglo-American Popular Culture, 1945–2020, 105, 110 <https://link.springer.com/chapter/10.1007/978-3-030-79221-3_7> accessed 24 February 2025.

for a brand, Boy London, was declared invalid by the EUIPO over its alleged association with Nazi symbolism.[20]

IV. Ways In Which Neo-Nazi And Extremist Content Is Disguised To Bypass Content Moderation On Social Media

Two prominent tactics used by neo-Nazis and extremists to bypass content moderation are the use of symbolic imagery and coded language. The *Kolovrat* and the *Odal Rune* are often used interchangeably with the *Black Sun*, a symbol of esoteric Nazi beliefs. For example, a photo of the Conservative Political Action Conference stage, an event organized by far-right extremists in the US, was widely circulated on Twitter because its design resembled that of the *Rune*.[21] The image of the *Rune* was also seen on flags at a violent rally, organised by extremist groups in Virginia, was also circulated widely across social media.[22] In both these cases, while the images circulated on social media were not violative of moderation policies *per se,* they carried extremist messages that are meant to reach and incite a larger audience. Coded-language is yet another strategy to circumvent content moderation. An investigation revealed that there were as many as 120 pages on Facebook espousing white supremacist ideology and they have gained a total of 800,000 likes[23], with some pages being online for more than 10 years. One of those pages belonged to M8l8th, a black metal music act from Ukraine. The two 8s refer to the letter H, the eighth letter in the alphabet. Both 88 and the double H are common shorthand in neo-Nazi circles for Heil Hitler. While such content is generally violative of Meta's community guidelines and policies, the coded portrayal of the message prevents it from being detected by the algorithm. Moreover, there is another category of coded-language called *homoglyphs*, i.e., similarly shaped characters with different meanings, like the letter 'o' and the number '0'.[24] For instance, if the word *far-right* were to be banned, extremists would spell it as '*f4r-r!ght'* in order to avoid moderation mechanisms. Therefore, we deem it important for Meta to review and detect clandestine symbolic references and coded-language within the posts.

---

[20] Anglofranchise v EUIPO - Bugrey (BOY LONDON), Case T-439/21.

[21] Suzanne Rowan Kelleher, 'How A Nazi Symbol At CPAC Turned Into A Massive Hyatt Public Relations Disaster' (*Forbes*, 1 March 2021) <https://www.forbes.com/sites/suzannerowankelleher/2021/03/01/how-hosting-cpac-turned-into-a-massive-hyatt-public-relations-disaster/> accessed 26 February 2025.

[22] Makenzie Marsland, 'Why Pagans Need to Reclaim Runes from Nazis' (*CVLTNation*, 29 August 2017) <https://cvltnation.com/why-pagans-need-to-reclaim-runes-from-nazis/> accessed 26 February 2025.

[23] Ritzen, Y., & Unit, A. J. I. (2020, September 23). Exclusive: Facebook used extensively to spread neo-Nazi music. *Al Jazeera*. https://www.aljazeera.com/news/2020/7/10/exclusive-facebook-used-extensively-to-spread-neo-nazi-music

[24] Aurora Agnolon, 'AI Tools and the Alt-Right: A Double-Edged Sword for P/CVE' (*Global Network on Extremism & Technology*, 28 January 2025) <https://gnet-research.org/2025/01/28/ai-tools-and-the-alt-right-a-double-edged-sword-for-p-cve/> accessed 27 February 2025.