AUGUST 2026

SECOND HALF

Adversarial Threat Report

Meta

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Meta has been publicly reporting on adversarial threats since 2017, initially focusing on coordinated influence operations, and subsequently expanding to cover other high-risk harm areas, including: espionage, surveillance-for-hire, fraud and scams, dangerous organizations and individuals, and the adversarial use of AI. The objective of our reporting remains twofold: to inform our shared understanding of the threat landscape across industry, government, and civil society, and to impose accountability on the actors behind these threats by exposing their operations in detail. This report covers investigations and actions we took primarily in the first half of 2026 across five threat areas: Fraud and Scams, Coordinated Inauthentic Behavior (CIB), Security in the Age of AI, Terrorist Organizations, and Surveillance-for-Hire.

Key Trends, and How We Respond

Several patterns cut across the threat areas in this report:

Disruption displaces threats; it rarely eliminates them. Sustained enforcement against Southeast Asian scam compounds has coincided with signs of geographic displacement rather than disappearance, as operations surface in new jurisdictions. The actors behind Russia's most persistent influence campaign, Doppelganger, diversified their activity from a single high-volume operation into a series of smaller, tactically distinct campaigns after years of defensive pressure. Surveillance-for-hire vendors sanctioned or sued in one corporate form re-emerged under new names and subsidiaries. Terrorist networks forced off one platform may reconstitute on another. In each domain, disruption raises the cost of doing business for adversaries but does not necessarily deliver a permanent end state. For this reason, we continue to look for new and promising opportunities for cross-societal collaboration that can help raise the cost of reconstitution faster than adversaries can absorb it.

The supply chains behind threats are industrializing. Influence operations outsource work to freelancers and to third-party contractors operating across borders. Surveillance-for-hire companies license capabilities to government clients who would not otherwise possess them. The actors we investigate increasingly resemble markets rather than campaigns: modular, specialized, and transactional. In this report, we describe an operation that mass-produced approximately 575,000 accounts in Pakistan that we believe were likely warehoused to later be used as pre-aged inventory. This is a trend we will continue to monitor for potential enforcement insights.

AI is both accelerant and countermeasure. Across the internet, scammers use generative AI to mass-produce profiles, fabricate imagery, and localize lures across languages and formats. Influence operators use AI-generated personas to target diasporas in their native languages. Cybercriminals use AI to generate offensive malicious code. But the same technology powers detection at scale and enables new preventive tools: an AI-powered chatbot redirecting would-be criminal recruits away from instructional material, and a scam detection tool that helps people in Southeast Asia identify potential fraud in real time. The durable advantage goes to defenders who can share the signals AI surfaces across the ecosystem.

Section Overviews

Security in the Age of AI

This section focuses on AI in the security domain. We share early observations from monitoring cyber misuse attempts against Meta AI, where AI use appears to be heavily concentrated at earlier stages in the cyber kill chain, and describe our layered defenses from model-level safety training through system-level guardrails and open-source tools. On the defensive side, AI-driven vulnerability analysis found four times more high-severity vulnerabilities in native code than traditional scanners during our software development process, helping prevent exposure.

Fraud and Scams

Financially motivated threat actors remain among the most adaptable adversaries we and the industry face. In this report, we detail policy-violating enforcement action against a coordinated network of more than 500,000 accounts and Pages using AI-generated engagement bait to source targets for gift-card fraud; approximately 575,000 predominantly dormant accounts mass-produced in Pakistan that we assess were likely warehoused as pre-aged inventory; and a network of 5,000 Instagram accounts from Uganda impersonating charities to solicit fraudulent donations. Working with the US Department of Justice, the Royal Thai Police, and cross-industry partners, we convened our third and largest coordinated disruption operation, removing more than 1.4 million assets and sharing intelligence that contributed to 63 arrests.

Coordinated Inauthentic Behavior

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. We disrupted multiple covert influence operations this half, adding to a cumulative record in which at least 135 countries have been targeted and networks have originated from 81 countries.

This section covers new activity by the actors behind Doppelganger, the most persistent Russian-origin covert influence operation we track. Although its high-volume "memes to nowhere" ad campaign continued to target France and Germany, and expanded to Hungary ahead of that country's election cycle, we now see a portfolio of offshoot efforts targeting Armenia, Moldova, and the United States. We also share case studies of other operations originating from Russia, France, Spain, Bulgaria, Iran, Israel, and Ukraine, and continue to publish threat indicators through our GitHub repository for the broader security research community.

Terrorist Organizations

We detail Strategic Network Disruptions targeting terrorist groups across multiple ideologies and geographies, paired with Custom Messaging prevention campaigns that reach younger users

connected to accounts associated with these terrorist networks. We removed approximately 9,000 assets operated by a terrorist organization that sought to re-establish a presence on our platforms, deployed LLM-based automated enforcement pipelines to combat that reconstitution, and referred credible threats to law enforcement. We disrupted terrorist networks in Pakistan (approximately 2,000 assets) and Colombia (more than 1,000 assets), deploying AI-based auto-enforcement against the latter. Following each of these disruptions, we deployed Custom Messaging campaigns co-developed with local partner civil society organizations, reaching approximately 130,000 younger users in Pakistan and 20,000 in Colombia. We also deployed our first Custom Messaging campaign in Europe, reaching 8,000 primarily German youth. Across all of our disruptions targeting dangerous organizations, image comprehension has become a key force multiplier, allowing investigators to identify violating content, and at-risk users for potential Custom Messaging campaigns an order of magnitude faster.

Surveillance-for-Hire

The surveillance-for-hire industry continues to target users of our platforms, with established vendors persisting while newer entrants bring their own tooling to market. We disrupted recidivist activity from Intellexa/Cytrox (Predator spyware); enforced against an Italy-based company, ASIGINT; took action against two China-based companies specializing in network traffic interception; and disrupted Bindecy, an Israel-based vendor operating a mobile spyware implant. Alongside enforcement details, we're sharing user-facing protections across our apps.

01

SECURITY IN THE AGE OF AI

In our previous reports, we documented an evolving relationship between artificial intelligence and adversarial threats—one defined less by dramatic inflection points than by steady, compounding adaptation on both sides. Our core assessment from 2025 remains intact: platform defenses continue to serve as a bottleneck against AI-enabled adversarial operations. Detection built on behavioral signals and technical indicators is designed to catch threat actors regardless of how convincing their AI-generated content has become. Human, technical, and organizational bottlenecks continue to constrain full automation of adversarial operations, much as they do in legitimate enterprises adopting AI.

