MARCH 2026

FIRST HALF 2026

Adversarial Threat Report

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Meta has publicly reported on adversarial threats since 2017, initially focusing on Coordinated Inauthentic Behavior (CIB) and subsequently expanding our scope to include espionage, surveillance-for-hire operations, and other malicious actors. We report to enhance the shared understanding of the threat landscape among industry peers, government bodies, and civil society, and we hold threat actors accountable by shining a light on their adversarial behavior.

We continue the expansion begun in our report last half by broadening our focus into additional high-risk harm adversarial threat areas. In addition to our existing reporting on CIB and scams, we introduce new narratives on our work to combat 'Nudify' content, cartels and drug trafficking organizations. We highlight how we take action to detect and remove these adversarial actors from our platforms, and empower the broader community to respond, even as adversarial actors increasingly operate across harm areas, across platforms, and across borders. The case studies contained in this report span our work combatting adversarial threats throughout 2025, with our CIB reporting limited to networks we removed in Q4 2025.

Fraud and Scams

Financially-motivated adversaries remain one of the most persistent and adaptive threats affecting the online ecosystem. These actors range from opportunistic individuals to sophisticated criminal scam syndicates (CSS), and they routinely evolve their tactics, techniques, and procedures to appear legitimate and evade detection measures.

Key Trends, and How We Respond:

- **AI is accelerating the sophistication and scale of these operations.** We are adapting our behavior and technical-based detections to combat the growing use of AI for persona development, multilingual content generation, and other attempts to bypass our detection systems.
- **Scam operations are becoming more industrialized.** We are increasing real-world risk for criminals and abusive businesses by taking action against scammer operations as early as possible in the Fraud Attack Chain, including collaborative efforts with law enforcement and through the legal system. The report details how some criminal scam syndicates operate with business-like specialization, training, and operational security—and how their operational footprints shift in response to real-world dynamics (e.g., conflict, border activity, and law enforcement pressure).
- **Targeting is increasingly tailored**, and designed to prey on individuals in high-need moments (housing insecurity, grief, loneliness).

Case Studies

To illustrate these trends, we include case studies of adversaries we disrupted who sought to adapt their narratives, personas, and infrastructure:

- **Criminal Scam Syndicates in Southeast Asia:** We disrupted syndicates running long-form scams (including romance and investment scams) use convincing lifestyle content to establish legitimacy, then shift targets into direct messaging and off-platform channels. We also outline how syndicate activity patterns can shift geographically over time, including movement tied to real-world events and enforcement pressure, and how these shifts affect the kinds of scam narratives emphasized (e.g., romance/investment vs. gambling-style lures).
- **Romance scams evolving personas:** We disrupted a network that shifted from military personas toward impersonating international doctors and medical professionals (including humanitarian-adjacent cues) to quickly build credibility and trust
- **AI-assisted "animal rehoming" fraud:** We took action against a network that used generative AI to boost credibility, and give the appearance of being a qualified local in scam animal rehoming advertisements.
- **Emerging scams exploiting vulnerable moments:** We removed a network that used coordinated fake rental listings targeting people seeking affordable housing, as well as "fake funeral livestream" scams that exploit grieving families through impersonation of bereavement-related services and payment-gated off-platform lures.

Countering 'Nudify' Apps and AI-enabled Intimate Image Abuse

This report expands our transparency on a growing threat: 'nudify' apps—services that use AI to generate non-consensual nude or sexually explicit imagery. Meta has longstanding rules against non-consensual intimate imagery, and we have further clarified and strengthened policies to prohibit the promotion of 'nudify' services and similar activity.

- **Multi-layer enforcement against promotion and access.** We remove content and advertising that promotes 'nudify' services, take action against accounts and Pages tied to promotional networks, block links to related websites, and restrict certain search terms to reduce discoverability.
- **Escalating consequences beyond routine enforcement.** We use legal action to increase real-world cost for entities behind 'nudify' services, including filing a recent lawsuit against Joy Timeline HK Limited, and issuing cease and desist letters to other entities violating our terms of service by seeking to advertise 'nudify' apps.
- **Ecosystem-level disruption beyond any single platform.** Because these services are marketed broadly across the internet (including beyond social media), we share signals -

starting with URLs through cross-industry mechanisms (e.g., the Tech Coalition's Lantern program) so other tech companies can investigate and take action. Lantern is a signal-sharing program that enables participating tech companies to share indicators about accounts, behaviors, and related signals that violate their child safety policies, so each company can investigate and take action on its own platform.

Cartels and Drug Trafficking Organizations

Cartels and drug trafficking organizations present pressing and complex risks, driven by real-world violence and persistent attempts to use online services for recruitment, coordination, extortion, trafficking, facilitation, and propagandizing.

We focus on how we evolve our tools and workflows to increase efficiency, reduce recidivism, and improve durability over time to detect and remove cartel activity from our platforms.

Core toplines on our evolving disruption toolkit:

- **We deploy newer discovery methods** that help subject-matter experts identify emerging cartel-linked and illegal narcotics activity even when adversarial actors use coded language and indirect signals.
- We use **improved image understanding and analysis workflows to surface relevant evidence** faster across large volumes of content.
- We integrate **insights from expert-led disruptions to inform automated detection and enforcement** pipelines, helping sustain pressure against reconstitution and repeat offenders.
- **We develop and deploy youth-focused messaging and educational interventions** designed to reduce vulnerability to recruitment and exploitation, developed with regional partners and refined based on measured learnings.

Cartels / Drug Trafficking Organizations Case Studies

The report includes case studies that illustrate both the threat and our evolving disruption toolkit:

- **Mexico-focused drug selling networks surfaced through improved discovery:** We describe how concept-driven discovery methods can surface clusters of content and accounts associated with attempted drug sales and promotion in Mexico—illustrating how better discovery can reduce reliance on reactive reporting.
- **Cartel-linked online activity tied to recruitment, extortion, and glorification:** We detail investigations into networks that displayed indicators consistent with organized criminal activity, including recruitment-style behaviors and content that promotes or glorifies

violence and criminal conduct—paired with enforcement approaches designed to remove the network, not just individual accounts or pieces of content.
- **Cross-border trafficking and adjacent high-severity harms:** We include an example of a drug trafficking network operating across the US and Mexico that used social surfaces to promote distribution.
- **Automation that sustains disruption pressure over time:** We describe how automated enforcement pipelines can be developed to counter recurring patterns—particularly where scammers exploit people seeking drugs (e.g., "drug-related scams" where buyers are defrauded), and where continuous enforcement can reduce reconstitution.

Coordinated Inauthentic Behavior (CIB)

CIB remains a foundational pillar of Meta's adversarial threat reporting. CIB involves coordinated networks of assets working together to mislead people using false identities. Our enforcement is rooted in deceptive behavior and coordination, not the content posted—regardless of the narratives promoted or whether an operation is foreign or domestic. We also continue to monitor for reconstitution and remove assets connected to previously disrupted networks. Our reporting in this section contains details of CIB networks that we removed in the fourth quarter of 2025.

CIB toplines this half include:

- **Two deep dives illustrating evolving tradecraft:**
    - We **investigated and removed an influence operation originating in Iran and targeting the US,** which used sophisticated fake personas on Instagram to misrepresent identity and amplify narratives. The case illustrates how influence operators increasingly blend amplification with attempted relationship-building.
    - We **disrupted a Pakistan-based domestic-targeting operation anchored in off-platform websites and coordinated amplification by pages and personas.** The case highlights the use of AI-assisted content generation across languages to increase output quality and consistency, and the use of identity misrepresentation (including stolen imagery) to appear locally authentic while promoting nationalist narratives and disparaging perceived adversaries.
- Beyond the deep dives, we **continued to investigate and remove other CIB networks, including operations originating from Russia, China, and Iran**, disrupting recurring patterns such as impersonation of grassroots media, localized branding to appear domestic, and cross-platform presence used to reinforce credibility.

INTRODUCTION

Meta has publicly reported on adversarial threats since 2017, initially focusing on Coordinated Inauthentic Behavior (CIB) and subsequently expanding our scope to include espionage, surveillance-for-hire operations, and other malicious actors. The objective of our reporting has consistently been twofold: to enhance the shared understanding of the threat landscape among industry peers, government bodies, and civil society, and to hold threat actors accountable by shining a light on their adversarial behavior. Over the years, our reporting has served as a valuable resource for industry partners, researchers, and security professionals seeking to understand and counter adversarial threats across the internet. We have consistently shared threat indicators, tactical analysis, and detailed case studies to support broader defender community efforts to detect, investigate, and disrupt influence operations.

At the same time, we recognize that the evolving threat landscape requires our public reporting to adapt. Threat actors continue to change their tactics, techniques, and procedures (TTPs), leverage multiple platforms and services, and exploit seams across the broader ecosystem. They increasingly use automation and artificial intelligence to improve the scale, speed, and credibility of their operations, including by generating more convincing personas, content, and infrastructure designed to evade detection. These dynamics reinforce a central lesson we have emphasized since our earliest reporting: countering adversarial threats requires coordinated, whole-of-society responses that extend beyond any single company's efforts.

In the previous edition of this report, we introduced a restructured format intended to support deeper analysis and more durable toplines, including new spotlights on fraud and scams, and AI-enabled threats. In this report, we continue that expansion by branching into additional high-risk adversarial threats where we see determined, adaptive adversaries attempting to exploit online services and target people in ways that can lead to serious real-world harm. This includes expanded transparency on our work against fraud and scams, as well as a discussion of our work against AI-enabled 'nudify' services and other forms of intimate image abuse, as well as a focus on our work to combat cartel and drug-associated gang activity. The case studies contained in this report span our work combatting adversarial threats throughout 2025, with our CIB reporting limited to networks we removed in Q4 2025. We are sharing these findings because we believe they can help industry, civil society, and governments better understand how these actors operate—and how their behavior evolves as defenders respond.

Because the threat landscape is dynamic, we expect the themes and specific focus areas of each report to continue to change. We select topics based on a combination of factors, including severity of harm, observed adversarial adaptation, and opportunities to share insights that can strengthen the broader ecosystem. With the exception of CIB- where we aim to report comprehensively on the networks we disrupt and to share supporting indicators- this report is not intended to document the full universe of our enforcement across every harm area we address. Instead, our goal is to provide a snapshot which highlights pressing adversarial harm areas and

shares the most meaningful observations we have made about how adversaries are adapting, scaling, and refining their TTPs.

We publish these snapshots to support practical action across the broader ecosystem. For other technology companies and defenders, we hope the findings help identify manifestations of similar behavior on other platforms and services. For researchers and civil society organizations, we aim to provide transparency into evolving threats and the tradecraft behind them. For governments and law enforcement, we hope these insights support investigations and other efforts to increase real-world costs for malicious actors. And for our users, we hope the examples and trends we describe help people recognize and avoid common scam patterns, deceptive approaches, and other adversarial behaviors designed to manipulate trust.

Across all sections, our core objective remains consistent: to increase awareness and strengthen collective defense against adversarial threats by sharing what we are seeing, what we are learning, and how threat actors are evolving. We hope this report- and our ongoing efforts to share relevant threat signals where appropriate- contributes meaningfully to whole-of-society efforts to better recognize, mitigate, and disrupt adversarial harm.

For a quantitative view into our enforcement of our Community Standards, including content-based actions we've taken at scale and our broader integrity work, please visit Meta's Transparency Center here: https://transparency.fb.com/data/.

01

FRAUD AND SCAMS

The threat posed by financially-motivated adversaries - from opportunistic scammers to sophisticated, professional criminal scam syndicates (CSS) - continues to represent one of the most persistent and evolving challenges facing the online ecosystem. As these actors adapt their tactics, techniques, and procedures (TTPs), defenders across the communities continue to watch as these operations react to our various responses, becoming more organized, more technologically capable, and more difficult to distinguish from legitimate activity.

In our previous Adversarial Threat Report, we introduced the Fraud Attack Chain - a framework for understanding the end-to-end lifecycle of financially-motivated scams. This model enables our teams to deconstruct complex scam operations into distinct phases, providing a clearer picture of how adversaries operate from initial infrastructure acquisition through to target exploitation and eventual fund extraction and beyond. We use this framework to organize our defense and mitigation strategy, identifying the most effective intervention points on our platforms - and where to partner effectively with other defenders across the broader ecosystem, upstream and downstream of where our visibility extends.

