MAY 2026

Tapping into America's Distaste for Forever Wars

The Spread of Iranian Narratives on Bluesky

By Jose M. Macias III and Nico Vacca

Introduction

The United States and Israel have made battlefield gains in their conflict against Iran, but the United States is struggling to counter Iranian propaganda. Operational successes have **removed** Iran's authoritarian supreme leader, dismantled its **defense leadership apparatus**, and **degraded** its **missile capabilities**. However, the opportunity cost of military success for the United States is the loss of ground in the information war for the hearts and minds of American audiences and Western audiences more broadly. While Iran is losing on the battlefield, it is competing effectively in the information space through an aggressive, multiplatform disinformation campaign.

Analysis by the Futures Lab of more than 9,000 Bluesky Social posts finds that messages seemingly designed to exacerbate public divisions, which compose 23 percent of posts in the dataset, are the highest performing, averaging 150 reposts, 470 likes, and 28 replies per post. These same posts have been viewed by an estimated 293,666 users and are statistically significantly associated with a higher sharing volume, with an estimated 41 percent increase compared to other posts.[1] In addition, network and association analysis identifies 19 core accounts spreading Iranian war narratives and finds them to be active in 15 communities, with 11 communities estimated to be relying heavily on a singular account and 3 communities interacting with more than a single account.[2]

Iran's disinformation playbook has successfully targeted Israel in prior conflicts with Hamas and Hezbollah. Following the October 7 attacks, Microsoft reported that Iran **increased** cyber operations

1. The estimated user view is calculated as the cumulative sum of reposts. Within the dataset, the 2,127 posts (23 percent) categorized under the third theme accounted for 293,666 views. The total dataset consists of 770,431 cumulative views.
2. This was calculated using association rules analysis. Of the 15 network communities, #13 is below the minimum sample threshold for statistically reportable rules.

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

against Israel. Moreover, Iranian influence operations rose from roughly six in 2021 to eleven in October 2023 alone, and Iranian actors repeatedly reused old footage while falsely presenting it as Israeli attacks. As a result of this, Meta reportedly **closed** Iran-linked accounts that pushed narratives on religious divisions within Israel, liberal and conservative critiques of Israel's war with Hamas, and criticism of Prime Minister Benjamin Netanyahu in likely efforts to inflame tensions within Israel. Following **the 2025 U.S.-Israeli strikes on Iranian nuclear facilities**, narratives from Iranian accounts generated **over 126,000 engagements and an estimated 224 million views**. Now Iran's disinformation machine appears to be targeting Western audiences to undermine support for the U.S.-Israeli conflict with the Iranian regime.

Research from **Microsoft**, **Cyabra**, and ***The New York Times*** shows Iran's wider social media influence campaign aimed at the United States. To evaluate a portion of this campaign, the Futures Lab leveraged open-source data from Bluesky Social, BERTopic, and Claude's Sonnet 4.6 model to classify narratives on Bluesky feeds of suspected Iranian disinformation accounts. The research team identified three core themes that underpin Iran's disinformation campaign: (1) portraying the Iranian military and leadership as both victimized and successful; (2) framing the conflict as an Israeli war of choice; and (3) using rhetoric to amplify divisions surrounding the conflict within the United States and among its allies and partners. Applying statistical analysis, the research team identified that the third theme is positively associated with larger reach. The research team also used network analysis, which showed that the 19 core accounts are deeply embedded in 15 Bluesky communities. To counter these narratives, the U.S. government, working with social media platforms, should craft a visual counter-messaging campaign to deflect AI fakes; establish a labeling regime for state-sponsored posts; and use agentic artificial intelligence (AI) to systemically detect and dismantle propaganda networks.

A Refined Playbook

Iranian information operations follow a familiar asymmetric playbook: When direct competition with U.S. and Israeli military power is unfavorable, the objective **shifts** from battlefield parity to a sustained campaign to erode support for the conflict by **raising** the associated political costs. Russian and Chinese disinformation campaigns established the playbook that **Iran has now adapted**. Russian campaigns generally focus on **amplifying** divisions within society and **eroding** trust in Western democratic institutions. Chinese campaigns have generally aimed to shape **positive perceptions of the Chinese Communist Party** and suppress narratives on sensitive topics that may delegitimize the regime, such as Taiwan, the South China Sea, and Xinjiang, but both countries use overlapping tactics. **Recent monitoring by Graphika** found that the Iran conflict has produced a crowded online threat environment marked by information warfare, including deepfakes and other **AI-generated media**, disinformation, cyberattacks, and coordinated inauthentic activity across platforms. A **2024 Microsoft report** on Iranian cyber-enabled influence activity during the Israel-Hamas war showed how Iran used false personas and manipulated content to turn a regional conflict into a source of political pressure in Western countries. **Advancements in AI** are accelerating this process by increasing the scale and iteration cycles for disinformation campaigns. Iran is applying an established asymmetric information warfare doctrine, accelerated by AI, to a domestic American audience that is already divided over the conflict.

