„Marg bar Khamenei“

Amelie Röhling 2023-02-04T10:13:00

A Protest Slogan reviewed by Meta and its Oversight Board

Despite the Iranian regime’s extensive efforts to control the online sphere, social media are still important platforms for dissent in Iran and beyond. The limits of criticizing the Iranian government – or, more specifically, a slogan popular with the current protest movement and its supporters – also played a role in a mid-January decision by Meta’s Oversight Board (OB), the appellate body for content moderation decisions, whose rulings Meta considers binding.

The case involved a Facebook post from July 2022 in a group dedicated to supporting freedom in Iran. It featured a cartoon of Iran’s Supreme Leader, Ayatollah Ali Khamenei, holding a chained, blindfolded woman with hijab in his fist-shaped beard. Below the image, a caption written in Farsi reads „marg bar“ the „anti-women Islamic government“ and „marg bar“ the „filthy leader Khamenei“. „Marg bar“ is literally translated as „death to“, but in the political context of Iran it is often used to mean „down with“.

After another user reported the post, Meta removed it for violating its community standards against violence and incitement. The user account also received a warning and two functional restrictions: The user was unable to create new content on Facebook for one week. For 30 days, he was prevented from posting or commenting in groups, inviting new members to groups, or creating new groups.

An appeal against Meta’s decision was precluded by an automated prioritization system. After the user appealed to the OB, it overturned the platform’s decision on the grounds that the slogan was worthy of protection and should be understood in the political context of Iran.

Integrating Fundamental Rights into Content Moderation

In making its decisions, the OB explicitly considers international human rights standards.

At EU level, the recently adopted Digital Services Act (DSA) even brings the fundamental rights dimension of content moderation into secondary law. Art. 14(4) DSA provides for an unprecedented obligation, which requires social media platforms to act in a proportionate manner in applying and enforcing their terms and conditions, with due regard to the rights and legitimate interests of all parties

involved, including the fundamental rights of their users as enshrined in the EU Charter of Fundamental Rights.

But how can this rather vague provision be operationalized? How can it be applied to a concrete case like the one outlined above?

Identifying the Relevant Rights and Interests

The first step is to determine which fundamental rights need to be considered. Art. 14(4) DSA explicitly addresses the rights of the recipients of the service such as freedom of expression and freedom of the media. This points to the fact that enforcing the rules of communication leads to a restriction of users' expressive possibilities protected by Art. 11(1) CFR.

Since service providers hold fundamental rights as private entities too, their interests must be considered as well. This is not contradicted by the fact that Art. 14(4) DSA appears at first glance to unilaterally oblige the platforms to the users' fundamental rights. The open wording ("with due regard") and the requirement of proportionality leave room for a balancing process that also takes the interests of the intermediaries into account.

As the rules of communication are part of the user contracts, their establishment and enforcement fall under the freedom to conduct a business protected by Art. 16 CFR, which includes freedom of contract. Furthermore, content moderation aims to keep the network attractive to users and advertisers, thereby serving economic interests of the providers, equally covered by Art. 16 CFR. Whether the service providers can rely on freedom of expression is less straightforward to answer (see generally Schiedermair/Weil, DÖV 2022, 305). The ECJ has not explicitly addressed this issue yet. However, content moderation falls under the scope of Art. 11(1) CFR since by moderating certain forms of content, providers are expressing in an evaluative way which contents and actions they disapprove of on their platforms. It also seems plausible that the service providers may invoke freedom of the media, as they provide a mass communication forum and control it in a substantial way by implementing rules and enforcing them.

The multi-layered constellation on social networks involving numerous stakeholders requires that third parties are taken into account (Denga, EuR 2021, 569 [593]), which is why Art. 14(4) DSA explicitly addresses the rights and legitimate interests of all parties involved. Indeed, content moderation not only serves the interests of the providers, but also those of third parties. The community standards on violence and incitement, cited by Meta in this case, serve to protect the right to life and physical integrity. In addition, the protection of Khamenei's reputation, which the ECHR also grants to non-European heads of state, might be considered here. Similarly, third parties can be negatively affected. Other recipients are denied (potential) access to the user's information by removing the post and limiting the user's ability to express themselves (cf. Google Spain, para. 68 f. & 81).