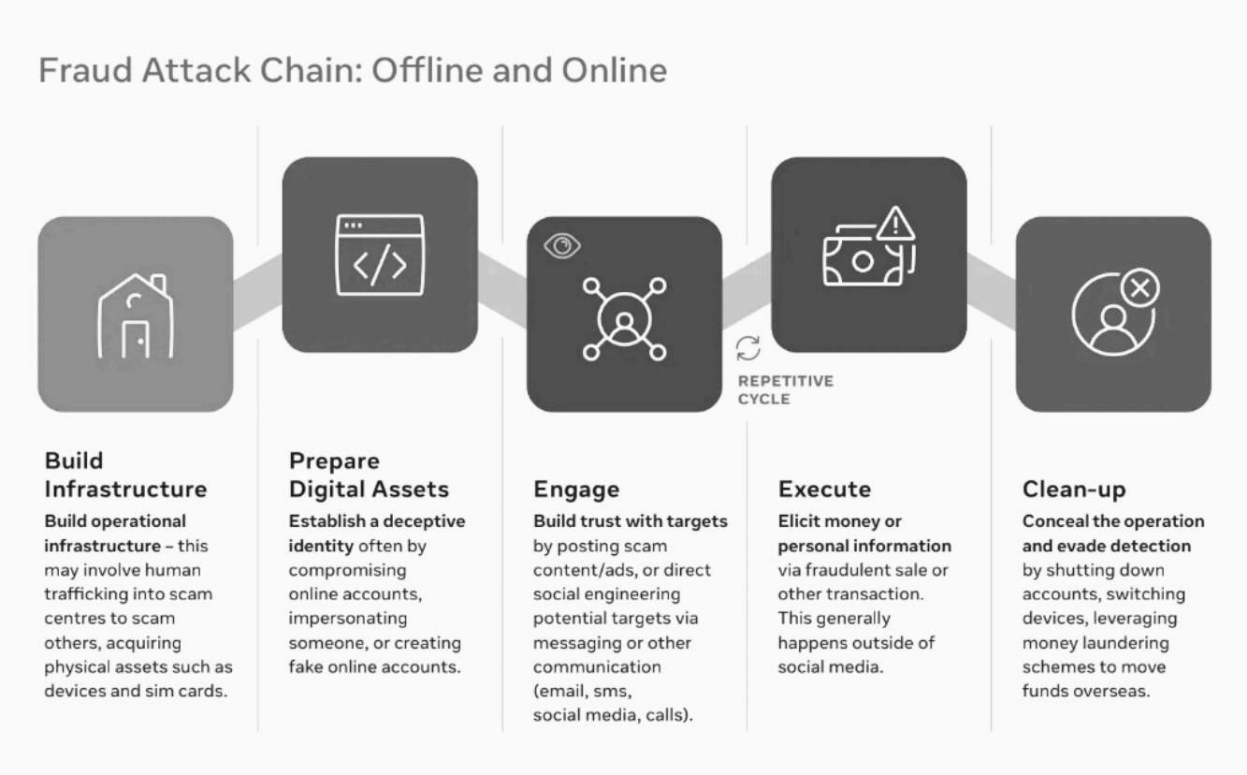

Image: An Overview of the Fraud Attack Chain.

The Fraud Attack Chain consists of five key stages, each representing a critical component of a scammer's operation:

1. **Build Infrastructure:** Operations begin to acquire the necessary physical and digital assets - including devices, access to the internet, domains and payment instruments - often using illicit marketplaces or by exploiting gaps in identity verification systems.
2. **Prepare Digital Assets:** Scammers create, purchase, or compromise online accounts to establish a facade of legitimacy. This may include aging accounts to build trust signals, acquiring verified profiles, or stealing credentials from legitimate users.
3. **Engage:** Adversaries initiate contact with potential targets, deploying social engineering tactics to build rapport and establish trust. Increasingly, this phase involves sophisticated lures - such as lifestyle content, fake job offers, or impersonation of trusted figures - designed to move targets to off-platform communication channels where oversight is limited.
4. **Execute:** The target is convinced to part with money, personal information, or access credentials. This stage represents the culmination of the scam and the point of direct financial harm.
5. **Clean Up:** Adversaries cover their tracks, launder stolen funds through complex financial networks, and often recycle their infrastructure for future operations - creating a persistent, repeating cycle of abuse.

The Strategic Value of the Fraud Attack Chain

The primary importance, and use of this chain, lies in its utility as a collaborative tool for the entire defender community and wider ecosystem. By mapping adversarial behaviors to this framework, we more clearly identify what and which type of defender, and intervention is best positioned to disrupt at each stage.

Throughout this, and future reports - we'll refer back to the chain, both directly and indirectly. Taking action against scammers as early in the chain as possible increases the costs for adversaries and decreases the risk to users. In this specific report we'll talk about our collaborative efforts with law enforcement and through the legal system, with the goal of increasing real-world-risk for criminals and abusive businesses. We continue to partner with external stakeholders with visibility in the "execute" phase - primarily financial service providers - to share signals, and receive insights to improve our mutual ability to detect and respond to scam activity earlier in the attack chain.

By operationalizing this framework, we aim to foster more effective, cross-sector collaboration - creating a more resilient, whole-of-society defense against the full spectrum of fraudulent activity.

Emerging Trends

Similar to our reports on other categories of harm, this document outlines new adversarial threats, trends, and adaptations. Scammers, as previously noted, are among the most persistent and adaptable adversaries the defense community encounters today - on par with state-funded adversaries covered in other sections of this report.

The challenges we face in combating fraud and scams are not unique to any single platform, sector, or geography. These are whole-of-society problems that require whole-of-society solutions. Scammers thrive in the gaps between defenders - exploiting seams in our collective visibility and coordination. As such we detail these emerging trends to support and inform that collective defense. Several trends we aim to highlight here:

1. **AI-Enabled Scam Operations:** Generative AI continues to lower the barrier to entry for scam actors while simultaneously increasing the quality, sophistication, and scale of their operations. Despite these advances in adversarial use of AI, we continue to adapt our behavior and technical-based detections to combat these more sophisticated adversaries. We are observing adversaries leverage AI for:
    - **Persona development:** Scammers continue to use AI to create convincing fake accounts with coherent backstories, authentic-looking profile pictures and video, which makes it increasingly difficult for the public to distinguish fake accounts from real ones. Deepfake video and voice cloning technology continues to grow as a vector, enabling threat actors to impersonate real individuals across a variety of channels.
    - **Content creation:** Attackers are now employing AI to generate multi-turn, believable conversations before delivering payloads. These AI-generated lures are increasingly 'hyper-personalized' and 'culturally nuanced,' mirroring the professional tone of target organizations or adapting seamlessly to local dialect and slang.
    - **Adversarial evasion:** We continue to track the use of AI in attempts to bypass platform defenses. Deepfake video is being used to evade likeness detection. Off-platform, we've seen the use of AI to facilitate rapid creation of fake businesses for purposes of scamming individuals - complete with websites, email trails, and fraudulent documentation.
2. **Professionalism of Criminal Scam Syndicates:** The industrialization of scam operations, particularly those originating in Southeast Asia, continues to expand. These syndicates have adopted the organizational structure and discipline of legitimate businesses, complete with specialized roles, training programs for scam center "employees" and sophisticated operational security. While criminal scam syndicate-associated human trafficking - where individuals are coerced or forced into working as scammers- is still common, this emerging vector indicates people are joining these operations willingly.

3. **Refocused Targeting:** We are seeing increased specialization in targeting, with certain scam archetypes (we'll detail these below) specifically designed to exploit vulnerable communities- including the elderly, those experiencing financial hardship, grieving families, and individuals seeking housing or employment. These targeted approaches reflect a calculated effort to maximize exploitation of those least equipped to identify or recover from fraud. We continue to see evolution across a range of vectors in these cases:
    - **Lifestyle and authenticity lures:** Rather than immediately pitching financial opportunities, adversaries increasingly cultivate long-term relationships through aspirational lifestyle content, building authentic-seeming social media presences before pivoting to scam narratives.
    - **Persona evolution:** We've observed shifts in the types of personas scammers adopt - including a notable move away from A-list celebrities, and other high-profile individuals, to local professionals and other trusted figures within smaller communities.

Understanding and Countering Scam Networks: An Overview of Scam Archetypes

A foundational understanding of common adversarial tactics and techniques is crucial to ongoing efforts to protect users and disrupt malicious activity across the internet. This section provides an overview of scam archetypes as a framework for categorizing and understanding the narratives and techniques employed by scammers. By familiarizing ourselves with these archetypes, and the attack chain (detailed above), we can collectively improve our ability to identify, report, and prevent scams.

Why, and When Do We Define an Archetype

We use archetypes as a way to classify recurring patterns of behavior and narrative that form the basis of common online scams. It's our attempt to represent the foundational playbooks that adversaries attempt to use to target users. Rather than focusing on the specific details of a single scam, which can change rapidly, archetypes allow us to recognize the underlying structure and intent of the malicious activity. Understanding how to mitigate against the core strategies scammers employ is a critical component of our layered defense approach.

Understanding these archetypes is beneficial for several reasons. For users, it provides a mental framework to recognize potential threats, even if they have not encountered a particular scam before. We publish these archetypes on our Scam Prevention Hub, along with other educational materials to help our users spot potential scams. For platforms and the security community, archetypes enable more effective and scalable detection and enforcement by targeting the root behaviors rather than just the surface-level indicators. This approach also allows us to anticipate and respond to new variations of old scams more efficiently.

Common Scam Archetypes

Our analysis has identified several prominent scam archetypes that are frequently employed by malicious actors across the internet. While the specific narratives and lures may vary, they generally fall into the following categories. We have observed these archetypes being used in a wide range of malicious activities, from simple phishing campaigns to more complex, multi-stage operations.

In-Depth Look at Scam Archetypes

Advance Payment Scams (also known as advance fee scams) often involve the impersonation of legitimate businesses or individuals to deceive targets into sending money before goods or services are received. These actors often create a sense of urgency (a sale, a required deposit) or a problem with a recent transaction to prompt immediate action. For example, a target might receive an email claiming an issue with a recent order, directing them to a phishing site to "verify" their payment details.

Investment Scams lure targets with the promise of high returns on "low-risk" investments. These scams often use sophisticated marketing and may even create fake investment platforms to appear legitimate. Adversaries will pressure targets to invest quickly, claiming that the opportunity is time-sensitive, and often use fake or impersonating personas to assert experience and hype past investment success. These scams have become increasingly prevalent with the rise of cryptocurrencies and online trading platforms.

Gambling Scams exploit the allure of easy money by promoting fake online gambling or betting sites. These sites are designed to steal financial information or to have users deposit money that they will never be able to withdraw. They often use deceptive advertising, such as "free bets" or guaranteed wins, to entice users.

Romance Scams are a particularly insidious form of social engineering where adversaries create fake online personas to build relationships with their targets. Over time, they cultivate a sense of trust and intimacy before manufacturing a crisis that requires financial assistance. These requests can range from needing money for a medical emergency to funds for travel to meet the target.

User Support Scams involve the impersonation of customer support representatives from well-known companies, including Meta. The adversary will contact the target, often through a direct message or email, claiming there is an urgent issue with their account. They will then direct the target to a phishing page to steal their login credentials, and in some cases, their two-factor authentication codes, leading to account takeover.

Celebrity Impersonation Scams leverage the public profiles of celebrities and other public figures to lend credibility to their fraudulent schemes. These scams can take many forms, from promoting fake investment opportunities to directing users to phishing sites. The use of a familiar face can lower a user's guard and make them more susceptible to the scam.

Government Impersonation Scams involve adversaries posing as government agencies or officials. They may offer fake government grants, tax refunds, or other benefits to lure targets into providing

personal and financial information. These scams often play on the authority and trust associated with government institutions.

The Evolving Threat Landscape: Adaptation and Innovation

While the archetypes discussed above provide a valuable framework for understanding the foundational strategies of malicious actors, it is crucial to recognize that the threat landscape is not static. Adversaries are constantly evolving their tactics, techniques, and procedures (TTPs) in response to our enforcement efforts and to exploit new technologies and social trends. They rarely operate within the neat confines of a single archetype, often blending elements from multiple categories to create more sophisticated and deceptive scams.

Our focus, therefore, is not just on disrupting the known threats, but on studying the behaviors of these adversarial networks to anticipate their next moves. We are seeing a continuous adaptation as they probe our defenses and seek to make their malicious activity more costly and difficult to detect. Below, we present several brief case studies that showcase some of these adaptations away from the traditional archetypes. These examples highlight the innovative and often complex nature of modern online threats and underscore the need for a dynamic and forward-looking approach to security.

Criminal Scam Syndicates

We removed over 10.9 million Facebook and Instagram accounts, 600,000 Facebook Pages, and 112,000 Ad Accounts in 2025 for violating our policies against Fraud, Scams and Deceptive Practices and Dangerous Organizations and Individuals.[1] This was part of our efforts to address Criminal Scam Syndicates. Analysis of these enforcements offers insights into the criminal networks responsible and trends in their locations and behaviors. Industry actions, political events, and law enforcement efforts in Southeast Asia have all forced adversaries to adapt.

We have previously described our efforts to address Criminal Scam Syndicates in Southeast Asia under our policies against Fraud, Scams and Deceptive Practices and Dangerous Organizations and Individuals. This is a problem whereby criminal organizations use forced labor to conduct scams from dedicated compounds. Scam centers exist at a variety of locations worldwide, but the problem is particularly prevalent in Myanmar, Cambodia, and Laos. In this report we offer insights into trends in Criminal Scam Syndicate behaviors during 2025 based on our investigations and enforcements.