RESEARCH DESIGN AND METHODS

To analyze narratives aligned with Iranian strategic interests likely contributing to Iran's information war, the research team conducted a combined content and network analysis, identifying patterns of messaging, engagement, and amplification. The unit of study was defined as Bluesky users. The researchers narrowed the scope of the study to Bluesky, using the **R programming language**, because of its free, open-access application programming interface (API). This platform choice also narrowed the audience under observation. As of July 2025, **Bluesky web traffic** was concentrated most heavily in the United States (about 50 percent), followed by Japan (about 6 percent), the United Kingdom (about 5 percent), Germany (about 5 percent), and Brazil (about 4 percent). With 70 percent of the platform's user base consisting of Western audiences, the analysis should not be understood to cover all regions of the world.

Next, the research team reviewed existing research to identify patterns of likely Iranian disinformation accounts and to draw data from their Bluesky feeds. Specifically, **previous research** has shown that Iranian disinformation accounts focus on policy topics spanning anti-Israel, pro-Palestinian, and anti-Trump administration narratives. More **recent research** has identified accounts using AI deepfakes and **multilingual campaigns**, as well as **Iranian state media** pushing a narrative of **false battlefield success** through exaggerated and fabricated claims of strikes on U.S. military bases, U.S. aircraft, and Israeli cities. Based on these tactics, techniques, and procedures (TTPs), the researchers analyzed the Bluesky discover feed and selected accounts that showed overlapping TTPs. These TTPs included bot-like posting behavior, sharing links to Iranian websites flagged by the U.S. Department of Justice, persistent post or repost activity in short time periods, limited organic engagement, rapid shifts across opportunistic narratives, and other likely inauthentic tendencies consistent with prior reporting. While these TTPs do not establish definitive Iranian attribution, they identify a high-confidence sample of suspicious accounts consistent with previously reported Iranian-linked influence activity.

The researchers then queried the posts on each user's feed between February 28, 2026, and March 31, 2026. This included posts by the core users, reposts, and data on the original authors of those posts—including reposts, likes, and comment history—to construct a dataset for narrative, sentiment, and network analysis. To note, not all posts in the data were generated by the core accounts. An estimated 58 percent come from the 19 core accounts, while the remainder are reposts by the core users from accounts that appeared on their feeds.

To classify the narratives of the suspected Iranian disinformation accounts, the team combined **BERTopic** sentence transformers with Claude's Sonnet 4.6 large language model.[3] BERTopic was used to identify key topics in posts while Sonnet analyzed the narrative themes found within. The research team then analyzed BERTopic's and Sonnet's groupings to create 10 narrative clusters and consolidated them into three major themes, which are presented in this paper. Research also shows that negative or hateful content can **overperform** and go viral on social media. To include measures of negative and toxic content, the team leveraged **vaderSentiment** compound scores on a scale of -1 (Negative) to 1 (Positive) and **Detoxify** toxicity scores on a scale of 0 (Not Toxic) to 1 (Toxic) to quantify post

3. BERTopic is a topic modeling technique that uses Hugging Face transformers and class-based term frequency-inverse document frequency (c-TF-IDF) to convert text into dense vector embeddings. These models perform well at understanding the contextual meaning of words within sentences, capturing complex linguistic patterns.

sentiments and levels of toxicity that could spread the Iranian narratives further.[4] In addition, the team constructed directed interaction edge lists linking non-core interactors to suspected core accounts using post-level likes and repost metrics, and then analyzed and visualized the spread using NetworkX and R.[5] Tabulations across 9,012 posts are shown in Table 1. The dataset was further analyzed using a regression model, as discussed in the results section and shown in the appendix. The final dataset includes the key topics of posts and narrative categories from the 19 core user feeds that were used to establish the three core themes unpacked in this paper. The 19 core accounts and users sharing and engaging with their content remain anonymous in this paper.

This study does not claim definitive attribution of the selected Bluesky accounts to the Iranian state, the Islamic Revolutionary Guard Corps (IRGC), or any coordinated foreign influence operation. Instead, it analyzes accounts selected because their posting behavior, narrative themes, link-sharing patterns, and amplification behavior resemble tactics documented in prior research on Iranian-linked influence operations. The findings therefore should not be read as conclusive proof that these accounts were part of an Iranian state-backed influence operation.