However, the environment in which these defenses operate is shifting. AI-generated content is now a near-standard component across many of the threat types we track, from influence operations to fraud and scams to attempted cyber misuse of our own models. The actors deploying these tools are adapting—and as with any industry undergoing technological transformation, some adapt faster than others. Looking ahead, the rapid growth of agentic AI models (systems capable of planning, executing multi-step tasks, and adapting to feedback autonomously) represents a qualitative shift that our defenses must be prepared to meet. We have not yet seen these capabilities deployed extensively at scale by threat actors, but we are monitoring the ecosystem and adapting accordingly. The window in which defenders hold an advantage is real, but likely not indefinite.

This section focuses on AI in the security domain: we begin by examining how we are working to secure our own AI models against cyber misuse, then we share some early insights and analysis into cyber activity on Meta AI. Finally, we explain how AI is accelerating capabilities across our security operations, highlighting the importance of AI for cyber defense. If you are interested in learning more about how we are seeing AI used in Influence Operations and Fraud and Scams, you can find details within those sections.

Consistent with our prior reporting, this section focuses on the adversarial use of AI, rather than AI-driven incidents or unintended behavior; that said, given the relevance to this topic, we wanted to highlight our reporting on a recent issue in which a misconfiguration in a third-party

cybersecurity evaluation environment allowed a pre-release version of Muse Spark 1.1 to interact with a real website during testing. The issue has been addressed and we detail our findings in the linked post.

Addressing Cyber Misuse of Our AI Models

Integrating Cyber Misuse Mitigations Across Our AI Tech Stack

Threat actors continue to actively test AI systems as force multipliers for malicious cyber operations, from attempting to generate malicious code to attempting to exploit model capabilities for scam operations. As we continue to expand the capabilities of our AI products—from Meta AI to our developer API and AI tool use—we have simultaneously deepened our investment in security mitigations designed to prevent the misuse of these systems for cyberattacks, fraud, and other adversarial purposes. We take a layered approach to security, spanning the model level, the system level around deployed products, ongoing monitoring and evaluation, and engagement with the broader defender community. Our aim is for them to work together to reduce the risk of cyber misuse and other adversarial threats over time.

The specific combination of mitigations we employ is dependent on the model and specific deployment risks; not every mitigation mentioned here is used for every model deployment.

Model-Level Mitigations

At the foundation, we apply safety training techniques during model development to reduce the likelihood that our models produce harmful cyber-related outputs. This includes refusal training (teaching models to decline requests related to hacking, spyware, or malware) as well as alignment methods that train models to prefer secure outputs even under adversarial pressure. These training-time interventions help reduce the rate of harmful outputs, forming the first line of defense against cyber misuse at the model level.

We have also invested in techniques to protect against data poisoning, including filtering training data to exclude content from known influence operations and espionage-linked sources. This helps prevent adversary-controlled content from degrading model integrity or embedding harmful biases at the training stage—a threat vector we first documented in our Q2-Q3 2025 report.

Capability Evaluation and Red Teaming

Before new models are deployed, we conduct capability evaluations to assess whether the model could materially contribute to catastrophic cyber outcomes, including scaled network compromise, automated vulnerability exploitation, and AI-enabled fraud. These evaluations employ a combination of industry benchmarks, diverse custom internal assessments, and realistic attack simulations developed with specialized third-party security firms. We also conduct manual and automated red teaming, engaging both internal and external security researchers to probe models for emerging adversarial techniques across agentic, multi-turn, and adversarial contexts. The

results of these assessments determine the model's risk level, inform release decisions, and have been documented in Safety and Preparedness Reports published alongside model releases.

System-Level Mitigations

Around our proprietary AI products, we maintain inference-time guardrails that operate independently of the model's own training. These defenses help ensure that even if an adversary finds a way to bypass training-time safety behaviors, a separate layer still stands between the attempt and a harmful output. They can include:

- Content Classifiers: Classifiers that filter inputs and/or outputs against defined risk categories in our policies, blocking and/or flagging harmful or policy-violating model outputs.
    - Cyber Misuse Classifiers: Classifiers designed to identify and block model responses to high-risk prompts seeking exploit code, malware generation, phishing, or social engineering assistance.
- Prompt injection/alignment classifiers, such as:
    - Prompt Guard: A classifier designed to detect known prompt injection instructions and jailbreak techniques in user inputs—the types of adversarial prompts threat actors use to bypass safety training.
    - Intent Guard: A classifier designed to detect indirect prompt injection—such as goal hijacking or malicious instructions hidden in content the model accesses during inference (e.g., emails or web search results)—that may cause the model to act in ways the user did not intend.

Post-deployment, we have monitoring and enforcement systems that identify usage patterns consistent with misuse, enabling us to respond to emerging threats.

Ecosystem-Level Collaboration

We believe that securing AI against adversarial cyber misuse is a shared challenge that requires the same whole-of-society approach that has proven effective against Coordinated Inauthentic Behavior and other cross-internet threats. No single company can address this alone. Some of our ecosystem investments include:

- **Open-source defensive tools:** We have open-sourced LlamaFirewall, Llama Guard, Prompt Guard, CodeShield, CyberSecEval, and other defensive tools, making these defenses available to the broader defender community.
- **Industry collaboration:** We participate in collective efforts to evaluate and understand frontier AI capabilities and risks. We also engage in information sharing with industry and government partners across a variety of threats, since threat actors typically use a variety of models and platforms to conduct their cross-internet operations.

- **Standards development:** We engage with bodies developing AI cybersecurity standards and shared best practices—such as the Frontier Model Forum—including contributing to baseline security frameworks and guidelines that can apply across the industry.
- **Threat Intel:** We publish detailed threat reports (such as this Adversarial Threat Report series) that document real-world adversary tactics we observe across our platforms. We also release Safety and Preparedness Reports alongside frontier model launches, disclosing pre-deployment evaluation results and residual risk assessments.

These layers are designed to complement one another: model-level training helps reduce the base rate of harmful outputs, system-level guardrails are built to catch what training misses, continuous evaluation identifies emerging gaps, and ecosystem collaboration ensures defenses improve industry-wide rather than in isolation. As adversaries adapt their tactics and AI capabilities continue to advance, we will continue evolving each layer of defense and sharing what we learn with the defender community to strengthen our collective resilience against adversarial cyber misuse.