Increasing Sophistication

The industrialization of scam operations, particularly those originating in Southeast Asia, continues to expand. These syndicates have adopted the organizational structure and discipline of legitimate

[1] For additional details on how we use our Dangerous Organizations policies to combat Criminal Scam Syndicates, please see https://about.fb.com/news/2024/11/cracking-down-organized-crime-scam-centers/.

businesses, complete with specialized roles, training programs for scam "employees," and sophisticated operational security. While we still observe the human trafficking dimension previously reported - where individuals are coerced or forced into working as scammers - this emerging vector adds a layer of complexity where people are joining these operations willingly.

Criminal Scam Syndicates have also diversified their approaches to run more sophisticated scams than the romance and cryptocurrency scams, often popularized as “pig-butchering,” that rely on high volumes of initial target engagement. The syndicates attempt investment, illicit gambling, and remote work scams, but increasingly use targeted approaches akin to “spear phishing” or “whaling” methods in cybersecurity. The syndicates identify wealthy individuals or people in positions of public or corporate trust and then impersonate regulators, law enforcement, attorneys, or investment advisors to elicit access to financial accounts or persuade the target to pay “fines.” One such type of scam is known as the “digital arrest.” The syndicates may support this type of scam through complicated backstopping including building websites, use of artificial intelligence, and video calls with actors in uniform on sets.

Geographic Trends

CSS in Myanmar during 2025 appears to have been affected by civil conflict. Our investigations confirmed reports from governments, non-governmental organizations, and victims that centers in Karen State near the towns of Myawaddy and Shwe Kokko were responsible for most scams originating from the country. The industrial park south of Myawaddy known as “KK Park” was the most notable of these, but there were numerous, similar complexes in the region. We continue to take proactive measures to identify and remove activity from these centers. These criminal syndicates attempted many varieties of scam, but these favored the long form romance, investment, and cryptocurrency scams that are commonly referred to as “pig butchering.” Starting in October 2025, we observed a significant decrease in the number of scams originating from Myanmar. In 2025, the US government also took steps to counter criminal organizations that facilitate large scale cyber scams, human trafficking, and cross border smuggling, including by imposing sanctions and creating a scam center strike force.[2]

Criminal Scam Syndicate activity in Myanmar fell to negligible levels by the end of 2025, but has not ended entirely. We assess that many of the large industrial parks dedicated to scams are no longer active, but are investigating the dispersion of the criminal actors. We assess that scam center activity continues as smaller operations in contested areas of Karen State. We are also aware of scam center activity in Shan State and are continuing to take proactive measures to identify and remove activity from these centers.

Criminal Scam Syndicates in Cambodia appear to have been affected by political events and law enforcement actions. At the start of 2025, scam centers in Poipet, Cambodia and elsewhere in northern Cambodia were responsible for the highest volume of scams on our platforms in the region. These centers attempted a variety of scams including romance, investment, impersonation,

[2] See https://home.treasury.gov/news/press-releases/sb0129 for additional details about US actions against criminal organizations engaging in scams.

and illicit gambling. The border conflict between Cambodia and Thailand that started in July 2025 and resumed briefly in November 2025 marked the beginning of a decline in these volumes, although scam centers in the region remain prolific to the present.

The focus of Criminal Scam Syndicate activity in 2026 is southern Cambodia. Scam centers have been growing in southern Cambodia throughout 2025, particularly in Svay Rieng Province. We assess that pressures in northern Cambodia have prompted this movement to the south, but the centers have advantages by attaching to support structures at the casino properties near the Vietnam border. These centers attempt a variety of scams including romance, investment, and impersonation. Illicit gambling is a particular focus of these scam centers.

Content Trends

We disable most Criminal Scam Syndicate accounts early in the lifecycle of a scam and before they ever attempt to contact a target, so often we cannot determine what type of scam they intended to run or who the scammers intended to target. We estimate scam types and intended targets based on accounts we disable later in the lifecycle or how a scammer configured an account.

The majority of Criminal Scam Syndicate accounts are configured with “lifestyle” media, meaning a scammer posts content that is intended to make an account appear authentic. This includes media that shows travel, food, fashion, luxuries, or other content intended to copy influencers. It may precede any type of scam, but accounts configured in this style may be used for scams involving one-on-one engagement such as romance scams.

The following is a set of “lifestyle” content posted to one account operated by a Criminal Scam Syndicate in Cambodia. The account represented itself as a UK-based influencer and was probably intended for romance scams. This content is typical of the archetype. It may be recognized for resembling stock photos and having relatively low social engagements, such as follows, likes, or comments. In this instance the scammers chose photos that did not show a face, probably to preserve options if they successfully engaged a target and were requested to show themselves on a live application.



Image: Sample of “lifestyle” content used by Criminal Scam Syndicate operators in Cambodia.

Approximately a third of Criminal Scam Syndicate accounts attempted to sell an item, which is indicative of no-delivery, advanced fee, or remote work promotion scams. Approximately one third of accounts posted gambling content. Approximately a fourth of accounts attempted to engage in investment scams. Less than a fourth of accounts attempted to engage in impersonation scams, including celebrity impersonation or government impersonation. Some accounts attempted multiple types of scams.

The following are examples of remote work scams operated by Criminal Scam Syndicates. The content typically offers high wages for “easy” side tasks and may target parents, students, caregivers or others expected to be at home. The scam occurs when victims are requested to pay upfront fees for equipment, access, tiers, or transactions on their “salary.” In some cases, the victim may be asked to perform logistics tasks that aid the scammers such as engagement farming or transshipment of merchandise or money.

Image: Examples of remote work scams operated by Criminal Scam Syndicates.

The following are examples of gambling scams operated by Criminal Scam Syndicates. They often redirect to casino applications or websites and can be recognized by flashy, AI-generated cartoons, people, and other imagery. It is difficult to determine whether their gambling applications are outright scams, seizing all money with no payouts, or whether they operate with rigged odds, fees, or other methods to obtain income. Regardless, the applications are operated by a criminal organization. Gambling scams typically target victims in Southeast Asia, Mandarin speaking regions, or the Pacific region.

Image: Examples of gambling scams operated by Criminal Scam Syndicates.

The following are examples of investment scams operated by Criminal Scam Syndicates. They often redirect to mobile messaging groups purportedly where investment advice is offered and can be recognized by promises of insider tips, unrealistic gains, and unbacked claims of prior successes. They may also impersonate or show media depictions of celebrities or prominent financial advisors, executives, and officials.

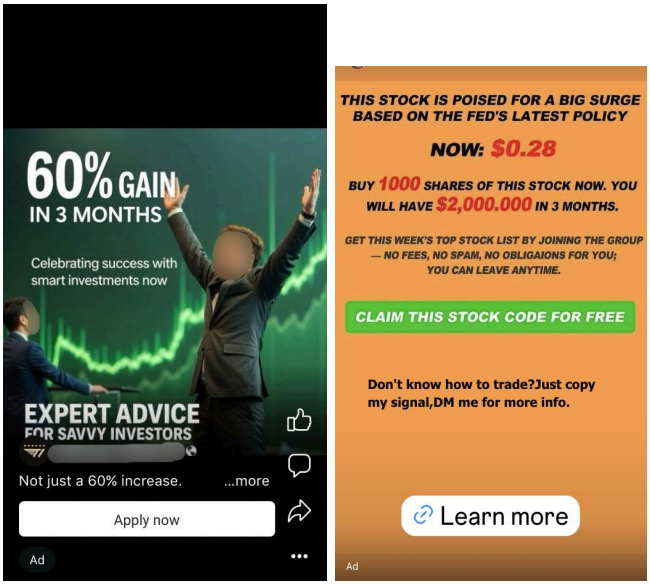

Image: Examples of investment scams operated by Criminal Scam Syndicates.

The following are examples of impersonation scams, specifically those intended to further recovery scams. Our previous Adversarial Threat Report focused on these scams, which impersonate

government officials, law enforcement, regulators, attorneys, or advocacy organizations with promises to help those who have been previously scammed recover their funds. These can be recognized by content that purports to originate from a government, but not connected to their official, verified social media presence. They may also reference non-existent or generic organizations, but some scammers are sophisticated with the impersonation.

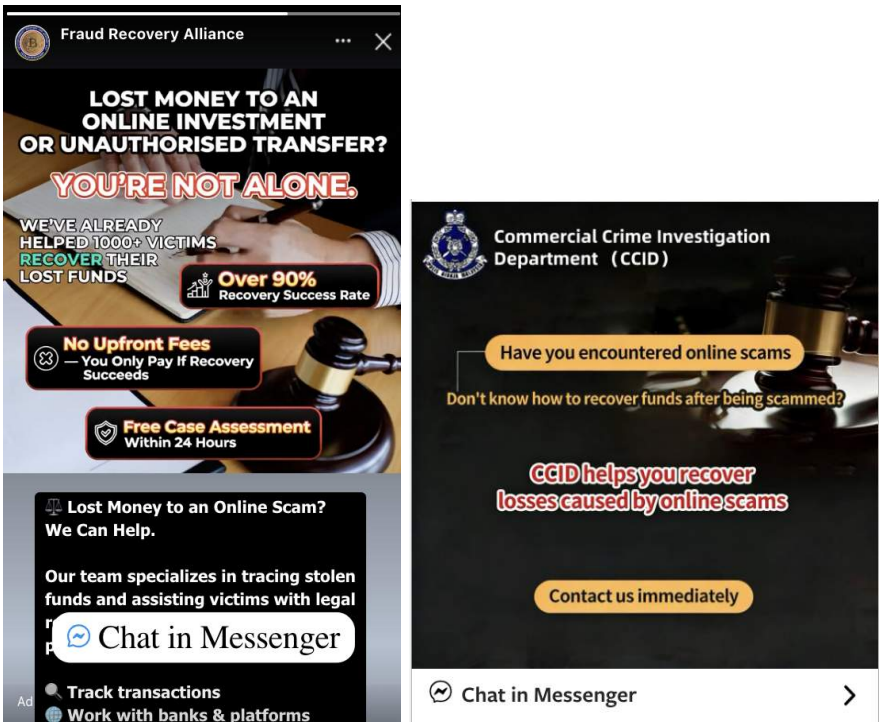

Image: Examples of impersonation scams operated by Criminal Scam Syndicates.

Target Trends

Our investigations surfaced that Criminal Scam Syndicates most frequently targeted English-speaking users across the US, followed by users in India. Mandarin speaking audiences including those in China, Taiwan, Hong Kong, and Singapore were the next groups targeted, with targeting of audiences in Japan and South Korea close behind.

These targeting concentrations likely reflect both the expected returns of larger, higher-income markets and the operational realities of these compounds where coerced or recruited workers are typically trained to conduct outreach in English, or in their own major regional languages - which enables higher volume targeting across these countries.

Criminal Scam Syndicates are opportunistic and may target individuals anywhere, and not exclusive to these countries. We also observe efforts to target other English-speaking populations, including in Australia and the UK, European Union members, and Brazil. There are reports that authorities in the UAE and Nigeria have also disrupted scam centers that operate on similar models, raising prospects that both the criminal actors and targets have spread outside Southeast Asia. As

scammers develop and adapt their methods, we continually seek to improve and enhance our enforcement processes to counter their efforts.

Romance Scammers Evolve from Military to Medical Professional Personas

We disrupted a network of about 23,000 users across Facebook and Instagram for violating our policies against Fraud, Scams and Deceptive Practices. The network employed deceptive personas—specifically posing as international doctors, military doctors, and in some cases claiming affiliation with organizations like Doctors Without Borders (Médecins Sans Frontières)—to gain targets' trust for romance scams.

This network showcases an evolution in romance scammer tactics. Previously, we have extensively observed and disrupted romance scam networks that primarily relied on military personas, leveraging the inherent trust and integrity associated with military service members to quickly build rapport with targets. However, threat actors have adapted their approach, shifting toward medical professional personas that similarly exploit public respect for healthcare workers and humanitarian organizations. In some instances, actors went beyond generic medical personas to impersonate real, publicly known doctors, adding an additional layer of credibility to their deceptive operations. This tactical evolution demonstrates the adversarial actors' continued adaptation to enforcement pressures and their strategic understanding of personas that engender trust from targets.

Our investigation leveraged a combination of behavioral and content signals to identify this network. Following disruption of the initial network, we implemented automated measures designed to detect and remove behavior that fits this pattern going forward. We continue to refine detection protocols and classifiers to address this evolving scam archetype.

Scammers Leverage Generative AI to Scale Animal Rehoming Fraud Operations

We disrupted a Cameroon-based network consisting of 100 groups and approximately 12,000 accounts across Facebook and Instagram that were leveraging AI to facilitate animal rehoming scams. The network used generative AI tools to enhance their operations by generating convincing communications with targets and product descriptions, assisting with translations, building cultural fluency and contextual understanding for targeting specific regions, and creating realistic animal imagery for their deceptive listings.