Results

Theme 1: A Victimized but Successful Military and Leadership

The first theme that emerged centers on portraying Iranian offensive and defensive operations as evidence of military strength, capability, and strategic legitimacy. The spread of the first theme is depicted in Table 1. The Bluesky posts were on average reposted 45 times, liked 138 times, and commented on by at least seven users. In addition, an estimated 30 percent of these posts included images, and nearly 40 percent had a video embedded. However, further analysis shows that Theme 1 posts are associated with significantly lower repost volume. Regression results in Table 1A (see the appendix) show the performance of the three core narrative themes compared to all others in the dataset as a baseline; the coefficients show that Theme 1 posts are associated with significantly lower repost volume, with about 35 percent fewer reposts compared to the baseline.[6]

The data includes posts highlighting Iranian battlefield successes, including both real operations and fabricated or exaggerated military achievements. This is consistent with research reports spotlighting **AI deepfakes** depicting successful strikes on a U.S. military base and the U.S. embassy in Saudi Arabia. In addition, this theme is supported by Iran-nationalist and pro-regime posts, including claims that Iran

4. Multiple natural language processing tools exist to measure sentiment. For this study, the research team decided to identify a tool that performs well at quantifying negative sentiment due to the virality of negative social media posts. The first tool used was Valence Aware Dictionary and sEntiment Reasoner, or VADER. This is a lexicon and rule-based sentiment analysis tool that is attuned to sentiments expressed in social media, which makes it an appropriate selection for studying posts from Bluesky Social. For more details on the scoring, resources, and dataset descriptions, visit VADER's documentation here: https://vadersentiment.readthedocs.io/en/latest/pages/resource_description.html. The next tool used was Detoxify; this is a tool trained on comments and capable of detecting different types of toxicity such as threats, obscenity, insults, and identity-based hate. It provides a Toxicity Score in addition to other scores. For additional reading, visit the python library https://pypi.org/project/detoxify/, and from the creator, see Laura Hanu, "How well can we detoxify comments online?," Medium, November 13, 2020, https://medium.com/unitary/how-well-can-we-detoxify-comments-online-bfffe5f716d7.

5. The packages used in R were a combination of igraph and ggraph. The library Network X provides algorithms for random and classic networks and tools to analyze network structure and build network models. For additional reading, see the library's documentation and examples here: https://networkx.org/nx-guides/content/exploratory_notebooks/facebook_notebook.html#network-communities.

6. Because the dependent variable is transformed as log(1 + repost_count), coefficient magnitudes are interpreted as approximate percentage changes using 100 × (exp(β) – 1) where exp stands for exponential function/equation. For example, for Theme 1 (β = –0.42895), exp(–0.42895) – 1 = –0.3487, indicating about 35 percent fewer reposts relative to the baseline category, holding other covariates constant.

is on the morally correct side of the conflict. Prior reporting also documented **AI-generated content** depicting Ayatollah Ali Khamenei as a martyr in the face of Western aggression, as well as depicting global pro-Khamenei protests. Recurring content also justified Iran's deterrent posture by portraying military strength and nuclear capability as legitimate expressions of sovereignty and necessary means of self-protection against foreign aggression.

Table 1: Average Reposts, Likes, Replies, and Likelihood of Containing Image or Video by Theme

Theme	Percentage of Posts	Average Reposts	Average Likes	Average Replies	Contains Image	Contains Video
1	19%	45.2	138	7.81	30.7%	39.7%
2	26%	54.2	142	7.65	30.9%	30.2%
3	23%	150	470	28.4	28.8%	26.7%
Other	32%	94	281	15.6	37%	27.2%

Note: The full dataset contains 9,012 posts from the 19 core accounts' user feeds between February 28, 2026, and March 31, 2026. Not all posts were created by the core accounts; only 5,263 posts, or 58 percent of posts, share the core account and authorship handle. Average engagement numbers are skewed due to the virality of some posts.

This theme also incorporates narratives of Western weakness to reinforce the core claim of Iranian effectiveness. The data included posts depicting U.S. decline, Western failure, and the inability of U.S. and Israeli militaries to impose decisive costs, framing Iran as resilient and strategically effective. Three days before the initial U.S. strikes, Graphika detected **a coordinated social media campaign** that warned of the high cost of war while fabricating or exaggerating claims of Iranian capabilities. This same logic continued after the conflict began, relying on a **fabricated stream of content** focused on U.S. military casualties, declining morale, and the suffering of servicemembers and military families. For Western audiences, the narrative function appears to frame war with Iran as producing visible failure and pain for its adversaries, thereby reinforcing the perception of Iranian battlefield credibility.