Early Insights from Cyber Misuse Monitoring on Meta AI

As we expand the AI products and capabilities we offer, we are investing in systems to identify and address potential misuse of our models for cyber-related purposes. One component of this ongoing work is developing cyber misuse classifiers, mentioned above. Because many of the same capabilities that enable misuse also support legitimate security research, our systems must carefully balance precision and recall, minimizing false positives that would hinder legitimate use while maintaining broad enough coverage to catch genuine malicious intent.

In the interest of advancing shared understanding of AI-enabled threats, we are sharing some early observations from Meta AI traffic. This analysis is grounded in established cyber threat intelligence frameworks and prior research on offensive tool taxonomies and as a result does not represent the full spectrum of potential misuse attack vectors in the wild. As with any early-stage threat intelligence effort, what we observe is shaped by where and how we look.

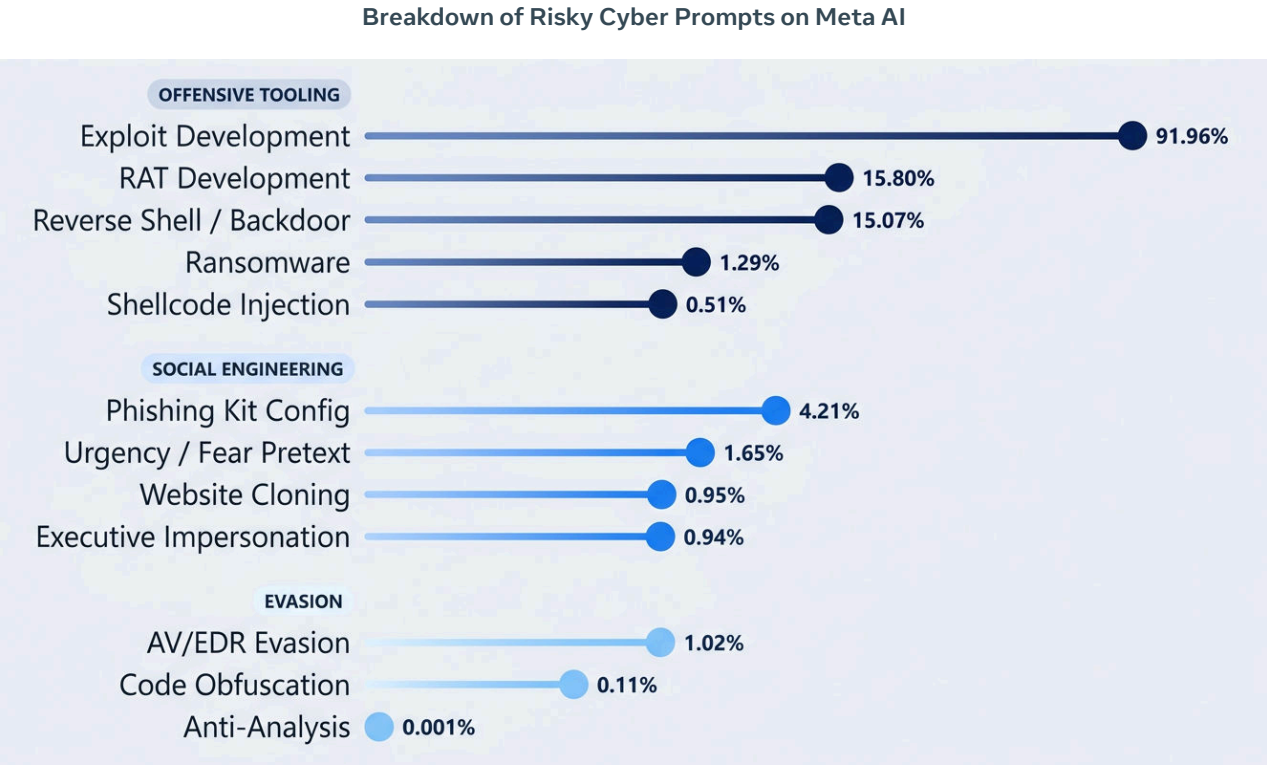


Chart: Breakdown of prompts flagged by our cyber misuse classifier, grouped by threat category. Each bar shows the percentage of all flagged prompts that triggered that particular indicator. Because a single prompt can trigger multiple indicators (e.g., a request for a RAT that also involves shellcode), percentages do not sum to 100%.

Breakdown of Risky Cyber Prompts

We grouped some of the most commonly occurring threat indicators our classifier is designed to track into three broad categories: offensive tooling (requests for exploit code, remote access trojans, reverse shells, ransomware, and similar), social engineering (phishing kit configuration, impersonation pretexts, website cloning, urgency-based lures), and evasion (techniques to bypass antivirus, evade analysis, or obfuscate malicious code). We then analyzed activity across these categories over the period of approximately one month.

The distribution reveals a pattern consistent with a predominantly **pre-operational threat actor population**. Exploit development—the initial capability-building step—appeared in nearly 92% of all flagged prompts, dwarfing all other categories. Requests for remote access trojans and reverse shell code each appeared in roughly 15%. The remaining categories—ransomware, shellcode injection, social engineering techniques, and evasion—each appeared in fewer than 5% of flagged prompts.

Several aspects of this distribution are notable. First, requests to enable social engineering appeared at far lower rates than offensive code requests, an inversion of traditional threat landscapes where social engineering is the dominant initial access vector. Second, evasion techniques appeared in just 1% of flagged prompts, suggesting that few actors are attempting to

use our models at the deployment stage of the attack chain; most are still in the capability-development phase. Third, ransomware-specific requests were notably low (1.3%).

Taken together, these patterns are consistent with the enforcement findings described below: AI appears to be expanding the threat actor population by lowering the minimum viable skill threshold for developing offensive capabilities.

Insights from Actor-Level Cyber Misuse Enforcement

Because many of the same tools and techniques serve both offensive and defensive purposes, we allow educational and defensive uses of our AI models. However, we may enforce when we identify particularly persistent or adversarial attempts to get our models to produce violating cyber outputs.

In a recent enforcement against cyber misuse of our models, we saw actors attempting to use generative AI across several stages of the cyber attack chain, from reconnaissance and tooling to social engineering and operational deployment. The most common pattern involved actors attempting to coerce our models into generating functional offensive tools: phishing pages designed to steal Instagram or Facebook credentials, working remote access trojans with command-and-control infrastructure, and Metasploit payloads with step-by-step deployment instructions. We continue to iterate on our training methodology and classifiers designed to refuse this kind of activity.