This case represents a significant evolution in scammer tactics, demonstrating how threat actors are adapting to incorporate AI-generated content into traditional fraud archetypes. The network posted fake pet adoption content—both organic posts and paid advertisements—to lure targets, then requested upfront payments under the guise of covering veterinary care, transportation, or rehoming fees. Our investigation revealed that scammers used AI to scale their operations, generating location-specific content and culturally appropriate messaging to improve targeting in the United States, Canada, Australia, New Zealand, and the United Kingdom. To expand their reach

and identify targets, the network actively joined and created animal and pet-related Groups, using these communities to identify targets. The operation employed a multi-faceted approach, running targeted click-to-message advertising campaigns focused on initiating conversations and directing targets to off-platform messaging services.

Following the initial network removal, we implemented scaled detection and enforcement measures that continue to identify and remove similar fraudulent activity. We used specialized techniques to detect and analyze the content and behavior of these scams, which enabled us to find links across the network. We have used these signals to enhance our detection capabilities associated with AI-assisted scam operations. This case underscores the importance of adapting our enforcement strategies to address the evolving landscape of AI-enhanced fraud, where traditional scam archetypes are being augmented with generative AI to increase operational efficiency and deceptive capability.

New scam archetypes: Fake Rentals and Funeral Livestreams scams preying on particularly vulnerable populations:

Scammers Target Vulnerable Housing Seekers with Coordinated Fake Rental Listings

We disrupted a coordinated network of approximately 63,000 accounts operated primarily from Bangladesh and Nigeria that were posting fake rental property listings on Facebook, specifically targeting vulnerable populations including low-income families and unhoused individuals seeking affordable housing in the United States. The network employed deceptive landlord personas, often presenting themselves as emergency housing providers with names like "HOMELESS Emergency Rentals" and "Low income house for Rent." These accounts posted content promising no-down-payment rentals and explicitly targeting users searching for subsidized housing assistance, exploiting the desperation of those in precarious housing situations.

The operation demonstrated sophisticated coordination and infrastructure sharing across the network. Our investigation revealed that scammers were directing targets to externally-hosted websites through seemingly legitimate property listing links, where users were prompted to complete surveys requesting personal and financial information including income levels, debt amounts, and contact details. Despite the initial appearance that these were disparate rental scams, we identified a significant amount of coordinated activity within this network. The collected data was subsequently sold to marketing services, or used to send targets low-quality affiliate marketing campaigns. We partnered with external researchers at Graphika to confirm that, within weeks of providing their information, individuals received hundreds of follow-up communications including deceptive or low-value financial offers.

Following our initial network disruption, we implemented both immediate enforcement measures and long-term product safeguards to prevent recurrence of this abuse pattern. We disabled approximately 63,000 accounts through strategic network enforcement and established automated detection systems that continue to identify and remove fraudulent accounts engaged in similar rental scam activities. We also implemented monitoring systems to detect potential migration of these tactics elsewhere on our platforms.

Scammers Target Grieving Families with Fake Funeral Livestreams

We identified and disrupted an emerging scam archetype targeting one of the most vulnerable populations: families and friends of recently deceased individuals. In this specific case we disabled a Bangladesh-based network consisting of approximately 8,600 accounts, 20 Pages, and took enforcement action against 7,300 associated account operators which were conducting fake funeral livestream scams on Facebook. This operation represents a particularly egregious form of fraud that exploits grief and the desire of loved ones to participate in memorial services, demonstrating how threat actors continue to identify new ways to monetize human vulnerability and emotional distress.

The network employed a systematic approach to target bereaved individuals through coordinated social engineering tactics. Scammers created deceptive profiles and Pages impersonating funeral homes, celebration of life services, and other bereavement-related organizations, often incorporating those keywords in their names and biographical information. The operation involved threat actors conducting searches for the names of recently deceased individuals and subsequently targeting their social networks by sending unsolicited friend requests to family members and friends. Once connections were established, the scammers would direct individuals to external websites disguised as legitimate livestreaming platforms, where users were prompted to enter credit card information under the guise of paying fees or making donations to access the purported funeral broadcast. Analysis of the network revealed operational security measures, including the systematic deletion of previous scam posts to avoid detection by future targets and the use of link shortening services to obscure malicious URLs.

Following the initial disruption, we implemented additional safeguards to prevent recurrence of this abuse pattern. We have established ongoing monitoring systems specifically designed to detect the behavioral signatures and content patterns associated with fake funeral livestream operations, including automated detection of profiles using bereavement-related keywords combined with suspicious geographical and behavioral indicators. Additionally, we have enhanced our detection capabilities to identify the specific link patterns and off-platform redirection techniques employed by these threat actors. As with all adversarial threats, we continue to adapt our enforcement strategies as threat actors evolve their tactics and develop new techniques to circumvent detection systems.

Signals Sharing- FIRE Case Studies

In our last report, we highlighted the importance of information sharing partnerships across the defender community to mutually advance efforts against fraud and scams, and enable action to be taken earlier in the Fraud Attack Chain. Our Fraud Intelligence Reciprocal Exchange (FIRE) program broadens defense community collaboration in combating scams by providing a platform for bilateral information sharing between Meta and the financial industry. Below, we present two case studies of network disruptions we undertook as a result of information shared through the FIRE program.

Disruption of Scam Network Involving Relationship Scams Targeting Millennials and Older Men

We removed, disabled and unpublished over 15,000 assets on Facebook and Instagram that used deceptive personas claiming to be Japanese women seeking relationships with millennials and older men, with some accounts also promoting gambling-related content. These groups violated our policies against Fraud, Scams and Deceptive Practices. These assets engaged in violating behaviors that encompassed relationship scams, suggestive content, and online gambling. Our investigation into this network began based on information we received from the FIRE program. The investigation revealed operations primarily based in China, Myanmar, Colombia, Philippines, Indonesia, the US, and Nigeria, targeting users in Japan and other regions. The network's primary target audience was users, particularly millennial or older men, in Japan and across the region.

The key tactic involved creating deceptive personas: fake Instagram and Facebook profiles impersonating young Asian women seeking romantic relationships. These scammers utilized tactics such as: developing fake profiles featuring attractive Asian women in revealing clothes, including bios specifically targeting men over the age of 40, and directing conversations off-platform to private channels like LINE or Messenger.

This operation demonstrated a significant scale of reach, effectively engaging millennial and older male users through the use of deceptive personas and off-platform communication.

Disruption of Scam Network Involving Military and Bank Impersonation and Investment Scams

We investigated and removed a network for violating our policies against Fraud, Scams and Deceptive Practices. Our investigation resulted in enforcement actions against over 3,000 assets, including Facebook profiles, Instagram profiles, ad accounts, and pages. Our enforcement included disabling accounts, unpublishing pages, and removing assets related to deceptive personas, and investment scams. Our investigation began based on information we received through the FIRE program. The investigation revealed operations primarily based in Nigeria and the United States, targeting audiences in the US, Canada, and other regions. This network involved military and bank executive impersonation and investment scams.

Scammers used tactics such as creating fake profiles impersonating US military officials and bank executives, posting offers with urgent language and images of cash, and promoting investment scams with vague promises of high returns. They disguised their identity using VPNs and proxy servers, set profile locations to the US, and redirected users off platform to evade detection. The network also leveraged pages and ad accounts to amplify reach and target specific countries with tailored scam content.

Other Efforts to Combat Fraud and Scams

Legal Actions Against Scam Advertisers

Meta continues to take legal action as part of our multi-layered approach to combat scams, including filing lawsuits against four deceptive advertisers who impersonated well-known celebrities and brands to deceive and defraud people on our platforms. These cases highlight the evolving sophistication of scammer tactics, from "celeb-bait" schemes to promote fraudulent healthcare products and investment scams, to cloaking techniques that circumvent ad review systems by showing different content to our automated systems versus real users. We also issued cease and desist letters to eight former Meta Business Partners who offered abusive services, including phony "unban" or account restoration services and renting access to trusted accounts that helped clients evade our enforcement systems.

These legal actions complement our technical enforcement measures, which include suspending violators' methods of payment, disabling related accounts, blocking domain names for websites used in scams, and sharing this information with industry partners so they can take action on their platforms as well. The cases demonstrate how we combine on-platform detection and removal with off-platform legal pressure to increase real-world costs for bad actors and disrupt their ability to reconstitute their operations. This approach reflects our broader strategy of using multiple intervention points across the Fraud Attack Chain to create sustained pressure against financially-motivated adversaries who continue to adapt their tactics, techniques, and procedures to evade detection and enforcement.

The Role of User Reporting in the Attack Chain

As adversarial actors, scammers are constantly adapting their tactics and exploring new potential attacks. This adaptation makes measuring scams and progress against scams particularly challenging — focusing primarily on content or behavior will lag behind adversarial shift, leaving us with an incomplete (often overly narrow) impression of what is happening. One metric we have found to be resilient to this constantly shifting landscape has been user reporting. Calculated by tracking the number of user reports for scam ads relative to every million views, Scam Reports per Million Views (SRMV) provides a direct signal of how our users experience potential scams, encourages our users to serve as valuable resources in helping to identify potential scams, and allows us to build countermeasures to help better identify and remove scams.

User reporting thus becomes a valuable component of our measurement strategy precisely because it integrates our users directly into the Fraud Attack Chain. For certain scam types, like longer-tail scams (ones that take place across different surfaces, and take longer to appear) Users provide an essential line of defense. When a user reports a scam ad, they are actively participating in the Disruption phase of the attack chain, providing vital, real-time data that our automated systems and human reviewers might have missed and allowing us to measure the real-world impact of adversarial campaigns that seek to exploit our advertising systems for financial gain.

This direct feedback from our community, quantified by this approach, is instrumental in driving product and enforcement improvements. For example, a rise in reports for ads that misuse images of public figures to bait people into engaging with scams ("celeb-bait") directly informed our decision to develop and deploy advanced facial recognition technology to detect and block ads that leverage the images of public figures without authorization.

While user reporting is a vital tool, it is not an all-encompassing measure of the threat posed by scammers or the impact on our users. Users tend to under-report suspected scam activity. As such we must complement our user reporting with other metrics to gain a more comprehensive understanding of the scams threat environment, with each metric tailored to a different facet of the adversarial activity we are seeking to disrupt. Additionally, user reporting relies on the efficacy of the reporting tools available to our users, and we are continually working to improve our reporting systems.

Driving Down Scams with AI:

Fraud and scams are constantly evolving. Bad actors can use any archetype or otherwise legitimate activity to perpetrate a scam, change their tactics, use language designed to evade detection, and operate across different formats—text, images, and video—and different platforms across the online ecosystem. Increasingly, scams are not apparent from any one source; as the attack chain shows, scammers implement different components of a scam across multiple different locations, from aspects of our platforms to off-site locations. Scammers also rely on subtle tricks that are hard to catch with traditional technology such as deceptive framing or small inconsistencies between an account's name, bio, and behavior.

That's why Meta is investing in advanced AI systems to strengthen our ability to detect, and take action against scams. AI is especially valuable because it can analyze multiple signals together, such as text, images, and the surrounding context - similar to how a human would reason if something they saw was a scam. This helps us spot more sophisticated scam patterns that might slip past conventional approaches - at scale and speed.

Detecting Identity Abuse: AI is particularly effective at detecting identity abuse, where scammers pretend to be celebrities, public figures, or brands. These scams often use subtle signals that traditional systems may miss, such as fake fan sentiment, misleading bios, or associations with public figures or brands. For example, imagine a public figure who owns a winery. A scammer might create an account using the celebrity's first name and a photo of their wine bottle. A dedicated fan

might immediately recognize this - but a human reviewer unfamiliar with this celebrity might not. AI has potential to process far more contextual information about public figures than any individual reviewer, helping us catch these deceptive accounts.

Deceptive Links and Domain Impersonation: A particularly effective mitigation tactic is to focus on delivery vectors – we've been focusing on areas like domain name and landing page impersonation as a common path for scams which may seek to direct individuals to deceptive webpages posing as legitimate websites. These tactics are designed to deceive users and undermine trust. To address this challenge at scale, we have deployed advanced AI systems to proactively detect and enforce our policies against such content. This technology allows us to identify a much broader range of brands and deceptive behaviors with high precision, moving beyond a small set of known entities to protect against thousands of brands being impersonated.

Advertiser Verification

We are always working to make our platforms safer for people and businesses. Today, advertisers may be required to verify depending on factors like where they deliver ads and whether they have a history of not following our rules or the type of ads they run are more susceptible to abuse.

We are expanding advertiser verification, to help ensure that verified advertisers drive 90% of our ads revenue by the end of 2026, up from 70% today. It will cover the highest-risk categories, while the remaining 10% will come from low-risk businesses, like your local ice cream shop. The verification process helps promote greater transparency, limiting attempts to misrepresent advertiser identity. It is an important part of our multi-layered approach to help protect people on our apps from scams.