Special Obligation for Service Providers

How are the rights involved to be balanced in the context of Art. 14(4) DSA? Although the provision conceptualizes a unilateral obligation of the provider, it is clear that it does not envisage a strict proportionality test as applied to state measures, since the providers themselves have fundamental rights. Still, Art. 14(4) clearly places the emphasis on the platforms‘ obligation to fundamental rights, which may be understood as a regulatory response to the asymmetrical relationship between providers and their users. This places a stronger obligation on the platforms to their recipients than is usually the case in the relation of private actors, who are, in principle, equally situated as non-state entities. However, to recognize the platforms' interests and set them apart from the state, they must be granted a wider margin of discretion for classifying statements under their communication rules and the proportionality of the chosen measure (Denga, EuR 2021, 569 [594]). In line with the tiered approach of the DSA, this margin should vary depending on the platform's reach, market power, and importance to the public debate (cf. Recital 47, 57 & 75 f. DSA).

Reviewing the Statement in light of European Case Law

Apart from that, the usual principles for the legal relationship between private parties apply. As required by the ECJ, a fair balance must be struck between the conflicting fundamental rights of all parties involved. For this purpose, it is useful to take a look at the speech-specific criteria developed by the ECJ and, especially, the ECHR to assess the statement's worthiness of protection.

A key aspect is the democratic significance of the expression: While there is little room for restrictions on issues of public interest, artistic, or scientific expressions, purely commercial or entertaining statements are less worthy of protection. The post, including the cartoon, clearly falls under the special protection of political speech as it is intended to criticize the Iranian regime and the Supreme Leader, in particular the oppressive policies towards women and the mandatory hijab law.

Call for Violence or Legitimate Protest?

Meta's classification of the post as call to violence refers to a special class of cases. When reviewing statements that allegedly stir up or justify violence, the ECHR considers various factors. If the statement is made against a tense political or social background, this may suggest that some kind of restriction is permissible. The post was published in July 2022 – so before the mass protests and related clashes between demonstrators and Iranian state forces started in September. But that should not matter here anyway: It would be contrary to the nature of freedom of expression if speech against the government could be suppressed by referring to a tense situation largely caused by the state itself. After all, it is Iranian security forces that have used excessive and lethal force against even peaceful protests.

Probably the most important factor is whether the statement, fairly interpreted and considering its context, can truly be understood as a call to violence. As Meta's Oversight Board pointed out, the disputed slogan „marg bar … Khamenei" literally means „Death to Khamenei", but „marg bar" has been used in the context of protests in Iran mostly in the sense of „down with". The latter meaning seems more plausible here too, since the slogan is equally directed at the Iranian government as an institution, and because of the political nature of the post as a whole. So, the Oversight Board is correct in stating that this is "a rhetorical, political slogan, not a credible threat".

Finally, the nature and form of the expression as well as possible consequential damages must be considered. In addition to the threatening potential of an expression, the prospect and extent of dissemination may also be relevant. With regard to online communication, the ECHR emphasized that it poses a potential danger because content can be disseminated worldwide in a matter of seconds and sometimes remains permanently available online. On the other hand, as the Court stated, the ability of users to create their own online content represents an unprecedented form for the exercise of freedom of expression. Both Meta and the Oversight Board concluded that the posting was not likely to cause actual harm. Rather, in the absence of real alternatives beyond the control of the Iranian regime, the importance of social media as a platform for Iranian dissidents and their supporters, both in Iran and in Europe, to freely exchange views must be considered.

In light of European case law, balancing under Art. 14(4) DSA leads to the conclusion that Meta did not sufficiently take into account the author's freedom of expression when applying its violence and incitement policy. What is at issue here is political speech, outweighing the predominantly commercial interests of the provider and limiting its discretion, which is already narrowed by Facebook's market-leading position. As for Khamenei, the protection of his reputation can still be invoked. However, the limits of permissible criticism for politicians in their official capacity, in which the Supreme Leader is addressed by the post, are much broader and – in this case – outbalanced by the public interest of the topic.

Proportionality of Moderation Tools

The next question is whether the concrete measures taken by the platform are proportionate within the meaning of Art. 14(4) DSA. Central to this are the nature and severity of the moderation act. In addition, the statement's worthiness of protection is decisive, which is why in the present case, due to its political significance, higher requirements must be met.