Many actors employed elaborate jailbreak frameworks, assigning the model fictional personas like “DarkForge-X” or “OMEGA-DIOS.” These actors showed high levels of persistence, cycling through multiple bypass techniques in a single session when initial attempts were refused. The technical sophistication of the actors themselves was often low; several could not achieve basic tasks like rooting an Android device, and appeared to be relying on the model to close gaps in their own capabilities rather than to scale existing expertise. This is consistent with the broader pattern we observe in this space: AI is expanding the threat actor population by lowering the minimum viable skill threshold, not merely accelerating actors who already have advanced capabilities.

Two patterns in this cluster are especially significant because they indicate actors moving beyond exploratory misuse into operational preparation. We observed actors applying social engineering techniques against the model itself, framing requests around school cybersecurity competitions or wrapping malware requests inside educational research pretexts. We also saw actors using our models not to generate attacks from scratch, but to refine and improve operations they had already begun: one user iteratively workshopped fake Meta security alerts through the model, requesting edits to remove emojis for a more professional appearance and to add specific device and location details that would increase victim believability. Another submitted hundreds of lines of a functioning RAT and used the model as a technical reference to understand and operate code they already possessed.

These cases suggest that for some actors, generative AI may serve primarily as on-demand operational support rather than a source of novel capabilities. In practice, this means

troubleshooting, translation, and refinement of existing operations that previously required either higher skill or coordination with other humans. As our model capabilities advance and AI is more deeply integrated into workflows, we expect the potential for cyber misuse will grow in scale and sophistication. Continued investment in monitoring, detection and investigation is essential to stay ahead of this threat landscape that is constantly evolving.

AI for Defenders

Public discussion of AI and adversarial threats tends to focus on how threat actors are using this technology to improve their operations. These conversations are important, but the other side of the equation matters just as much: how AI is changing what defenders can do. Right now, defenders have an opportunity to get ahead of adversaries, but that window depends on cross-societal collaboration and equipping the defender community with the right tools and knowledge.

At Meta, how we build and secure our products has changed substantially. More and more of the code across our platforms is written by or with AI. We have adapted our security posture accordingly, rebuilding how we identify, assess, and remediate threats with AI at the center. This spans vulnerability discovery, intrusion detection, identity and access management, and incident response.

Threat actors we track—state-sponsored espionage networks, criminal scam syndicates, surveillance-for-hire vendors—now face defenders operating at a speed and scope that was not possible through manual effort. We share these developments as concrete proof points for what AI-enabled security looks like at enterprise scale. You can find other examples of AI-powered defenses throughout this report, including scam detection chatbots and search interventions.

AI for Security: Securing Code at Machine Speed

As frontier AI models lower the cost for adversaries to find software flaws at scale, we launched a coordinated program designed to identify and remediate those same flaws first. This work spans code we write and ship, the open-source libraries we depend on, and our cloud infrastructure. We deployed AI-driven vulnerability analysis across our highest-risk attack surfaces, including thousands of open-source libraries in critical infrastructure and hundreds of Meta-published open-source projects, surfacing thousands of potential vulnerabilities. In April 2026, our AI-driven code analysis tools identified four times more high-severity vulnerabilities in native code than traditional scanners during our software development process.

From May to July of this year, we ran AI-powered reviews across large portions of our deployed code to identify vulnerabilities, with AI generating the fixes for over 40% of the highest priority vulnerability remediations. Where it once took weeks to move from discovery to fix, our mean exposure time for critical vulnerabilities is now measured in days.

These are the categories of vulnerability that surveillance-for-hire vendors and state-sponsored espionage actors seek to exploit. Finding and fixing them before adversaries can weaponize them raises the cost of exploitation for attackers and shrinks their window of opportunity.

Detection, Response, and Access Management

We have applied similar approaches across detection and incident response. SEVMate, integrated into our incident management system, automates root cause analysis by scanning recent code changes and ranking the most likely culprits, reducing mitigation time substantially. Issues that previously took weeks to resolve can now be addressed in hours.

In parallel, we developed and deployed HODOR, a detection agent purpose-built for emerging threats like prompt injection against AI systems, in under one month, compared to a traditional timeline exceeding six months. This speed of deployment matters: defenders need to build detection capabilities at the same pace new threats emerge.

Maintaining the Advantage

Ensuring that defenders can maintain an advantage over attackers requires sustained investment. As frontier models become more accessible, the gap between attackers and defenders will likely narrow. The window in which defenders can build and mature AI-enabled capabilities ahead of widespread adversarial adoption will not stay open indefinitely.

We continue to share our findings, tools, and benchmarks with the broader community. Our open-source releases of defensive models and frameworks reflect our commitment to ensuring the defender community's collective capabilities scale alongside the evolving threat environment.

Looking Ahead

Platform defenses remain the critical bottleneck against AI-enabled adversarial operations. Across influence operations, fraud, and attempted cyber misuse of our models, behavioral detection, coordination analysis, and network-level enforcement continue to identify and disrupt threats — regardless of how sophisticated the AI-generated content layer has become. The human and organizational constraints we document throughout this report provide defenders a structural advantage: threat actors cannot automate what requires real-world infrastructure, institutional knowledge, and operational coordination that AI alone cannot replicate.

That advantage is real but not permanent. As agentic AI models mature and become more accessible, the equilibrium between offense and defense will face new pressures. We are investing accordingly in layered model safeguards, in AI-powered detection and enforcement, and in the open-source tools and cross-industry collaboration that strengthen the defender community as a whole. The trajectory of this technology demands that defenders move faster than adversaries adopt it, and we are committed to maintaining that pace.

02

FRAUD AND SCAMS

Introduction

Financially motivated threat actors continue to represent some of the most adaptive and persistent threats facing the online ecosystem. The case studies that follow highlight how these adversarial actors are adapting, and how they have invested time and resources in attempts to avoid the signals that usually expose them. We see operators posting their own anti-scam "warnings" to earn the trust of the people they were about to defraud. Fake animal-rescue charities turned strangers' generosity into payments that most donors never realized they had lost. And when threat actors created more than half a million accounts designed to look legitimate, a novel linguistic signal exposed the operation before those accounts could engage. Meeting adversaries like these means surfacing the harm they work to hide, and that is the focus of this report.

No single defender sees the whole of a scam. Scammers operate in the gray between platforms, sectors, and borders, and use AI to move faster and appear more credible than ever before. We introduced the Fraud Attack Chain in our Q3 2025 Adversarial Threat Report to map that lifecycle end to end. We use it again here to show where each disruption lands, and where we rely on partners upstream and downstream to both see and disrupt what we cannot.

Countering these threats requires disrupting scam networks on our own platforms, and sharing signals and intelligence across the defender community—extending our reach upstream, to the infrastructure and efforts that precede an attack, and pressuring the scammers downstream, collaborating with the financial sector and with support from government partners.