02

TAKING ACTION AGAINST ‘NUDIFY’ APPS

What are ‘Nudify’ Apps

We are committed to fighting all forms of sexual exploitation on our platforms. This includes tackling the concerning growth of so-called ‘nudify’ apps, which use AI to create fake non-consensual nude or sexually explicit images. Meta has longstanding rules against this type of imagery, and more than one year ago we updated our policies to make it even clearer that we don’t allow the promotion of nudify apps or similar services. We remove ads, Facebook Pages and Instagram accounts promoting these services when we become aware of them, block links to websites hosting them so they cannot be accessed from our platforms, and restrict search terms like ‘nudify’, ‘undress’ and ‘delete clothing’ on Facebook and Instagram so they don’t show results.

Taking Action Against ‘Nudify’ Apps On Our Platforms

Like other types of online harm, this is an adversarial space in which the people behind these apps and the crimes affiliated with their use—who are primarily financially motivated—continue to evolve their tactics to avoid detection. For example, some people who make and market these apps use benign imagery in their ads to avoid being caught by our nudity detection technology, while others quickly create new domain names to replace the websites we block.

That is why we are continuously evolving our enforcement methods. We have worked with external experts and our own specialist teams to expand the list of safety-related terms, phrases and emojis that our systems are trained to detect within ‘nudify’ app ads. In addition, we have developed new technology specifically designed to proactively identify these types of ads—even when the ads themselves do not include nudity—and use matching technology to help us find and remove copycat ads more quickly.

As part of this work, last year, our teams ran six in-depth investigations to expose networks of accounts that attempted to run ads promoting these services. As a result, we actioned 771 Facebook Users and 230 Pages for violating our policy against Adult Sexual Exploitation. Unlike some of the areas highlighted in this report, this case study is not based on a single network, but includes aggregate actions resulting from six investigations of networks engaged in spreading ‘nudify’ content. In addition, between November 2025 and January 2026, we removed over 344,000 ads across Facebook and Instagram that attempted to promote ‘nudify’ apps.

Fighting 'Nudify' Apps Across the Internet

With 'nudify' apps being advertised across the internet—and available in App Stores themselves—removing them from our platforms is not enough. For nearly one year now, when we remove ads, accounts or content promoting these services, we share information—starting with URLs to violating apps and websites—with other tech companies through the Tech Coalition's Lantern program, so they can investigate and take action too. Lantern is a signal-sharing program that enables participating tech companies to share indicators about accounts, behaviors, and related signals that violate their child safety policies, so each company can investigate and take action on its own platform. Since we started sharing this information at the end of March 2025, we have provided more than 5,500 unique URLs to participating tech companies.

We also bilaterally share signals about violating child safety activity, including sextortion, with other companies, and this is an important continuation of that work.

Taking Legal Action

Beyond removing 'nudify' content from our platforms and enabling industry partners to take action, we also pursue legal action against those who abuse our policies on sexual exploitation. We have sued Joy Timeline HK Limited, the entity behind CrushAI apps, which allow people to create AI-generated nude or sexually explicit images of individuals without their consent. We filed the lawsuit in Hong Kong, where Joy Timeline HK Limited is based, to prevent them from advertising CrushAI apps on our platforms.

Additionally, in September 2025, we issued cease and desist letters (C&Ds) to 46 other companies that violated our terms by advertising 'nudify' apps—such as Undressly and Crushlove—on our platforms.

In November 2025, we also issued C&Ds to nine developers and/or advertisers in China, Hong Kong, Singapore, Vietnam, and the United Arab Emirates (UAE), linked to six so-called "kissing apps"—AI tools advertised on our platforms as enabling users to generate fake images of people kissing.

These legal actions underscore both the seriousness with which we take this abuse and our commitment to doing all we can to protect our community from it.

Supporting Prevention and Response

We welcome legislation that helps fight intimate image abuse across the internet, whether it is real or AI-generated, and that complements our longstanding efforts to help prevent this content from spreading online through tools like StopNCII.org and NCMEC's Take It Down or CyberTipline. That's why we championed the TAKE IT DOWN Act, an important step to addressing this abuse across the internet and supporting those affected, and we expect to fully implement its requirements several months ahead of the deadline.

We also continue to support legislation that empowers parents to oversee and approve their teens' app downloads. As well as sparing them the burden of repeated approvals and age verification

across the countless apps their teens use, this would allow parents to see if their teen is attempting to download a ‘nudify’ app from an App Store, and prevent them from doing so.

03

CARTELS/DRUG-TRAFFICKING ORGANIZATIONS

Drug cartels operating in Latin America, and the associated networks they play a key role in supplying and supporting, present pressing and complex risks. The global drug trafficking market was estimated to have a value of between $840 billion and $1.44 trillion in 2024. The White House recently estimated that in 2023, the cost of the illegal opioid trade to Americans alone was $2.7 trillion, factoring in loss of life and quality of life. High overdose death rates in the United States and Canada, and continuing cartel-related violence in Mexico and other Latin American countries, add to the significant human cost of these activities. In light of the harms caused by cartel violence and the illicit drug trade, the US government has taken the unprecedented step of designating a number of cartels as Foreign Terrorist Organizations and Specially Designated Global Terrorists.

Combatting the criminal organizations that try to exploit our platforms through propagandizing, recruitment and coordination of criminal activities remains a central focus of Meta's integrity efforts. Under our Dangerous Organizations and Individuals policy, we deploy the same level of enforcement against criminal organizations that we do against terrorist groups.

Most of Meta's enforcement against these groups comes from routine content enforcement undertaken at scale. However, the adversarial behavior of these networks necessitates the use of novel investigative techniques and subject-matter expertise. Online networks operated by drug cartels and illegal narcotics networks are highly adversarial, agile and persistent - incentivized to maintain online presence by the enormous profits that their illegal activity can generate. On-platform, these actors use obfuscation, coded language and other deception methods in an attempt to evade enforcement. Meta uses a combination of AI and human review teams to detect, contextualize, analyze and exploit video, images, audio and text data related to cartels and illegal narcotics networks. As well as leading directly to enforcement actions, this tailored approach allows us to observe the ways in which these groups seek to evade our enforcement, so we can learn to counteract those methods at scale.

One important tool that Meta uses to counter dangerous organizations, including cartels and illegal narcotics networks, is Strategic Network Disruptions - an expert-led, targeted approach that detects and removes whole networks of bad actors. Meta uses Strategic Network Disruptions to identify and remove networks of actors who seek to abuse our policies across a range of harm areas, including those mentioned elsewhere in this report.

This section examines a series of case studies of Strategic Network Disruptions deployed under our Dangerous Organizations and Individuals policy in 2025. It details innovative detection,

investigation, enforcement and prevention methods Meta has recently begun to deploy, to more efficiently counter the specific challenges posed by cartels and illegal narcotics networks. In 2025, we significantly advanced our Strategic Network Disruption capability by establishing a dedicated team of experts focused exclusively on illegal narcotics networks and Criminal Organizations. The use of new techniques by this team contributed to a significant increase in Criminal Organization account removals resulting from Strategic Network Disruptions under our Dangerous Organizations and Individuals policy, between 2024 and 2025.

Criminal organizations use a wide range of social media platforms and online services. Effectively disrupting their operations requires a decisive, collaborative effort from private companies, civil society organizations, government, and law enforcement. We hope that in sharing these enforcement details, we can help others understand this threat and strengthen their efforts against the cartels.

Below, we describe some of the new tools and methods we have applied, and examine their impact through exploring a series of case studies of disruptions we have undertaken. We also describe a growing element of our Dangerous Organizations and Individuals prevention efforts - youth-focused messaging campaigns designed to deliver educational content, in order to help protect young people from potential harm.

Case Studies - Enhancing Strategic Network Disruptions

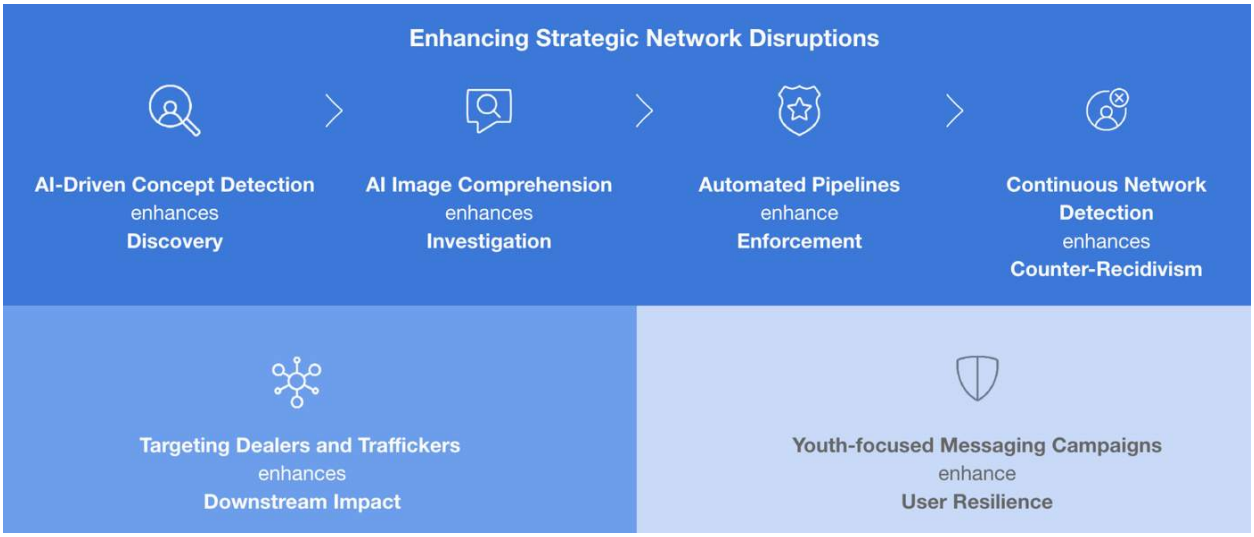

Image: An overview of novel methods now being deployed in Strategic Network Disruptions.

AI-Driven Concept Detection

Dangerous Organizations and Individuals Strategic Network Disruptions often begin with the identification of emerging threats and adversarial behaviors on our platforms. AI-Driven Concept Detection is a new content discovery method that greatly improves the efficiency with which we

can identify violating text, and helps us overcome adversarial obfuscation techniques. This means Meta is better able than ever before to discover adversarial actor networks.

Crystal Meth Dealing Network in Mexico

Meta completed an investigation into the promotion of crystal methamphetamine sales in Mexico. Using AI-Driven Concept Detection, Meta surfaced accounts and content associated with violations of our High Risk Drugs Policy, attempting to sell, buy, or promote drugs, such as crystal methamphetamine. We actioned about 51,000 objects for violating our High Risk Drugs Policy.

Cartel Networks

Meta has also used this method against cartel networks directly. In 2025, Meta actioned 101 Facebook accounts, 1 Group and 54 Instagram accounts for violating our Dangerous Organizations and Individuals policies. A number of these accounts, operating in Mexico, appeared to be linked to recruitment efforts and the promotion and glorification of criminal activity. Meta also conducted a spin-off investigation into another cartel faction, which we similarly identified through AI-driven concept detection. As a result, we actioned 62 Facebook accounts, 4 Groups, 16 Instagram accounts and 1 Page for violating our policies against Dangerous Organizations and Individuals, which were seeking to potentially engage in violent extortion. This method has also been deployed in the United States, where Meta detected and enforced a network of accounts linked to cartel-affiliated actors, primarily along the Southwest border. In this instance, we actioned 197 Facebook accounts, 1 Group, 1 Page, and 191 Instagram accounts for violating our Dangerous Organizations and Individuals policies. Meta used AI-assisted analysis to identify cartel-related content signals and expand from initial leads to additional violating accounts, as well as uncovering content for further network exploration. Several accounts in this network displayed indicators consistent with attempted illegal narcotics trafficking activity.

AI-Driven Concept Detection provides significant advantages in the discovery phase of Strategic Network Disruptions, particularly to expert teams that are already deeply familiar with cartel behaviors. Given this, Meta has scaled-up our use of this method across our enforcement work against criminal organizations and illegal narcotics networks, as well as applying it to other Dangerous Organizations and Individuals problem areas.

AI Image Comprehension

When Meta examines potential cartel and illegal narcotics dealing networks, our subject-matter experts have to sift through huge volumes of data to identify important signals, such as evidence of criminal activity. AI Image Comprehension significantly increases the efficiency and recall of this process, by allowing Meta experts to better understand the content of images while significantly reducing the need for manual review.

United States-Mexico Drug Trafficking Gangs

Meta experts completed a review of a set of accounts operating primarily between the southwestern United States and Mexico, which seemingly sought to use Meta platforms to promote and coordinate the distribution of illegal narcotics. We also observed adjacent activity

consistent with attempted weapons trafficking and financial fraud. To rapidly identify the full extent of potentially violating accounts, Meta used AI-enabled discovery workflows to expand from initial leads to additional associated accounts, as well as content that would have been difficult to uncover through manual review alone. Meta also used AI Image Comprehension to help surface and document relevant visual evidence, improving investigative efficiency and recall. Following this review, Meta actioned 55 Facebook accounts and 248 Instagram accounts for violating our policies against High Risk Drugs.