Both measures taken were serious. Removing a post is the most severe content-related action as the statement is completely suppressed. And even though the account's functional restrictions were only temporary, they were particularly severe because they extend to the Iranian government's proclaimed "Hijab and Chastity Day", which is often used by critics of compulsory hijab to protest. The author was denied this opportunity in that he could not express himself via Facebook. As the OB rightly points out, account restrictions „can shut people out of social movements and

political discourse in critical moments, potentially undermining calls for action gaining momentum through Meta's products."

Such grave measures should be reserved for content that infringes the rights of others significantly or constitutes a serious violation of the terms of service.

The Issue of Automation

A special constellation of problems arises when AI systems are used for automatically detecting and classifying content: AI systems have shown deficits in accurately understanding the context of an expression, which is essential to correctly interpreting speech, as evidenced by the phrase "marg bar".

These inadequacies in classification may lead to *overblocking* of content. Even if the error rate is low, given the large scale of social media content, this can result in huge impairments of the fundamental rights of users and content providers. On the other hand, the advantages for the service providers and potentially affected third parties are obvious. The automated systems allow a comprehensive review of online content that is significantly faster and less expensive than purely human moderation (see Finck, Artificial Intelligence and Online Hate Speech, 2019, 6 f.).

In the present case, the rejection of the complaint submitted to Meta by the author was automated. As described by the OB, a complaint was automatically closed by Meta if a certain threshold was not reached. This was influenced by, among other things, the type and virality of the content, the severity of the violation, and the amount of time that had passed since the content was published. Such a prioritization system seems inherently inappropriate, if it filters out highly political posts, which is why OB Meta recommends revising the indicators. In the context of the DSA, such a practice is not even permissible. Art. 20(4) DSA requires a diligent substantive examination of each complaint. And Art. 20(6) DSA stipulates that the decision on the complaint against a moderation act must not be made solely by automated means.

As long as the error rate seems tolerable, Art. 20(6) DSA provides an appropriate solution to the conflicting interests concerning automated moderation, which remains possible at least in the first decision. If the error rate is too high, very large online platforms, as defined in Art. 33 DSA, are required under Art. 34(1) DSA to assess systemic risks posed by their algorithmic systems at least yearly. This includes negative effects on freedom of expression, originating from a moderation system with a significant error rate.

Furthermore, Art. 35(1) DSA requires platforms to take measures mitigating the identified risk, including adjustments to the moderation process (lit. c) and algorithmic systems (lit. d). In the case of automated moderation, this not only includes reducing the error rate, but also implementing other safeguards. For example, a human decision maker could be involved at the time of the initial decision to review the automated result rather than leaving it entirely up to the AI system.

However, since these obligations usually apply only annually and the threshold for classifying a platform as „very large“ is – considering the average of 45 million active users per month (Art. 33(1) DSA) – set very high, protection gaps will likely arise in this regard. With regard to upload filters for copyright infringing content, the ECJ has recently emphasized that these are only compatible with Art. 11(1) CFR if sufficient safeguards ensure that the systems can adequately distinguish between permissible and impermissible content (see generally Reda, VerfBlog, 2022/5/02). If the system is not sufficiently accurate and Art. 34 f. DSA do not apply for the aforementioned reasons, from a fundamental rights perspective, a corrective may be called for, which can be based on Art. 14(4) DSA and could consist of, for example, involvement of a human decision-maker in the initial decision-making stage.

Conclusion

With the help of European case law on freedom of expression, some substance can be attributed to Art. 14(4) DSA. Within the territorial scope of the DSA (Art. 2(1) DSA), the democratic-functional conception of the European approach to freedom of expression guarantees robust protection for political speech, which resonates with the current Iranian protest movement and its supporters. Against this background, the decision of the OB seems very convincing in its reasoning and result, which are based on a profound human rights approach. But in terms of automation, the OB only sticks to recommendations, while the DSA provides some actual procedural safeguards.

Nevertheless, due to the European courts' multifaceted approach, a certain degree of uncertainty remains with respect to Art. 14(4) DSA. This is ultimately unavoidable, as the area of freedom of expression is best decided on a case-by-case basis, making concrete regulation difficult.

This article contains extracts from Röhling/Weil, Die Grenzen privater Normsetzung durch soziale Netzwerke, in: Schrör/Keiner/Müller/Schumacher (eds.), Entscheidungsträger im Internet, 2022, 151, https://doi.org/10.5771/9783748934981-151, licensed under CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/legalcode).

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig“, project identification number: ScaDS.AI.