Scams pose a problem that no platform, bank, or telecommunications provider can solve alone. A 2026 survey conducted by Gallup, co-funded by a cross-sector coalition that included Meta, JPMorganChase, and AT&T, found that people in the US encounter scams across every channel they use: scams via phone calls and text messages are as common as those on social media, email, and messaging apps.

Meta continues to invest in building defenses across its platforms as well as supporting a shared defense in which platforms, banks, telecommunications providers, and governments exchange

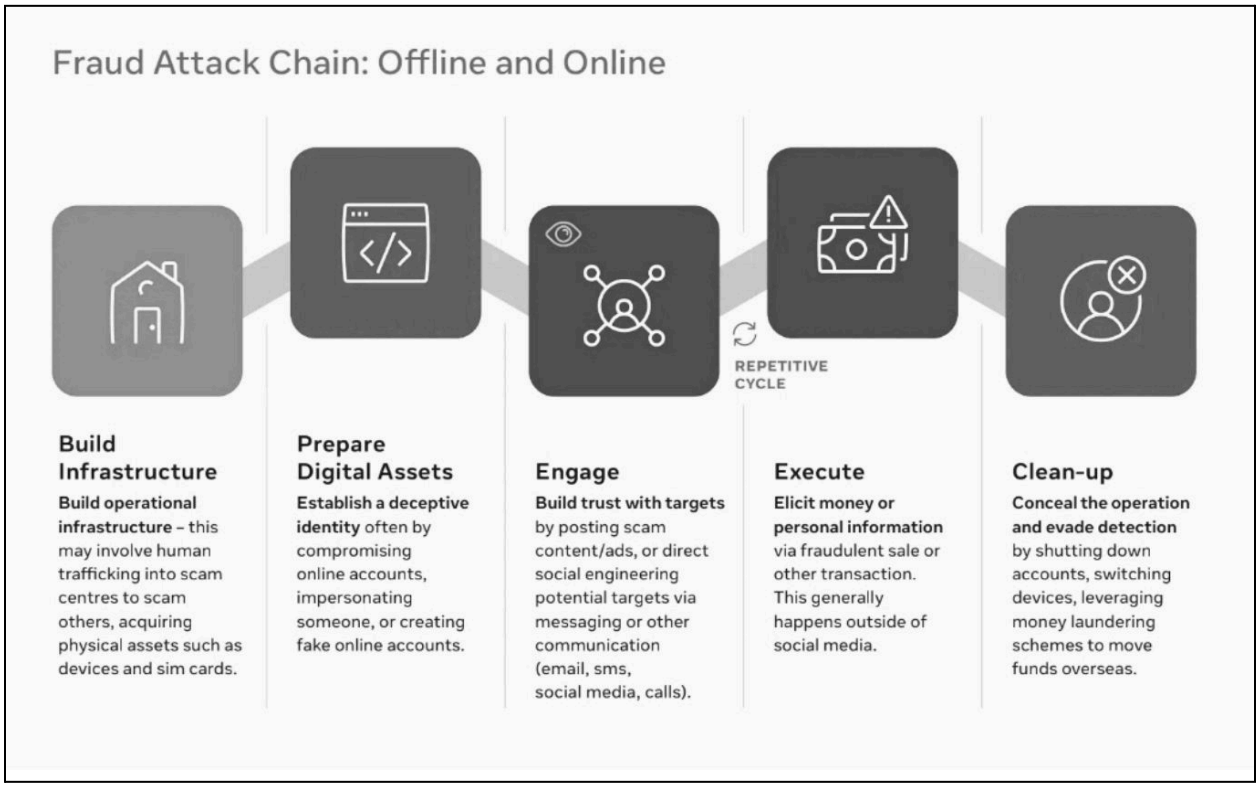
signals and support their respective investigations and enforcement actions. That principle underpins our participation in the Industry Accord Against Online Scams and Fraud, the ASEAN Public Private Partnership Principles for Tackling Online Scams, the Fraud Intelligence Reciprocal Exchange (FIRE), and the Global Signal Exchange (GSE), as well as the cross-industry operations and partnerships described in this chapter.

Consistent with previous reporting, the cases that follow illustrate how adversaries are adapting their tactics, techniques, and procedures, and how a whole-of-society response is adapting with them. We share them so that other platforms, researchers, financial institutions, and governments can detect and defend against the same behaviors in their own environments.

The Fraud Attack Chain: A Refresher

The Fraud Attack Chain describes the five stages of a financially-motivated scam: **Build Infrastructure** (organizing people, devices, and payment instruments); **Prepare Digital Assets** (creating, aging, or compromising the accounts and domains that make an operation look real); **Engage** (making contact and building trust, often before moving a target across platforms); **Execute** (extracting money, credentials, or data); and **Clean Up** (laundering proceeds, recycling infrastructure and closing accounts for the next cycle).

The framework's value is in collaboration. By mapping adversary behavior to the chain, defenders can see which intervention, by whom, best disrupts each stage. Acting earlier in the chain raises adversary costs and reduces harm to people. We use the chain below to frame how these operations work, then detail a set of case studies from the past half.



The Fraud Attack Chain maps the lifecycle of a scam operation so defenders can identify where each intervention lands.

Throughout, we note where each disruption sits in the chain and how the defender community can operate upstream and downstream—extending our reach beyond what we can see alone.

Cross-Industry Signal Sharing

In our last report, we highlighted the importance of bilateral information sharing between members of the defender community through Meta's Fraud Intelligence Reciprocal Exchange (FIRE) program. The FIRE program allows us to collaborate with stakeholders who have visibility across the stages of the Fraud Attack Chain, enabling more thorough disruption of fraud operations across platforms. As part of this program, we liaise directly with industry partners through our ThreatExchange platform and with industry and government partners through the Global Signal Exchange (GSE). Across these platforms, Meta exchanges information with more than 95 partners from a variety of industries around the world, directly leading to the removal of hundreds of thousands of violating assets and scam URLs.

Threat Evolution

The trends we highlight this half are not unique to our platforms. Scammers operate across every surface and channel, as the Gallup research demonstrates, which means shared indicators surface across the ecosystem too. We share what we are seeing for that reason: tracking these common patterns and indicators is what allows partners to disrupt networks across the ecosystem and build on one another's intelligence.