Cartel Networks

Meta used AI Image Comprehension as part of a broader set of investigative techniques, to evaluate content we believed might have been linked to multiple Mexico-based cartels and affiliated actors. Through this method, Meta was able to identify a cluster of cartel-linked networks which were violating our policies against Dangerous Organizations and Individuals by seeking to exploit our platforms to engage in credible threats of violence, as well as activity consistent with drug trafficking, drug production, and extortion. This investigation led to the identification of cartel-linked networks of accounts operating in the United States and Mexico. As a result of this work, we actioned 649 Facebook accounts, 59 Groups, 7 Pages, and 89 Instagram accounts for violating our policies against Dangerous Organizations and Individuals.

The use of AI Image Comprehension in Strategic Network Disruptions is not just leading to greater efficiency and recall, allowing our subject-matter experts to do more in less time. It's also facilitating the discovery of entirely novel networks of Dangerous Organizations and Individuals, and rapidly becoming a core method for our expert teams.

Automated Pipelines

The experts leading our Strategic Network Disruptions gather insights from the examination of on-platform criminal networks, and use those to create automated detection and enforcement pipelines. Advances in our automation methodology have improved the precision of this work, in turn increasing the number of pipelines that can be used not just to detect violating accounts, but to enforce on them, as well as on high-confidence similar accounts. In fact, 2025 was the first year in which automated pipelines created as a result of Strategic Network Disruptions removed more users than manual Strategic Network Disruptions themselves.

Automated Disruption of Drug-Related Scams

In Q4 2025, Meta successfully implemented a sophisticated automated enforcement pipeline specifically designed to remove accounts engaged in drug-related scams. In these scams, users seeking to buy drugs are defrauded by accounts they seek to buy drugs from.

This pipeline has demonstrated remarkable efficacy in detecting and actioning these illicit activities. Since its launch, we actioned about 156,000 objects for violating our policy against High Risk Drugs. This disruption pipeline is projected to have a sustained impact on the prevalence of drug-related scams within the Meta ecosystem, and underscores Meta's commitment to utilizing advanced technology to proactively address complex adversarial threats.

This is just one example of the use of automated enforcement pipelines, and our subject-matter experts are regularly creating and deploying similar new pipelines. This strategic investment in automated threat detection and remediation represents a significant advancement in our commitment to user safety and platform integrity.

Automated Disruption of Criminal Organizations

Meta's subject-matter experts are also developing systems to automatically find and remove criminal organization users globally using a range of diagnostic signals. In 2025, we actioned about 26,000 objects for violating our policies against Dangerous Organizations and Individuals by automated enforcement pipelines operated directly by our subject matter experts.

Continuous Network Detection

Preventing recidivism - the return of bad actors whose accounts have been actioned using new accounts - is a vital part of ensuring that the impact of Strategic Network Disruptions is not short-lived. Meta has developed new AI-assisted automatic counter-recidivism methods, to better predict and prevent recidivist behavior, based on an examination not only of actioned accounts but those in their networks. These new tools have seen our counter-recidivism efforts become more effective, making our enforcements stick better than ever before. Additionally, the introduction of these methods into our disruptions of illegal narcotics drug accounts - not only those directly tied to designated Criminal Organizations - means that for the first time, we can conduct continuous enforcement against those networks.

Criminal Organization Recidivists

Over many years, Meta has executed multiple Strategic Network Disruptions targeting Mexico-based cartels. This past year we spent time investigating users with network links to users we previously removed. This work enabled the discovery of hundreds of cartel-affiliated users, many involved in propaganda, human smuggling operations, coordinating violence, and drug-trafficking operations.

As a result of these new counter-recidivism methods, in 2025, Strategic Networks Disruptions against Criminal Organizations saw a significant increase in the number of accounts being removed for recidivism compared to the previous year.

Targeting Dealers and Traffickers

Preventing the sale or trafficking of illegal narcotics is a key integrity focus for Meta, and we're always iterating on our enforcement methods, seeking to improve the effectiveness and scalability of our enforcement. Since mid-2025, through a new, dedicated expert team, Meta has expanded its actor enforcement efforts against illegal narcotics dealers and traffickers.

By targeting these dealers and traffickers and enforcing on them through subject-matter expert driven actions, we can more effectively combat efforts to facilitate the trafficking or sale of illegal narcotics using our platforms.

Potential Drug Trafficking in the United States and Mexico

This approach was deployed in our investigation and disruption, discussed above, of Instagram and Facebook accounts which violated our policies against Dangerous Organizations and Individuals and High Risk Drugs by seeking to use Meta platforms to promote and coordinate the distribution of illegal narcotics across the southwestern United States and Mexico. We identified that these accounts potentially facilitated the trafficking and distribution of fentanyl, cocaine, and crystal meth to customers in the United States, and demonstrated overlap with additional high-severity harms including possible weapons trafficking and ATM fraud. This model of enforcement is also what allowed us to launch the investigation into accounts involved in the promotion of crystal methamphetamine in Mexico, mentioned earlier in this section.

Youth-focused Messaging Campaigns

Young people are frequently the victims of cartel violence and exploitation. Cartels seek to recruit minors by exploiting their economic and social needs, and use them to conduct criminal activity, including using them as look outs, 'drug mules', or even to conduct lethal violence. Youth-focused prevention efforts and educational campaigns can play an important role in protecting young people from this kind of exploitation.

Since 2024, Meta has undertaken a series of youth-focused messaging campaigns, delivering educational content and online resources to help users protect themselves from potential harm. These campaigns are designed and operated in partnership with regional experts and local NGOs, ensuring our interventions are both effective and deeply relevant to the communities we serve. For example, in Mexico, we partnered with Reinserta to deliver youth-focused messaging campaigns. In Colombia, we worked with Tiempo de Juego to tailor our interventions to local needs.

Our youth-focused messaging campaigns have since reached young people in countries around the world, delivering educational content and online resources to help users protect themselves from potential harm. These custom messaging interventions are a key part of our broader platform safety strategy, reflecting our commitment to proactive harm reduction.

In 2025, Meta launched targeted messaging campaigns in Mexico - before applying what we learned through this effort to new campaigns in Colombia, Haiti and Brazil.

Mexico Campaign

Meta's pilot campaign in Mexico was designed for youth vulnerable to recruitment by criminal organizations, aged 13-24. The campaigns were designed to inform users about platform violations, educate them on the risks of engaging with criminal organizations, and empower them to make safer choices online.

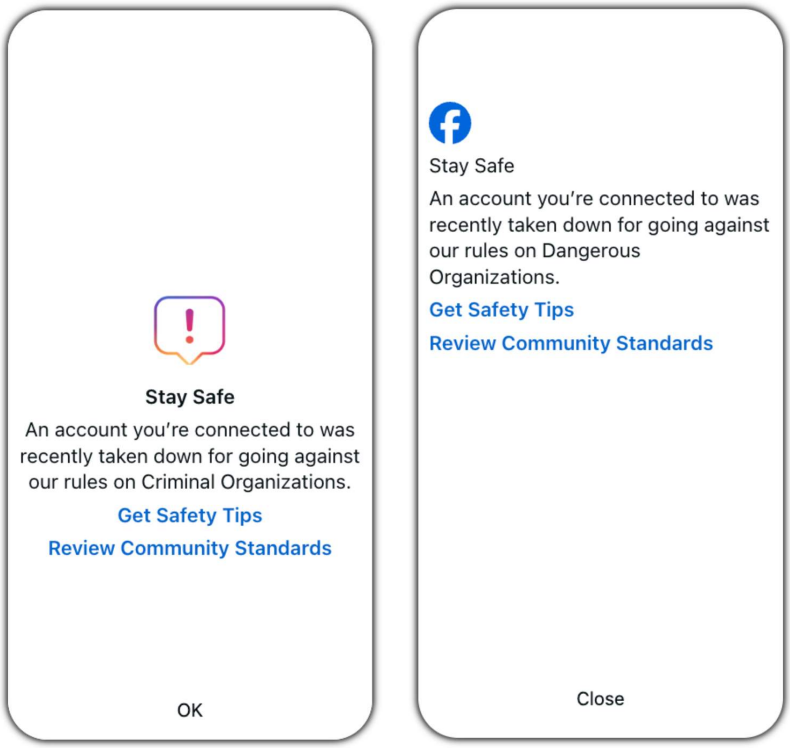

Image: Sample of a message from our youth-focused messaging campaign.

Our approach prioritized young people who, while not currently violating policies, were connected to accounts previously identified and removed for involvement with criminal organizations. Through a phased rollout, we systematically measured the campaign's impact and refined our strategy to optimize message effectiveness. The campaign resulted in a demonstrable increase in account deactivation and unfriending behavior. Account deactivation effects were most pronounced among highly connected users in Mexico. Unfriending behavior increased primarily among male users and those aged 18 or younger. Engagement rates were highest for messages delivered directly to support inboxes, which substantially boosted click-through rates.

Expansion Across LATAM

Later in 2025, Meta expanded on the success of the Mexico campaign by launching similar initiatives in Colombia, Haiti, and Brazil. This messaging campaign aimed at users aged 13-24 was designed to give resources to young users to protect themselves from recruitment and influence by dangerous organizations.

For this expansion, Meta developed new custom messaging methods, including the use of AI techniques to more effectively identify at-risk youth. This approach utilized nuanced behavioral signals to pinpoint the accounts of young people who may be more susceptible to recruitment, thereby increasing the precision of audience selection.

The custom messaging program demonstrated effectiveness in driving meaningful behavioral change and mitigating DOI risk in strategically important markets such as Colombia. The intervention produced a robust causal effect on account deactivation, and the two-round design confirmed that follow-up messaging can substantially amplify behavioral change outcomes.

Deactivation effects were concentrated among users with the highest DOI connections and prior strikes, validating the precision of our targeting model.

Meta is actively considering expanding these programs to new countries and regions around the globe, while also building out a comprehensive suite of resources to support a diverse range of at-risk youth. We continue to test and refine these interventions in countries where young people are exploited by Dangerous Organizations, leveraging data-driven insights to maximize their effectiveness. By combining Strategic Network Disruptions with targeted prevention efforts, we aim to more effectively disrupt criminal organizations' recruitment efforts.

The Future

The methods described in this section are not the only new techniques that our subject-matter experts are applying to Strategic Network Disruptions against cartels and illegal narcotics networks. We are regularly developing new AI-powered tools, enhancing our ability to detect and enforce on adversarial actors. As our adversaries seek to develop new ways of evading enforcement, Meta's investment in new tooling and methodologies allows us to counter these efforts.

04

COORDINATED INAUTHENTIC BEHAVIOR

What is Coordinated Inauthentic Behavior (CIB)

CIB involves coordinated networks of Meta assets working together to mislead people using false identities. Our enforcement of CIB violations is based on the deceptive behavior these networks engage in, and not the content they post. In each CIB case that we disrupt, network operators coordinate to use false identities in order to mislead others about who they are. When we investigate and remove these operations, we focus on behavior, not content — no matter what they post or whether they're foreign or domestic.

We monitor for previously removed networks that attempt to reestablish themselves on our Family of Apps (FOA). Using both automated and manual detection, we continuously remove accounts and Pages connected to networks we took down in the past.

In addition to creating inauthentic accounts and Pages, adversarial networks engaging in CIB may attempt to run ads to amplify their messaging. In these cases, networks typically take steps (oftentimes extensive and sophisticated steps) to hide their true identity and location. While our automated systems block many of these attempts before the ads ever run, ads associated with CIB networks we remove also come down as part of the enforcement.

Key Insights

This report details our disruption of multiple covert influence operations that violated our Coordinated Inauthentic Behavior policy. We provide a detailed case study discussing an Iranian operation targeting US and English-speaking audiences, as well as a Pakistan domestic targeting case study utilizing advanced AI-assisted persona development to advance the operation. Additionally, we share case summaries on six other covert influence operations originating from Russia, China and Iran.

Threat Indicator Sharing

In support of the global security research community, we are sharing threat indicators related to covert influence operations detailed in this report, through our dedicated GitHub repository. We hope that by sharing indicators of compromise and behavioral signatures our industry partners and the broader security research community can enhance detection and mitigation of similar adversarial activities across platforms and the internet.

Partnering to Counter CIB Networks

At Meta, countering CIB requires strong partnerships beyond our internal teams. CIB actors are adversarial, persistent, and well resourced. They often attempt to reconstitute their networks after we disrupt them and seldom limit their activities to a single platform. Since we started this work in 2017, Meta has been clear that we cannot counter these adversaries alone. By collaborating with government agencies, law enforcement, cybersecurity experts, and other technology companies, we enhance our ability to detect, investigate, and disrupt threat actors that seek to mislead our community. Sharing information and working together with these external partners is essential to understanding evolving threats and protecting the integrity of our platform, especially during critical moments like elections. Below, we summarize how we partner with different stakeholders in government, private enterprise, and civil society to counter CIB activity on our platforms and across the wider internet.