Theme 2: Israel's War of Choice

The second theme centers on narratives portraying Israel as the principal aggressor in the conflict and the driving force behind U.S. involvement in the conflict. The spread of the second theme is depicted in Table 1. The Bluesky posts were on average reposted 54 times, liked 142 times, and commented on by at least seven users. In addition, an estimated 31 percent of the posts included images, and 30 percent had a video embedded. In Table 1A (see the appendix), the regression results show that Theme 2 posts are associated with higher repost volumes, with an estimated 20 percent more reposts compared a baseline of other themes.[7] Evidence from recent reporting, also captured in the dataset, points to three distinct but overlapping discursive patterns that define this theme.

7. For Theme 2 ($\beta = 0.18597$), $\exp(0.18597) - 1 = 0.204$, indicating that Theme 2 posts received approximately 20.4 percent more reposts than the baseline category, holding other covariates constant. "Other themes" refer to a baseline set of posts not used in the regression analysis as a categorical variable for themes one through three. This baseline is found in the log transformed constant/intercept of reposts and is omitted from the table for readability.

First, Israel is portrayed as responsible for drawing the United States and its allies into conflict with Iran. In one example from Clemson University that mirrors posts examined in this study's dataset, a Bluesky post questions why the United Kingdom is **eager to fight** in "Israel's war" while showing an AI image of Trump stuck in Iran with Netanyahu looking on from above. This reflects a real critique of the conflict but creates a narrative in which Western governments are acting in Israel's interests rather than their own. This messaging is likely designed to resonate with Western audiences already skeptical of a war with Iran or close alignment with Israel, who would be receptive to claims that their governments are sacrificing national autonomy.

The second discursive pattern is built on human suffering and atrocity framing possibly intended to intensify anti-Israel sentiment. This line of messaging may be designed to strengthen Iran's sovereignty by portraying Western governments as betraying their principles by aligning themselves with states willing to wage war with impunity. **One example** highlighted in external research appears to have amplified a real incident at Gandhi Hospital in Tehran. The posts framed the evacuation and damage from **nearby strikes** as a direct attack on a neonatal ward, contributing to widespread harm and fear among the Iranian public. In this study's dataset, this narrative was reflected with posts centered on the **Minab school strike**, which formed 6.8 percent of BERTopic's non-noise data. Even where the underlying events were real, posts used civilian suffering to support a broader moral narrative portraying Israel as genocidal.

The third pattern challenges Israel's legitimacy through an explicitly ideological anti-Zionist frame. Posts in this cluster repeatedly describe Israel as a terrorist, apartheid, or genocidal state and use the destruction in Gaza following **Hamas's October 2023 attack** to support that narrative. These posts use toxic and hateful stylized spellings to depict Israel, such as "Jizzrael," "zi0," "sionazi," and "IsraHell." The regression results in Table 1A (see the appendix) highlight that a one unit increase in positive sentiment score, as measured by VADER, is associated with an estimated 36.6 percent decrease in reposts.[8] This outcome means that posts that are more negative in sentiment perform better on the platform. There is an inverse outcome for toxicity, as a one unit increase in toxicity score is associated with an estimated 62.4 percent decrease in expected reposts.[9] The results show that negative sentiment is estimated to support the spread of a post narrative compared to toxicity. Taken together, this messaging frames Israel as an illegitimate and destabilizing actor whose influence over Western governments and role in civilian suffering justified broader hostility toward both Israel and its allies.

Theme 3: Exacerbating Public Divisions over the Conflict

The third theme captures narratives that are likely designed to make the conflict politically corrosive within the United States and other Western countries. This theme generated the strongest engagement on Bluesky. Posts in this category averaged 150 reposts, 470 likes, and 28 replies, with roughly 28 percent including images and 26 percent including video. The regression results from Table 1A (see the appendix) also highlight that posts in Theme 3 are associated with an estimated 41 percent more

8. VADER compound score is a normalized composite score spanning -1 to 1 where -1 is the bottom scale of negative sentiment and 1 the top scale of positive sentiment. Estimating the effect can be done as (β = -0.456), exp(-0.456) – 1 = -0.366 or -36.6 percent. Shifting a VADER compound score from 0 to 1 is associated with about 36.6 percent fewer reposts; in contrast, a shift in VADER compound score from 0 to -1 exp(+0.456) – 1 = +0.577 or 57.7 percent.