AI-Enabled Scam Operations

Over the past six months, we have seen scammers use AI at every stage of an operation: to mass-produce profiles, to fabricate imagery and to generate and localize lures across languages and formats. That same automation now extends to upstream infrastructure: domain generation algorithms that produce brand-mimicking lookalike domains at scale are increasingly AI-assisted, outpacing traditional signature-based defenses. That being said, the same technology remains central to our defense. It powers detection and enforcement at scale and helps create new preventive tools developed in-house and delivered with partners at speed. AI lowers the cost of attack everywhere at once, so the most durable countermeasures are the ones defenders can share: a single detection signal or scam URL surfaced by one platform can protect people on others.

We wanted to share a few insights we have identified across the AI-Enabled Scams space this half:

Deepfakes and Celebrity Bait as a Dominant AI-Enabled Fraud Vector

AI-generated deepfake videos impersonating public figures have become a primary tool for scam operators promoting fraudulent investment schemes. These operations target government officials, finance ministers, bank executives, and media personalities across multiple markets. Our detection systems contributed to a reduction of over 80% in user reports of the most frequently

impersonated celebrities. As these defenses have improved, we have observed a clear evolution in targeting: operators have moved beyond impersonating global celebrities and now increasingly pose as local professionals and trusted community figures, including doctors, military personnel, humanitarian workers, and financial advisors, whose identities are harder for victims to verify independently. We have invested in more advanced AI-powered detection systems to counter this activity.

AI-Generated Engagement Bait for Target Recruitment at Scale

We have seen increasing use of mass-produced AI imagery at the top of a scam funnel, where operators flood Groups with AI-generated content designed to provoke engagement and then route responders off-platform to complete a financial scam. We disrupted two large networks built on this model: the first paired AI-generated religious imagery with manipulative prompts to surface receptive users, who it appears were then moved to off-platform chat groups, and websites to solicit "donations"; the second used AI-generated imagery in a fake giveaway scheme where users who engaged were told they had "won" and asked to purchase gift cards to claim their prize (see Case Studies below). We removed these networks at scale.

The Fraud Attack Chain Is Still Not Fully Automated

Despite these tactical advances, our investigations confirm that AI has not automated the fraud attack chain end-to-end. Bottlenecks remain at key stages, particularly interaction with real-world financial systems: setting up convincing financial entities capable of accepting funds, managing money movement to maintain victim trust, and laundering proceeds through complex financial networks. Scaling infrastructure online also remains constrained, given that many established platforms have longstanding experience detecting automation and scripting.

Geographic Dispersion of Scam Operations

Sustained pressure on Southeast Asian scam compounds, the focus of extensive action detailed in our reporting last half, has coincided with signs of displacement rather than disappearance. Public reporting and law enforcement action this year point to activity surfacing in new locations. Authorities in Sri Lanka arrested more than a thousand foreign nationals tied to cybercrime. Joint operations involving law enforcement in the United Arab Emirates and Thailand raided multiple scam centers and seized hundreds of millions of dollars. Investigators have documented alleged expansion into further jurisdictions. In the networks we disrupted directly we continue to see significant footprints in West Africa and South Asia.

In May, we co-hosted a third Joint Disruption Week alongside the US Department of Justice's Scam Center Strike Force and the FBI and, for the first time, brought cross-industry partners into this coordinated operation, which we describe in more detail below. Working with the Royal Thai Police and law enforcement partners from Australia, Canada, Indonesia, Japan, Malaysia, New Zealand, the Philippines, South Korea, Sri Lanka, and the United Kingdom, this event contributed to

the arrest of 63 people connected to scam centers, and continues to put a real face on the machinery behind these sites.

Targeting of Specific Communities

The operations we disrupted this half evidence how scammers may seek out specific communities. Below we share archetypes of scammers that concentrate on: unemployed younger people seeking work, donors moved by humanitarian crises and older adults looking to engage with government services. Targeting specific communities is also where partnership pays off most directly. Our partnership with the National Elder Fraud Coordination Center (NEFCC), for example, is aimed squarely at protecting older adults.

Case Studies: Disrupting Scam Networks

AI-Generated Engagement Bait

We removed a network of more than 500,000 accounts and Pages for violating our Fraud and Scam Policies. This network used engagement bait on Facebook to source targets for gift-card fraud, targeting people predominantly in the United States.

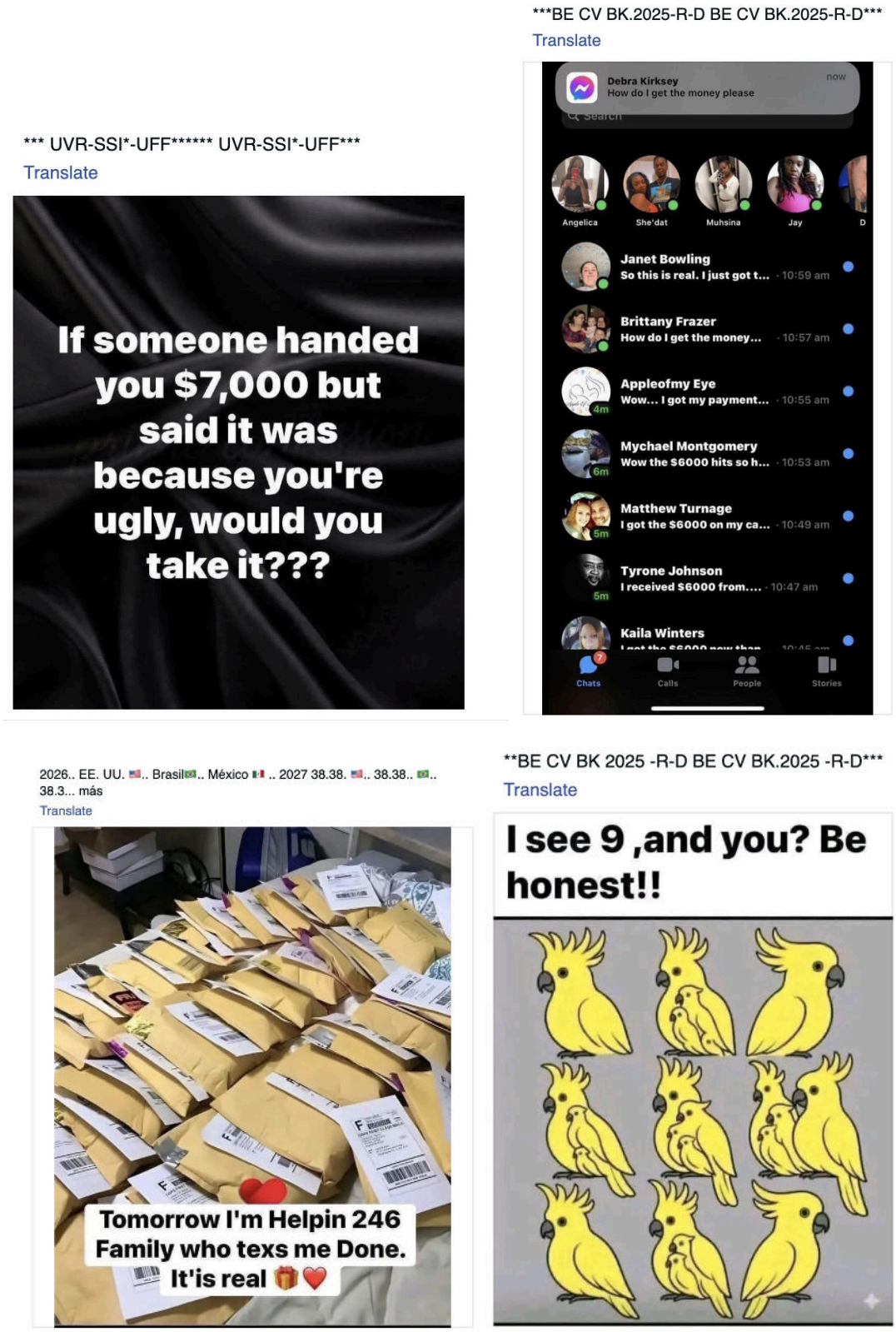
The network operated at scale across large, often distinct Facebook Groups. Operators generated or manipulated the content with AI-generated images, celeb "bait," and emotive religious or political posts designed to attract clicks and comments in the public Groups, where the vast majority of the posts were concentrated.

