Integrity Institute

Response to Meta's Potential State-Backed Influence Campaign Cases Under Oversight Board Review

The Integrity Institute

We are a member-powered think tank and global nonprofit organization fueled by more than 600 Trust and Safety professionals across 23 countries and 80+ major platforms, all with direct experience preventing digital harms. This feedback was drafted via input from Integrity Institute members with expertise on state-sponsored content influence, coordinated inauthentic behavior, anti-terrorism efforts on social platforms. It does not reflect the views of all Integrity Institute members.

Summary

State-backed influence operations should be understood as coordinated efforts to shape the information environment rather than as collections of isolated pieces of content. While individual posts may appear benign when viewed independently, coordinated campaigns rely on networks of accounts, amplification strategies, timing, and narrative reinforcement to manufacture consensus, suppress dissent, intimidate vulnerable communities, and influence public perception.

These cases highlight the importance of evaluating not only whether individual pieces of content violate platform policies, but also whether they form part of broader coordinated behavior capable of producing real-world harm. Effective responses therefore require behavioral analysis, network-level detection, proportional interventions, and heightened consideration of offline human rights impacts during periods of political unrest and armed conflict.

How should social media platforms detect and address online influence operations, particularly those carried out on behalf of states?

State-backed influence operations are fundamentally behavioral problems rather than purely content moderation problems. Individual posts rarely provide sufficient evidence that an

influence campaign exists. Instead, platforms should prioritize identifying coordinated patterns across networks of accounts, shared infrastructure, synchronized posting behavior, common amplification pathways, and repeated interactions that collectively indicate coordinated inauthentic behavior.

Graph-based detection should therefore complement traditional content moderation. Relationships between accounts, shared technical infrastructure, repeated engagement patterns, coordinated hashtag usage, and synchronized activity often provide stronger evidence of influence operations than any single post viewed in isolation. These systems should operate proactively, continuously identifying emerging networks rather than relying solely on user reports after campaigns have already gained traction.

State affiliation alone should not automatically trigger content removal. Governments routinely communicate through social media, and platforms should distinguish between legitimate government communication and coordinated influence operations. Where state affiliation is confirmed, transparency measures—including clear disclosure of state affiliation—allow users to better contextualize information without unnecessarily restricting access to lawful speech.

Platforms should also adopt graduated interventions rather than relying solely on the binary choice of leaving content online or removing it entirely. Transparency labels, reduced algorithmic amplification, temporary distribution limits while investigations are ongoing, and removal where coordinated inauthentic behavior or incitement is confirmed provide a more proportionate enforcement framework that balances user access to information with mitigating manipulation.

How do state-sponsored online influence operations typically operate, especially with respect to Meta's platforms?

State-sponsored influence operations typically rely on coordinated ecosystems rather than isolated accounts. These campaigns combine authentic state communications, state media outlets, coordinated networks of inauthentic or compromised accounts, automated amplification, and sympathetic users to reinforce narratives while obscuring the appearance of centralized coordination.

Rather than persuading audiences through individual posts, these campaigns seek to shape the broader information environment. Common tactics include flooding platforms with coordinated messaging, amplifying preferred narratives through coordinated engagement, hijacking or overwhelming hashtags used by opposition movements, manufacturing uncertainty through conflicting information, and introducing false or misleading narratives that erode trust in independent sources.

On Meta's platforms specifically, the greatest leverage often lies not in the content itself but in recommendation systems, account authenticity, and network effects. Recommendation

algorithms can unintentionally amplify coordinated campaigns when engagement signals are artificially inflated, making behavioral detection particularly important. Platforms should therefore prioritize identifying coordinated amplification networks, monitoring suspicious engagement patterns, and evaluating relationships between accounts rather than focusing enforcement solely on individual posts.

Campaigns also evolve over time. Networks frequently replace removed accounts with new ones, repurpose older "sleeper" accounts that have established credibility, or coordinate across multiple platforms simultaneously. Effective disruption therefore requires continuous monitoring of network behavior rather than one-time enforcement actions against individual accounts.