9. Detoxify scores are on a 0 to 1 scale and the regression estimates (β = -0.978), exp(-0.978) – 1 = -0.624 or 62.4 percent decrease as the score grows more toxic.

reposts compared to a baseline of other themes.[10] Rather than persuading audiences that Iran is in the right, this messaging likely sought to make continued support for the conflict feel costly, deceptive, and contrary to national interests using several discursive patterns. The first depicts the conflict as the product of manipulative and self-interested U.S. leadership disconnected from the public. This aligns with a narrative reflected in an open letter by Iran's president to the American public, which asks **whose interests the war actually serves**. It strikes at the core of recent polling, which finds that **61 percent** of Americans disapprove of the conflict.

One example of this theme contained in the study's data exploits the **partisan divide** surrounding the conflict by framing the conflict as a distraction measure. Specifically, similar posts in the data mirror reported posts that argue that the conflict serves as **a distraction** from the Epstein files and the people involved with the scandal. Additional messaging suggests that the president started the conflict to "**distract us from the real issues**." There is also a separate effort that personalizes Iran's disinformation campaign by targeting the U.S. president's family to inflame further partisan tensions; in one example, posts spread disinformation suggesting that Barron Trump **bought** oil futures shortly before the conflict. These narratives frame the conflict as the product of corrupt leadership serving private or political interests and not the American people.

A second narrative pushed within this theme emphasizes broader American institutional and elite dysfunction by targeting American institutions and corporations. Foreign influence campaigns use major U.S. companies to **deepen internal division and undermine confidence in the U.S. government itself**. This includes Iranian groups **targeting institutions** such as tech companies**, banks, and other high-visibility entities** with cyberattacks, as well as opportunistically claiming responsibility for server outages. The effect is to pair visible disruption at iconic U.S. companies and institutions **closely associated** with American power and prosperity with narrative amplification, reinforcing the perception that the conflict is producing disorder for the United States beyond the battlefield.

A third narrative translates the conflict into direct economic and human costs for Western audiences. Messaging focuses on disruption in the Strait of Hormuz, which the accounts use to frame the war as an **immediate burden** for the American public rather than a distant military operation. The internal dataset shows posts targeting pocketbook issues meant to make the conflict feel immediate to Americans. These include emphasizing rising gas and energy prices, food costs, inflation, and trade-offs between war spending and healthcare or Supplemental Nutrition Assistance Program (SNAP) funding. Through this view, the conflict becomes a source of household insecurity and further evidence that U.S. and Israeli leaders are acting recklessly and against their own publics' interests. On the human cost side, this pressure campaign has also involved **large-scale disinformation efforts** using AI-generated videos of U.S. flag-draped coffins, servicemembers suffering and regretting the war, and children pleading with parents not to fight.

Finally, the last framing extends beyond the United States, likely intending to widen divisions across Western allies. The internal dataset included a distinct BERTopic cluster centered on Australian

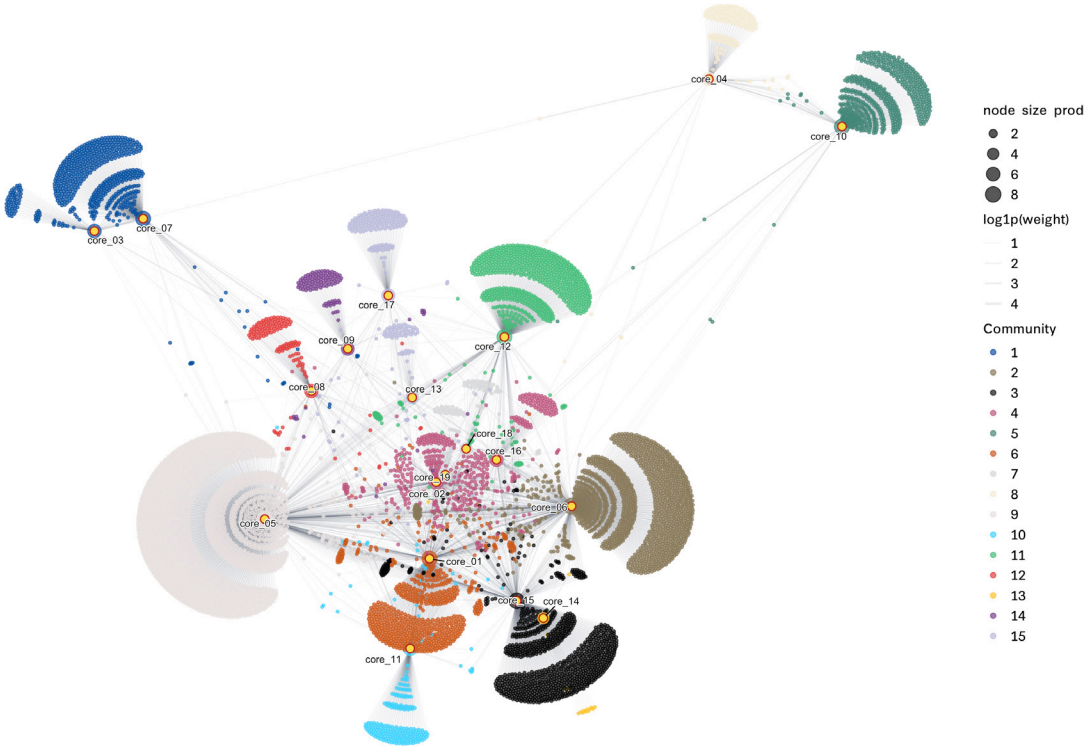
10. "Other themes" refers to a baseline set of posts not used in the regression analysis as a categorical variable for themes one through three. This baseline is found in the log transformed constant/intercept of reposts and is omitted from the table for readability.