Two features distinguished this network. Operators embedded coded strings in their posts that we believe functioned as campaign-tracking markers, coordinating automatically generated content across the network—a practice that mirrors the delivery-tracking used in legitimate marketing. Second, the operation ran on shared infrastructure, with a small number of devices connected to a disproportionately large number of accounts. Although the accounts presented as being spread across many countries, we determined that the core of the network was linked to individuals in Nigeria.

The scam cycle was as follows: a user who engaged with a bait post was contacted by the network and told they had won a giveaway prize. They were then directed off-platform and asked to provide personal details, before being told of a fee to release the prize. Users often read the initial bait as generic or unwanted content rather than as a scam, so they may never report it — a reminder that platforms cannot rely on user reports as a single deterministic signal.

When we look at the Fraud Attack Chain itself, ‘Engagement Bait’ sits in the gray zone between ‘engage’ and ‘prepare digital assets’. To make determinations at these stages, defenders rely on behavioral signals: coordination patterns, device clustering, and the downstream funnel to off-platform payments. This is why the framework emphasizes acting across the full lifecycle rather than at any single stage.

Following our investigation we removed the network's accounts and their associated pages for violations of Meta’s Terms of Service.



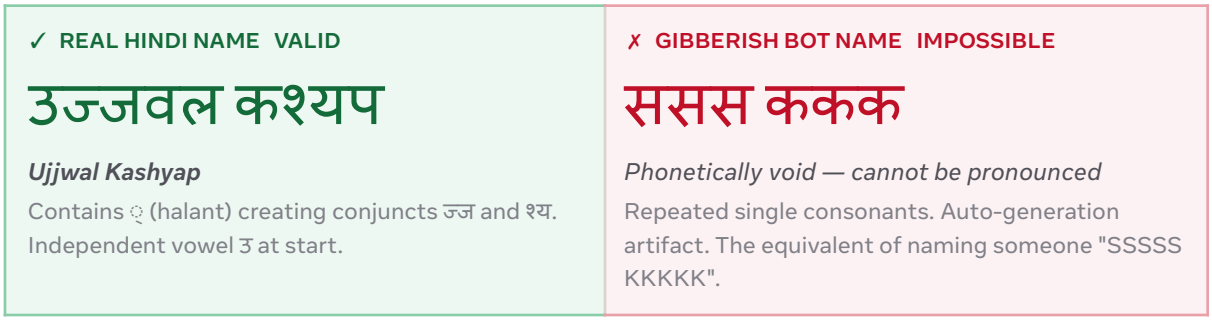
Images: Examples of AI-generated engagement bait used by the network: operators posted manipulative content in Facebook Groups, embedding coded strings (visible in the captions)

Industrial Identity Creation

We disabled more than 575,000 Facebook accounts for violating our Fraud and Scam policies. This network was created predominantly in Pakistan and was mass-produced as dormant scam inventory.

Dormant inventory presents a distinctive detection challenge: accounts that generate no or very limited content or connections produce none of the behavioral signals that conventional classifiers rely on. That is why structural and linguistic approaches represent a critical complement to behavioral detection.

In this specific case, the accounts' names appear to have been generated programmatically, without regard for the conventions of the language—in this case Hindi—producing combinations a native speaker would be very unlikely to use. Several signals surfaced this network, including indicators present at registration that set these accounts apart from the patterns we typically see in organic account creation.