What is the impact of pro-Iranian government propaganda and online influence operations on Iranian civilians, particularly in the context of the recent protests?

The harms associated with state-backed influence campaigns extend well beyond the visibility of individual posts. During periods of widespread protest and political repression, coordinated influence operations can contribute to intimidation, suppress civic participation, undermine trust between protesters, and reduce individuals' willingness to share information or organize collectively.

These harms are particularly significant in environments where independent journalism is restricted, internet access is disrupted, and reliable information is already difficult to obtain. Under these conditions, coordinated propaganda can shape public understanding not only by spreading favorable narratives but also by manufacturing uncertainty, creating confusion about unfolding events, and making citizens less confident in the information they receive from one another.

Platforms should therefore evaluate these cases through both a Trust & Safety and human rights lens. Where there is credible evidence of ongoing political violence, widespread arrests, internet shutdowns, or other forms of state repression, decisions about state-linked influence operations should account for their potential offline consequences. The question is not solely whether individual posts violate existing policies, but whether coordinated campaigns contribute to the suppression of fundamental rights, including freedom of expression, freedom of association, and access to reliable information.

Platforms should consider implementing conflict-sensitive review protocols that incorporate contextual expertise, document the rationale for enforcement decisions, and explicitly evaluate the potential human rights implications of leaving coordinated state-backed influence campaigns online during periods of active crisis.

How does the Iranian government and its supporters use social media, and do they employ coded threats or incitement to violence?

The Iranian government's online information strategy has evolved into a sophisticated ecosystem involving state media, affiliated organizations, cyber units, coordinated supporters, and networks of online accounts that work collectively to reinforce government narratives while undermining opposition movements. These efforts frequently extend across multiple social media platforms, with Instagram playing a particularly significant role given its popularity inside Iran.

Common tactics include portraying protesters as criminals, foreign agents, or terrorists; minimizing reports of state violence; promoting narratives favorable to government security forces; amplifying pro-government messaging through coordinated engagement; and deploying misinformation intended to create confusion or distrust among opposition groups. During periods of unrest, these campaigns often coincide with broader efforts to restrict access to independent information through censorship, internet shutdowns, and surveillance.

The use of coded language presents an additional challenge for enforcement. Messages invoking sacrifice, martyrdom, national defense, or historical and religious symbolism may carry substantially different meanings when viewed within the political and cultural context in which they are shared. Content that appears ambiguous in isolation may function as implicit threats or signals of support for violence when posted alongside ongoing government crackdowns or coordinated intimidation campaigns.

For this reason, evaluating these cases requires significant local expertise and contextual analysis. Platforms should avoid relying solely on literal interpretations of text and instead consider timing, surrounding events, intended audiences, historical references, and patterns of coordinated behavior when assessing whether content constitutes implicit incitement, intimidation, or support for state violence.

Conclusion

These cases demonstrate the limitations of evaluating state-backed influence operations through isolated content moderation decisions alone. Influence campaigns operate through coordinated networks, behavioral patterns, and cumulative effects that extend beyond the content of any individual post. Accordingly, enforcement strategies should increasingly prioritize behavioral detection, network analysis, and proportionate interventions alongside traditional content moderation.

The Integrity Institute encourages Meta to continue investing in graph-based detection capabilities, transparency around state-affiliated activity, and conflict-sensitive review

processes that incorporate local expertise and human rights considerations. Particular attention should be paid to how recommendation systems, amplification mechanisms, and account authenticity contribute to the reach and effectiveness of coordinated influence campaigns.

Finally, these cases underscore the importance of viewing platform governance as more than a question of removing individual pieces of content. Protecting the integrity of online information environments requires addressing the systems that enable coordinated manipulation while preserving access to legitimate political speech. By strengthening behavioral detection, improving transparency, and adopting proportionate interventions, platforms can better safeguard both public discourse and the rights of individuals during periods of political instability and conflict.