political hashtags, suggesting the presence of an Iranian-linked disinformation campaign in Australian social media circles. Clemson research also identified a parallel network of IRGC-affiliated accounts posing as users from England, Scotland, and Ireland while pushing domestic political content tailored to those audiences, along with a Spanish-language cluster posing as users in the Americas, including accounts claiming to be in Texas, California, Venezuela, and Chile. The internal dataset complements this finding with accounts posting in Spanish and Portuguese. Taken together, these patterns suggest that the campaign sought not only to deepen divisions within the United States, but to weaken public support for the conflict across the broader Western coalition and turn the conflict into a source of friction for the West.

ACCOUNTS ARE EMBEDDED WITHIN 15 COMMUNITIES

The core 19 accounts and their Iran-aligned narratives mainly circulate within standalone groups of users (network communities) and rarely bridge over into other communities. Their embeddedness represents a challenge for social media platforms to identify them at scale, but agentic AI could flag accounts for review by humans based on their community structure. Figure 1 shows a repost network in which everyday users are linked to the 19 core user accounts based on observed repost interactions. The network structure indicates that the 19 core users are connected to a set of 15 distinct network communities on Bluesky with strong internal density and few ties to other users, with some core accounts occupying more central network positions and receiving higher repost exposure than others.

Figure 1: Core Accounts Embedded Within Communities



Note: min edge weight 1; layout fr(niter 3000, y stretch 1.22); labels 19.

Source: Futures Lab analysis of Bluesky Social.

In addition, the analysis finds that post amplifications are concentrated in a small set of core accounts. Table 2 captures repost exposure and shows that it is highly uneven across the network. Core accounts 5 and 6 received the largest repost exposure (weighted_in = 5,489 and 5,301), followed by a second tier including core account 15 (2,470), core account 10 (2,268), and core account 1 (1,934). This concentration indicates that narrative circulation is not evenly distributed across all 19 core accounts. Instead, a small number of accounts absorb a disproportionate share of repost activity and therefore carry more distributional weight in the observed network period. Breadth and structural position suggest different roles across core accounts. Comparing repost breadth (in_degree) with exposure (weighted_in) shows that some accounts are broad and deep amplifiers, while others are narrower but still represent high-volume nodes.[11] For example, core account 5 combines the highest exposure with the widest repost base (3,048 distinct reposting accounts), whereas core account 10 has high exposure with a narrower repost base (587). Page rank values reinforce this hierarchy, with top accounts occupying more central network positions.[12] The edge betweenness is near zero for most top nodes, meaning that leading accounts function less as bridges between communities and more as primary endpoints of amplification within their own community structures.[13]

Table 2: Top Five Anonymous Core Accounts by Repost Exposure

Anonymous ID	Repost Exposure (weighted_in)	Repost Breadth (in_degree)	PageRank	Bridge Score (betweenness)
Core Account # 5	5,489	3,048	0.0825	0.000000
Core Account # 6	5,301	1,820	0.0714	0.000000
Core Account # 15	2,470	1,326	0.0524	0.000156
Core Account # 10	2,268	587	0.0190	0.000000
Core Account # 1	1,934	941	0.0673	0.000198

Note: Repost exposure represents weighted incoming repost interactions; repost breadth represents the number of distinct reposting accounts.

Additional network association-rule analysis (see Table 1A in the appendix) shows that amplification is highly segmented by community. Several communities map almost one-to-one to specific core accounts, with near-perfect confidence. For example, the analysis shows with 100 percent confidence that accounts in community 9 are likely to engage with core account 5, accounts in community 3 with core account 15 (0.999), accounts in community 6 with core account 1 (0.999), and accounts in community 11 with core account 12 (0.999). This pattern indicates that much of repost behavior is not random across the network but concentrated within distinct community-account pairings. The lift values reinforce

11. In_degree refers to the number of inbound links; for additional reference, see Andrew Disney, "Social network analysis 101: centrality measures explained," Cambridge Intelligence, January 2, 2020, https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/. Weighted_in reflects the strength of the relationship between nodes; for additional reference, see David Hevey, "Network analysis: a brief overview and tutorial," *Health Psychology and Behavioral Medicine* 6, no. 1: 301–28, http://doi.org/10.1080/21642850.2018.1521283.