Industry Peers

CIB actors rarely confine their activities to a single platform; instead, they operate across a range of online services to maximize their reach, target diverse audiences, and reinforce their inauthentic accounts. To effectively counter these cross-platform threats, Meta actively collaborates with other technology companies by sharing investigative leads when appropriate. We mutually share information with industry peers to help both Meta and our peers independently investigate and remove violating activity from our respective platforms. For example, in the case of a Russian-origin network detailed in this report, actors leveraged AI models to make their accounts appear more authentic, using sophisticated visual branding, profile photos, and influence materials. Meta shared these findings with OpenAI, who conducted an independent investigation, identified activity from the same network, and removed the network from their platforms.

The Disinformation Research Community

We partner with academic institutions and the broader disinformation research community to deepen our understanding of threat actors and strengthen our investigative efforts. When we disrupt a CIB network, we share information about the assets we have taken down through the Influence Operations Research Archive, a secure repository designed to provide vetted external researchers with access to previously public content from disrupted networks. By enabling researchers to study these materials, we help advance independent analysis, foster transparency, and support the development of new strategies to counter online influence operations.

Government and Law Enforcement

While Meta is able to proactively detect and remove CIB networks from our platforms, our efforts are limited by our visibility into the problem as it exists on our services. Effectively addressing these threats at their root requires cooperation with law enforcement agencies and other government stakeholders. Governments possess legal authorities—such as the ability to investigate and pursue prosecution—that can significantly increase the risks and costs for those engaging in covert influence operations. Meta engages with law enforcement when appropriate, sharing leads and collaborating to disrupt malicious activity. We conduct our own independent

investigation when reviewing leads from external partners and, when appropriate, take enforcement action in line with Meta's policies.

Deep Dive: Dissecting the Kill Chain of an Early-Stage Iranian Influence Operation

Iranian threat actors continue to target the US and other English-speaking audiences via CIB operations. These networks, which use inauthentic accounts to create and distribute their content, share several key narratives, including criticism of US policies in the Middle East, as well as support for Gaza in the Israel-Gaza conflict. Today we are sharing an insight into one such operation that we detected and disrupted.

We actioned 8 Facebook accounts, 2 Pages, and 294 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 15 accounts followed one or more of these Pages, and about 41,000 accounts followed one or more of these Instagram accounts. There was no ad spend associated with this network. The network originated in Iran and primarily targeted English-speaking audiences in the United States, with a smaller, earlier component targeting Arabic-speaking audiences in Iraq. Our systems detected the deceptive activity early, allowing us to disrupt the operation before it managed to build an authentic audience on our services.

While Iranian threat actors are persistent, this specific investigation offers the opportunity to analyze the lifecycle of a CIB network caught in its infancy. Our automated systems detected this operation early in its development, allowing us to observe how the network moved through its "kill chain," from establishing infrastructure and building personas to attempting distribution, before it could successfully build an authentic audience.

As we have previously reported, Meta uses the Online Operations Kill Chain framework to analyze many sorts of malicious online operations, identify the earliest opportunities to disrupt them, and share information across investigative teams. The kill chain describes the sequence of steps threat actors go through to establish a presence across internet services, disguise their operations, engage with potential audiences, and respond to takedowns.[3]

Acquiring Assets

We found this activity as a result of an internal investigation into suspected CIB in the region. The activity targeting the US on our platforms represented an expansion of a well-established influence operation that we assess began in 2024 on X, where dozens of interconnected fictitious personas portraying themselves as US residents shared each other's content.

[3] Please see https://www.jstor.org/stable/pdf/resrep48480.7.pdf?addFooter=false for additional details about the nine phases of the Online Operations Kill Chain.

Technically sophisticated adversaries, such as Iranian CIB actors, employ advanced operational tradecraft to evade detection by our automated systems. For example, these actors can invest in clean devices and cloud or proxy IP usage to attempt to obfuscate the origin and identity of authentic operators behind these networks. Despite these efforts, our deep-dive investigation identified signals that linked the activity to individuals in Iran.

However, CIB operations are structured to advance specific strategic objectives, developing content narratives and establishing inauthentic accounts tailored to target particular audiences. By analyzing both infrastructural indicators and behavioral patterns shared among these accounts, we were able to identify coordinated activity across a network, despite the barriers posed by actors' operational security.

Disguising Assets

Once the infrastructure was established, the network moved to a "prepare" phase, investing resources into creating credible personas. The network employed a two-tiered structure consisting of “creators” and “amplifiers”.

The core of the network consisted of a subset of high-complexity accounts (the “creators”) designed to produce original content. These personas had well-developed backstories, interests, and jobs, and included an American political scientist with a PhD, a women’s rights activist, and an Albanian satirical cartoonist. To increase credibility, these accounts utilized multiple, high quality AI-generated images for profile pictures. In addition to sophisticated human personas, the network also developed convincing brand identities to make their operation appear like a local, grassroots movement. The network adhered to a brand identity by interspersing strategic content with general news or non-civic content.

Surrounding the core creators was a larger ring of lower-complexity “amplifier” personas designed to boost engagement. The amplifiers engaged with creators’ content via comments, reactions, and shares, in an attempt to make their content appear more popular.

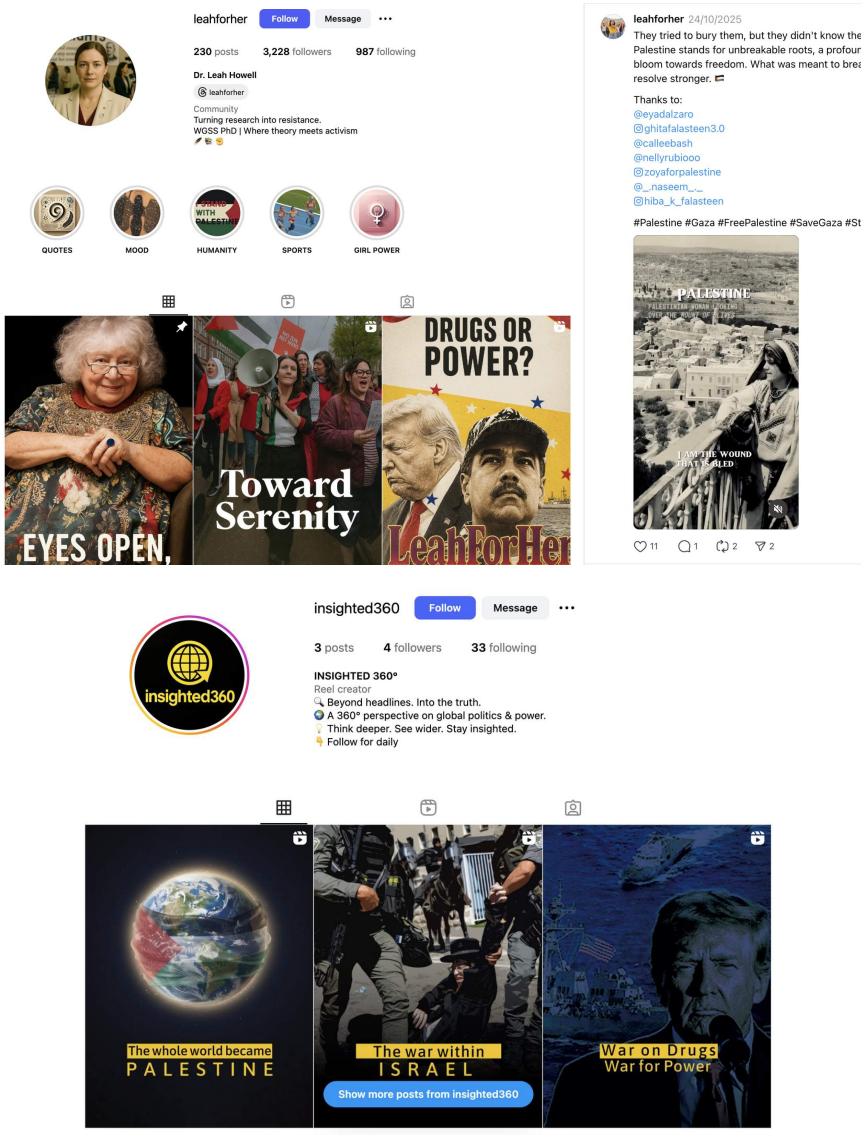

Image: Examples of a fictitious persona utilizing different profile photos of the same person on IG and Threads, using IG highlights functionality and tagging authentic users in their content; and fictitious media brand “Insighted 360.”

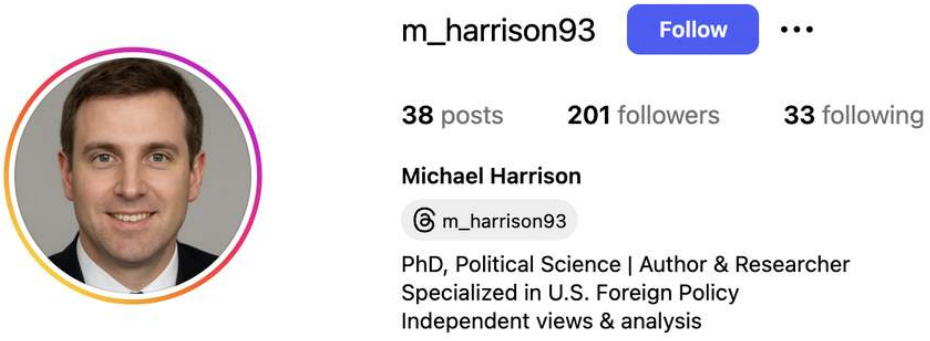

Image: Example Instagram account of a fictitious persona, using AI-generated profile photos, and its associated persona on X.

Image: Example of an amplifier that we assess to be the same influence operation on X, using AI-generated photos.

Targeted Engagement

After establishing personas, the network attempted to distribute narratives critical of Israel and US foreign policy, and supportive of Palestinians. Their distribution strategy relied heavily on co-opting authentic content.

Rather than solely relying on their own fabricated content, the "creator" accounts frequently reshared posts from authentic, high-profile pro-Palestinian voices, adding original captions and tagging the original authors. This tactic was likely an attempt to legitimize their presence and solicit reposts or engagement from real influencers.

The "amplifier" accounts were then tasked with boosting this content via likes and comments, creating a "manufactured consensus" around the posts. The network utilized the full suite of Instagram features to maximize visibility, including Stories, Highlights, and Channels, and even

expanded some personas onto Threads. Despite this multi-surface approach and sophisticated persona building, the operation failed to gain traction, achieving negligible engagement from authentic communities before our disruption.

Attempt to Enable Longevity

After we removed the network, the actors behind this activity quickly sought to reestablish their presence on Instagram. Recidivism is expected of CIB networks, which are often operated by well-resourced, persistent actors that will attempt to return to our services after enforcement. We monitor for recidivism via manual and automated measures, which in this case offered us insights on how these actors adapt to enforcement.

We observed a marked shift in the tactics behind this operation after we removed the network. Rather than investing in the elaborate persona development that characterized their initial operation, the actors pivoted to a strategy focused on speed and volume. The operators sought to create new accounts which fell into two distinct categories: one group turned its attention domestically, attempting to hijack conversations about local protests and government crackdowns within Iran, while the other consisted of dormant, US-centric personas that remained largely inactive, likely reserved for future activation.

Content quality was notably compromised in this phase. The new accounts engaged almost exclusively in commenting activity, abandoning original posting and substantial persona building. Our automated systems removed a portion of these returning accounts, and manual enforcement efforts eliminated the remainder of linked assets.

Deep Dive: Domestic Pakistani Activity Displaying Wide Use of AI

Under our CIB policy, we took down a network attributed to the Pakistani military's Inter-Services Public Relations (ISPR) wing, which we observed using novel tactics and generative AI to target domestic audiences with nationalist narratives promoting Pakistan's central government. Inauthentic networks manipulating public discourse are subject to enforcement regardless of their target audience, and the CIB policy applies equally to both domestic and foreign influence operations. Pakistan attempts influence operations on multiple fronts: state-linked actors aim to consolidate internal support while also conducting adversarial campaigns against regional rivals, as in the case of one we previously reported in April of 2019.

This case study highlights both a domestic influence operation, as well as the operation's novel use of generative AI tools. While threat actors have previously experimented with AI, this operation moved beyond simple text generation or GAN-generated profile pictures; operators leveraged multiple AI tools to research targets, craft sophisticated, multimedia content, and custom personas designed to evade detection and increase engagement.

As part of this network, we actioned 23 Facebook accounts, 8 Pages, and 14 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 64,000 accounts followed

one or more of these Pages, and about 5,000 accounts followed one or more of these Instagram accounts. The network also engaged in around $3,000 in spending for ads on Facebook and Instagram, paid for mostly in Pakistani Rupees. Our investigation into this activity began after we reviewed information shared by an industry peer.