12. Page rank refers to a score calculated based on the direction and weight.

13. Edge betweenness is a measure of traffic flow-through, also understood to measure the proportion of shortest paths containing a given node and edge (direct or undirected links between nodes).

that these are strong associations above baseline prevalence. High-lift rules—such as community 14 to core account 9 (lift = 68.8), community 12 to core account 8 (59.6), and community 8 to core account 4 (57.1)—suggest unusually strong affinity between specific communities and target core accounts.[14] At the same time, some communities split attention across two accounts, such as community 7 (core account 18 at 0.522 confidence; core account 19 at 0.478) and community 4 (core account 2 at 0.771; core account 16 at 0.263), indicating mixed amplification pathways rather than a single dominant target. The association rule findings suggest that digital communities share one to two core accounts spreading Iranian narratives that amplify the content. Removing accounts that are spreading likely disinformation narratives therefore requires systematic deployment to find and target accounts for review by human moderators to disrupt actors that feed similar posts in digital communities.

Recommendations

As the conflict ebbs and flows between the United States, Iran, and Israel, the United States needs to start fighting back in the information space. This report proposes three recommendations aimed at countering Iranian narratives by punching back with visuals and videos to counter-message, creating transparency around posts, and using agentic AI to counter disinformation accounts at scale.

EXPAND COUNTER-MESSAGING WITH VISUALS AND VIDEO

To deny Iran the American psyche and push back against Iranian narratives, the White House should reach across the aisle to garner bipartisan support for a messaging campaign. In this manner, the administration could gain additional support and buy-in from local political and corporate elites that could join the administration in a truthful, **coordinated messaging campaign**. The statistical analysis conducted as part of this study shows that, within the analyzed dataset, posts with video content are estimated to have performed better on Bluesky. Any U.S. agency supporting the mission in Iran could use this to their advantage to (1) counter-message using videos or images and (2) counteract the fake AI videos using confirmed combat footage suitable for a wider audience. This will help communicate to the U.S. and wider Western public in a manner that will spread further than text posts and deflate the deepfakes. The Trump administration has experience sharing such visuals, in recent months sharing videos of U.S. strikes that targeted suspected **drug smuggling boats** and destroyed Iran's navy in under **10 days**.

BRING BACK POST LABELING FOR STATE-SPONSORED ACCOUNTS AND POSTS

In addition, to dissuade engagement and increase transparency around Iranian-backed accounts, the U.S. government should establish a partnership with social media platforms to request a proactive response by industry to label state-sponsored propaganda posts. **Research** has shown that users are less likely to engage with content on social media platforms when they are better informed about its source. The **labeling regime** could take different forms but should balance the need to counter-message with First Amendment rights protecting freedom of speech on online platforms. For example, this could be accomplished using a combination of verified account badges and digital signatures for videos. To combat the speed of dissemination, posts could be labeled with warnings that indicate a story may be rapidly evolving, unverified by major news sources, or under dispute.

14. Lift evaluates the strength of interaction relative to random chance.

REQUIRE INDUSTRY TO DEPLOY AI AGENTS TO TAKE DOWN DISINFORMATION ACCOUNTS

The age of AI is here, and the U.S. government should work with social media platforms to deploy agentic AI not only for counter-messaging but also to identify accounts, posts, and bot farms for review and shut them down. **Research highlights** that the U.S. government uses AI to defend against propaganda but **limits offensive actions** during peacetime and is therefore disadvantaged. With the conflict between the United States and Iran on pause, now is an opportune window for social media platforms to coordinate with the Department of War to detect and dismantle troll farms and disinformation accounts and free digital communities from Iranian propaganda.

Conclusion

Information operations shape beliefs, exploit social fractures, and leave behind distrust that outlasts conflicts themselves. This research presents a likely pro-Iranian campaign on Bluesky by users organized around narratives of Iranian military resilience, anti-Israel framing, and division across Western publics. This reflects a strategic calculation that losses on the battlefield can be offset in the information space and shape public opinion against policy implementation. The significance extends beyond this war. By amplifying tensions that already exist, Iran is attempting not only to weaken support for the current conflict, but also to erode the political resolve and alliance cohesion the United States and its allies will need long after the fighting stops. That is why the United States should fight back through counter-messaging, transparent post labeling, and AI deployed at scale. This is the first of many fights to come to protect the American public from foreign influence during kinetic conflict. Thus, what the United States does today will shape the future of information warfare for its allies and partners; it's imperative that the United States shields its public and seizes the narrative. ■

Jose M. Macias III *is an associate data fellow in the Futures Lab within the Defense and Security Department at the Center for Strategic and International Studies in Washington, D.C.* ***Nico Vacca*** *is a research intern with the Futures Lab.*

The authors wish to thank Kelsey Hartman for her outstanding review and analytical copyediting in this white paper.

This report is made possible by general support to CSIS. No direct sponsorship contributed to this report.

Appendix

Table 1A: Iranian Narrative Regression Results

Variables	Repost Effect
Theme 1: Victimized but Successful Military & Leadership	-0.429***
	(0.057)
Theme 2: It's Israel's War of Choice	0.186***
	(0.052)
Theme 3: Exacerbating Public Divisions Over the Conflict	0.341***
	(0.054)
VADER Compound Sentiment score (-1 to 1)	-0.456***
	(0.044)
Toxicity Score (0 to 1)	-0.978***
	(0.067)
Post with Image	0.558***
	(0.046)
Post with Video	0.968***
	(0.047)
Observations	7,042
R^2	0.123
Adjusted R^2	0.122
Residual Std. Error	1.613 (df = 7034)
F Statistic	140.578*** (df = 7; 7034)

Note: *p < 0.1; **p < 0.05; ***p < 0.01. The data is from Bluesky Social, and the constant is the log value of reposts (1.51792) and is omitted from the results. The complete dataset spans 9,012 user feed posts on Bluesky, including 5,263 original posts from the original 19 seed accounts. The VADER compound score is a normalized, weighted composite score used to generate unidimensional measure of sentiment on a scale of -1 (Negative) to 1 (Positive). The Detoxify Toxicity Score has a scale of 0 (Not Toxic) to 1 (Toxic).

A note on VADER sentiment compound scores and Detoxify toxicity scores

Controlling for narrative theme, image, and video, VADER sentiment was negatively associated with repost count. Because higher VADER compound scores indicate more positive sentiment, this suggests that more negative posts received more reposts. A one-unit increase in VADER compound score was associated with an estimated 36.6 percent decrease in repost_count + 1. Conversely, a move from neutral to strongly negative sentiment was associated with an estimated 57.7 percent increase in repost_count + 1. Detoxify

toxicity was also negatively associated with repost count: A 0.10 increase in toxicity was associated with approximately 9.3 percent fewer repost_count + 1, and a full-scale increase from 0 to 1 was associated with approximately 62.4 percent fewer repost_count + 1. Thus, while negative sentiment appears to be associated with greater reposting, toxic language appears to be associated with lower reposting.

Table 1B: Network Association Rules for Community Engagements with Core Accounts

Left-Hand Side	Right-Hand Side (Anonymous ID)	Support	Confidence	Coverage	Lift
Community 14	Core Account # 9	0.0128	0.992	0.0129	68.84
Community 12	Core Account # 8	0.0145	1	0.0145	59.61
Community 8	Core Account # 4	0.0166	0.9938	0.0167	57.14
Community 7	Core Account # 18	0.0061	0.5221	0.0116	54.55
Community 7	Core Account # 19	0.0056	0.4779	0.0116	53.99
Community 10	Core Account # 11	0.0248	0.9959	0.0249	34.81
Community 15	Core Account # 17	0.0199	0.7311	0.0272	34.15
Community 15	Core Account # 13	0.0078	0.2879	0.0272	31.43
Community 5	Core Account # 10 (T5)	0.0588	1	0.0587	16.55
Community 4	Core Account # 2	0.0378	0.7710	0.0490	15.87
Community 1	Core Account # 7	0.0591	0.8844	0.0668	14.66
Community 4	Core Account # 16	0.0129	0.2626	0.0490	13.57
Community 11	Core Account # 12	0.0833	0.9988	0.0834	11.36
Community 6	Core Account # 1 (T5)	0.0762	0.9987	0.0763	10.31
Community 3	Core Account # 15 (T5)	0.1141	0.9991	0.1142	7.32
Community 2	Core Account # 6 (T5)	0.1635	1	0.1635	5.34
Community 9	Core Account # 5 (T5)	0.2786	1	0.2786	3.19

Note: T5 refers to accounts in Table 2's top five accounts. Community 13 did not meet a threshold cutoff in confidence. The associated rules analysis required minimum support of 0.5 percent and confidence of 20 percent; with about 9,716 transactions, that means a rule needs at least 49 observations. Community 13 has only 16 interactors.