Our investigation uncovered the central use of fake accounts, which the network used to manage its assets and distribute content. The individuals behind this activity invested significantly in developing "custom personas" tailored to their target audience. These accounts posed as fictitious individuals, journalists, and activists from Balochistan, utilizing stolen profile pictures of people in regional garb with Balochi names, who posted flattering commentary about Balochi culture, and shared imagery such as Balochi flags to embellish their inauthentic identities. These inauthentic accounts were used to administer Pages, place ads, and amplify content to give the operation the appearance of a grassroots movement of local citizens, as opposed to a military influence campaign.

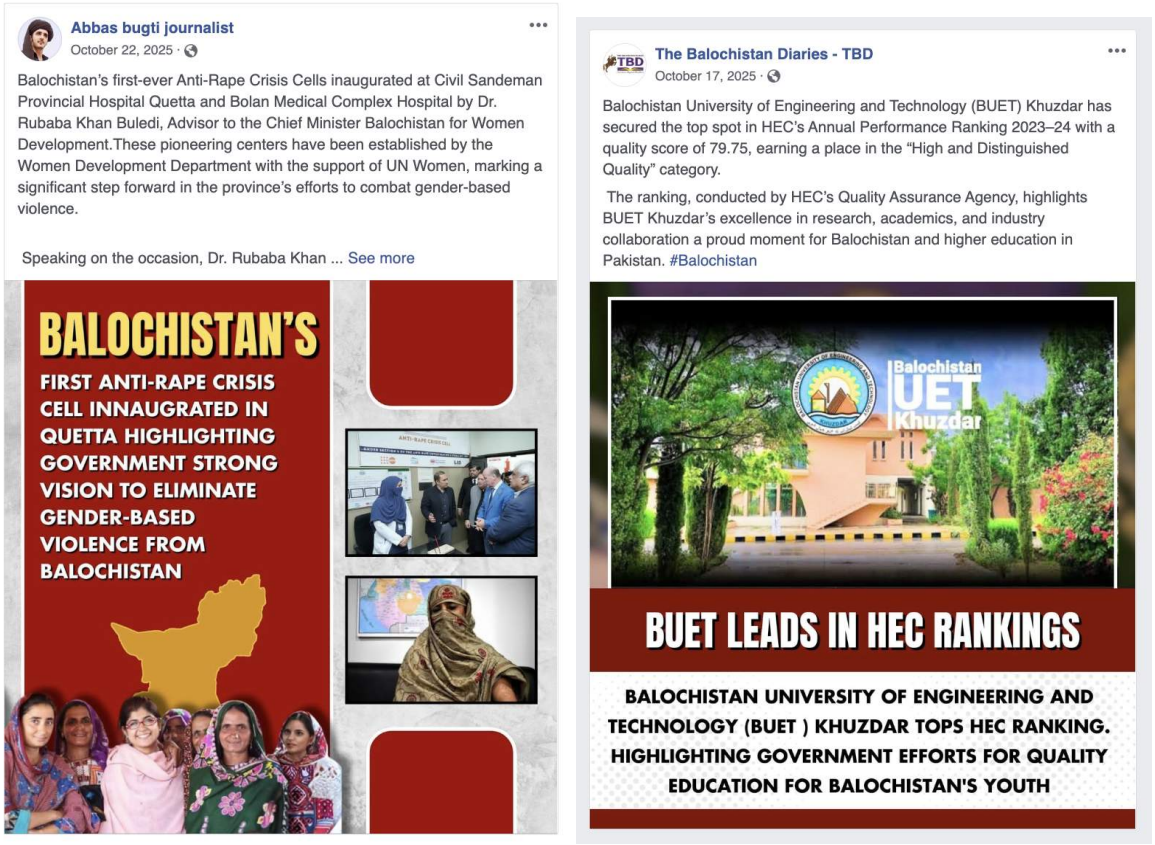

Image: Example posts from Abbas Bugti, one of the network's journalist personas, and from The Balochistan Diaries.

The network's primary objective was to promote Pakistani national sentiment. The operation anchored its narratives around established media brands created by the network, most notably a series of assets under the umbrella of a blog called "The Balochistan Diaries," which boasted a multi-platform presence. To obscure the link between these assets and the ISPR, the network

utilized a multi-layered distribution strategy, driving traffic to a standalone website, and maintaining a presence on other platforms such as X (formerly Twitter) and YouTube. To that end, this operation aimed to promote the central government, and spread nationalist messages of unity.

This network represents a notable volume and range of AI use for influence operations. In no particular order, operators leveraged:

- AI tooling in order to identify potential targets of their operation;
- LLMs to generate polished prose in multiple languages, including English and Urdu, for the their network's pages, websites, posts, and comments;
- AI video generation to create photorealistic "journalist" personas, which were featured in video interviews posted to the network's Pages, adding a veneer of professional reporting to the operation that might otherwise have required significant resources to produce;
- Used AI tools to develop and promote the website, thebalchistandiaries[.]com, which hosted writing that was likely AI-assisted. Text in posts sharing links to the website on Facebook and Instagram was also likely AI-generated. Users in this network appeared to share AI-generated comments on posts from entities both within and outside the network, praising Pakistan's government and criticizing its regional rivals in clean, consistent prose.

Other Case Studies

01 Russia

We disrupted a network originating in Russia that targeted audiences in Sub-Saharan Africa, including Angola, Ghana, Kenya, and South Africa. We actioned 37 Facebook accounts and 29 Pages for violating our policy against Coordinated Inauthentic Behavior. About 30,000 accounts followed one or more of these Pages. The network engaged in around $7,000 in spending for ads on Facebook and Instagram, paid for mostly in Euros and US Dollars. These assets posed as local, grassroots news sources. The campaign sought to influence public opinion by amplifying historical grievances against former colonial powers and other anti-Western narratives, as well as promoting Russia as a viable alternative economic and political ally.

The operation, which began in August 2025, promoted primarily authentic, off-platform news articles and social media channels that already supported narratives aligned with this campaign. In some cases, the network shared opinion articles published by legitimate news websites that carried bylines falsely attributed to real individuals or attributed to entirely fictitious personas. This tactic is consistent with external reports that Russian entities paid to place content in African outlets like Pulse Live Kenya. The network also used AI-generated content to make its accounts appear more authentic to local users, including in the Page's visual branding, such as profile photos and

advertisements promoting the Page, as well as in its influence operation materials. We shared this information with our partners at OpenAI who conducted an independent investigation into the network, leading to the removal of the network from their platforms.

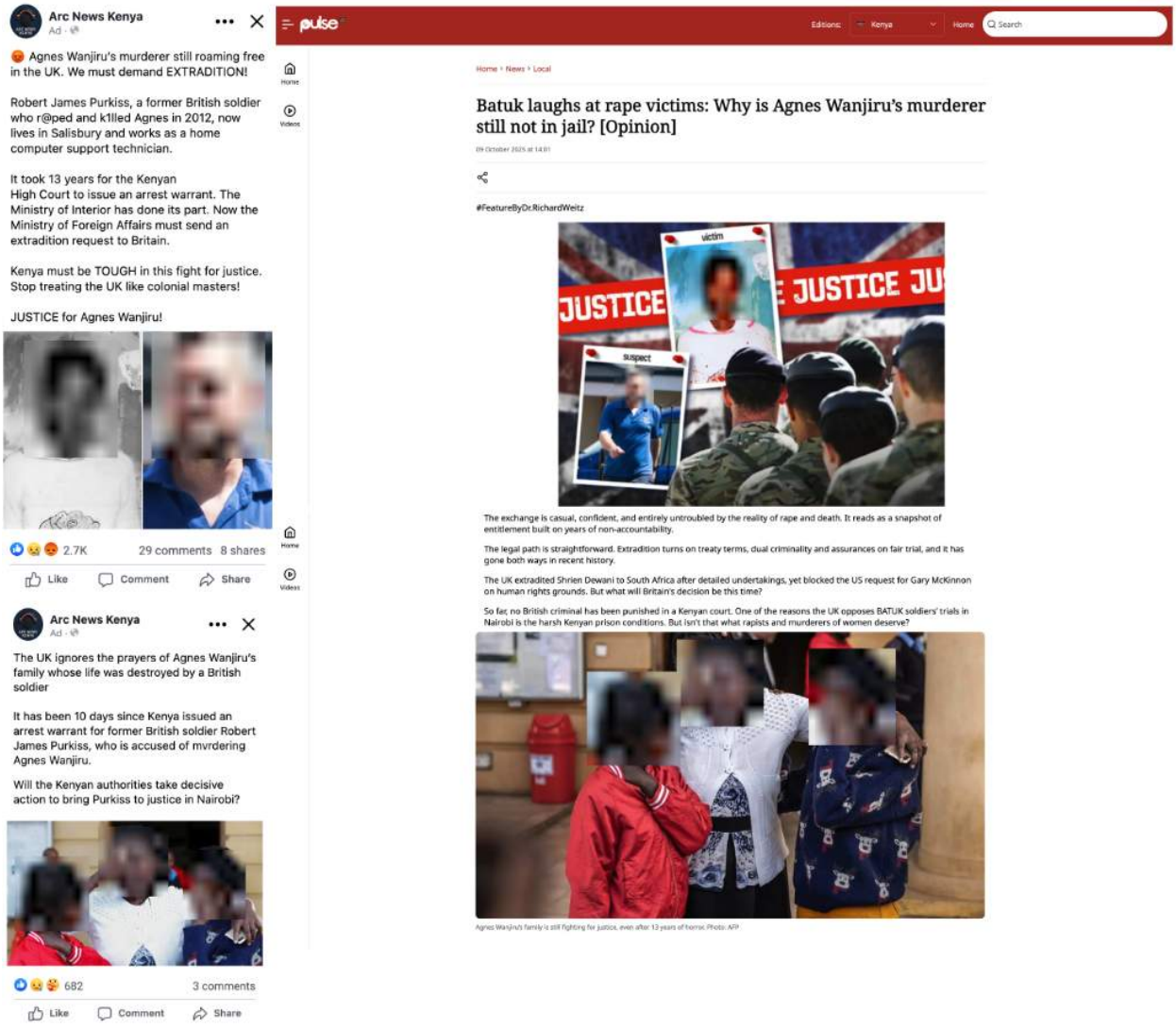

Image: Example ads run by Facebook Pages in the network and an associated Pulse Live Kenya opinion article with the byline #FeatureByDr.RichardWeitz.

02 Russia

We disrupted a network originating in Russia that targeted audiences in approximately 20 countries, including Ukraine, Moldova, Armenia, and several nations across Africa and the Middle East. We actioned 52 Facebook accounts, 134 Pages, and 35 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 238,000 accounts followed one or more of these Pages, and about 36,000 accounts followed one or more of these Instagram accounts. We attribute this activity with high confidence to Ruposters, a news aggregator based in Moscow with

a public history of servicing government-related contracts and which we previously reported on in August 2024. The network engaged in around $49,000 in spending for ads on Facebook and Instagram, paid for mostly in US Dollars, to amplify narratives designed to expand Russian influence while undermining geopolitical rivals.

This operation, which began in early 2025, was notable for its focus on short-form video content and significant organic reach across social media platforms. The network impersonated diplomatic entities like the UN, and created localized media outlets to disseminate pro-Russian messaging. To maintain persistence and evade detection, operators relied heavily on proxy IPs for location spoofing.

03 China

We disrupted a network originating in China that targeted audiences in Taiwan. We actioned 154 Facebook accounts, 23 Pages, and 1 Instagram account for violating our policy against Coordinated Inauthentic Behavior. About 93,000 accounts followed one or more of these Pages. This network promoted pro-Beijing narratives and criticized Taiwan's ruling party. The network engaged in around $15,000 in spending for ads on Facebook and Instagram, paid for mostly in Hong Kong Dollars, Chinese Yuan, Taiwan New Dollars, likely to appear more legitimate.

The network operated several pages, such as Taiwan Gossip Net and New Generation Rebellion, that claimed to be run by Taiwanese volunteers and nationals. These pages and their associated websites encouraged users to submit anonymous grievances about Taiwanese public affairs to foster domestic discord. The individuals behind this campaign used Taiwan-based proxy IPs and traditional Chinese script to make their operations appear to be of Taiwanese origin. Despite this obfuscation, our investigation uncovered that the network originated from China. The network maintained a presence on other platforms and a dedicated Android app to support its fake personas.

04 Iran

We disrupted a multi-year influence operation that targeted Azerbaijan. We actioned 367 Facebook accounts, 3 Pages, and 360 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 4,000 accounts followed one or more of these Pages and about 34,000 accounts followed one or more of these Instagram accounts. The network engaged in around $400 in spending for ads on Facebook and Instagram, paid for mostly in Euros. The network aimed to destabilize Azerbaijan politically, promote Iranian regional interests, and undermine Western influence. We attribute this network to Azerbaijani individuals located in Iran and Germany.

This operation, active since 2022, was characterized by high-quality persona building; fake accounts included detailed information such as relationship statuses, education, and specific professions. Some of the accounts in this network were designed to reflect the actual Azerbaijani diaspora in Russia, speaking Russian and claiming to live in Moscow. The network attempted to

inject narratives into the comment sections of prominent authentic news pages like the BBC and RadioLiberty.