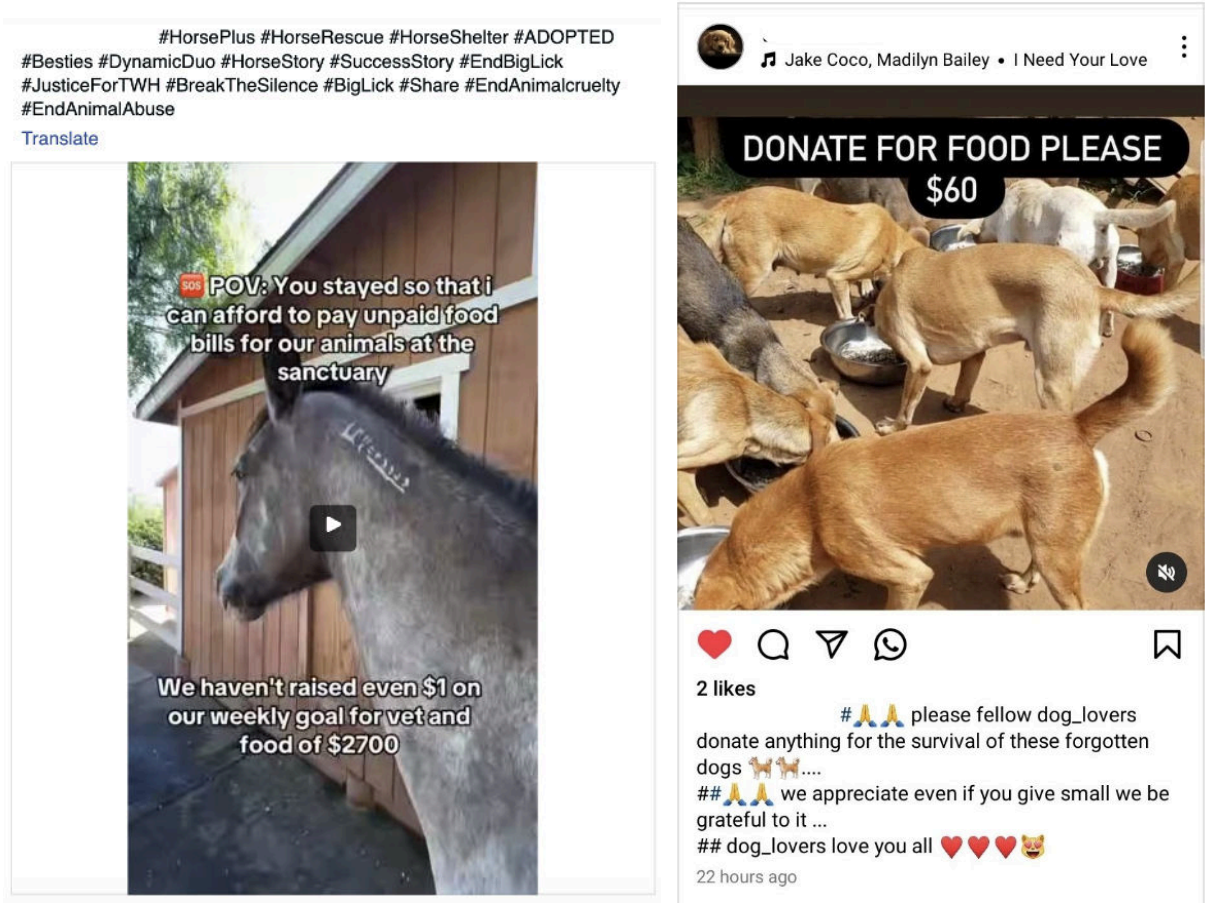
The operation is best understood not as a scam but as a scam supply/enabler operation involving a threat actor scripting the creation of accounts in bursts, then warehousing them, attempting to build the appearance of age and legitimacy. That pattern is most consistent with a division of labor in which one set of actors builds inventory and a separate set deploys it, though we cannot rule out vertical integration where the same operator does both.

Charity and Crisis-Donation Fraud

We disabled a network of approximately 5,000 Instagram accounts for violating our Fraud and Scam Policies. This network operated primarily from Uganda and impersonated charities and animal-rescue organizations to solicit fraudulent donations from users across the United States, United Kingdom, Australia, Canada, and the European Union.

The operation was not a set of isolated accounts but part of a persistent, recidivist ecosystem — one recently documented by BBC Africa Eye. Coordinated operators run parallel fundraising campaigns for the same fabricated causes, sometimes featuring the very same animals or footage across many accounts at once.

Several tactics made the network unusually resilient. Rather than depending on engagement or other activity such as follows or direct messages, they gained distribution by posting emotionally distressing videos, allowing content to organically spread while generating little of the connection-based signal enforcement typically relies on. The network collected 'donations' through off-platform sites, a single link in each account's bio that directed users to consumer payment services. The network also used real-world crises, including active humanitarian emergencies, to add urgency to its appeals.



Images: Examples of Instagram posts impersonating animal-rescue charities.

Images: Examples of Instagram posts impersonating humanitarian organizations.

This archetype highlights a specific risk that platforms face: because donors give without expecting anything in return, many never realize they have been defrauded—so users might not report it. Detection therefore has to rest on more than one signal.

Following this investigation, we enforced against the accounts for violations of Meta's policies and stood up continuous detection for this activity tuned on a combination of donation solicitation, off-platform payment links, and abusive coordination signals.

Domain Typosquatting

We identified and blocked more than 10,000 impersonation domains used to defraud buyers and sellers in Taiwan through lookalike parcel-delivery brand pages.

As the Fraud Attack Chain illustrates, before a scam reaches a person, it needs infrastructure, and domains are among the most valuable, reusable, and vulnerable assets an operation owns. Working from partner and platform signals, we disrupted a long-running effort to impersonate well-known convenience stores through lookalike domains used in attempts to defraud buyers and sellers in Taiwan.

The operators relied on rapid domain-cycling—where they registered slightly altered variants (a technique known as Typosquatting) as quickly as older ones were blocked and paired the domains with otherwise innocuous pages that built an audience, soliciting follows and page interactions following the posting of content that offered the latest deals. Once the target was engaged, the scam link was then shared to victims through private messages.

We identified newly registered domains from external registration signals, matching them against patterns associated with the impersonated brand before they were ever used against people on our platforms. Tracing the domains through content surfaced thousands of additional linked assets and operators across Threads, Instagram, and Facebook that were promoting the same scam, extending the disruption well beyond the domains themselves.

We know that fraudulent networks extend beyond Meta platforms, so when we block a scam-linked domain we share it with our signal sharing peers. We share blocked URLs via the Global Signal Exchange and the FIRE program so that a domain surfaced by one defender can be shared with many. That shared intelligence is what turns a single takedown into an ecosystem-wide block.

Investment Scams Funneling Users Off-Platform

We disabled over 2,000 Facebook and Instagram accounts for violating our policies on Fraud and Scams. This network attempted to move targets to third-party websites and direct messaging channels, where scammers could extort them privately.

This network primarily originated in Ghana and Nigeria, targeting members of the West African diaspora. The network operated across multiple fraud verticals including promoting inauthentic investment platforms, impersonating legitimate money transfer services, advertising a fake cryptocurrency, and impersonating religious figures to solicit donations. Using these tactics, the network funneled targets to other domains where we don't have visibility or the ability to moderate.

We discovered this network by investigating fraud-linked email addresses we received from the National Elder Fraud Coordination Center (NEFCC). The majority of activity on our platforms was preparatory rather than directly fraudulent, however the additional context provided by NEFCC allowed us to disrupt the network during the "engage" phase of the Fraud Attack Chain. This case illustrates the value of cross-sector signal sharing, where external partners can provide actionable leads that enable us to identify and disrupt networks even when the fraudulent activity itself occurs off-platform. By disrupting on-platform accounts, we interrupted the network as they attempted to engage targets, before the fraud itself could occur.

Government Benefits Scam Targeting Older Adults

We disabled a network of over 10,000 assets on Facebook for violating our policies on Fraud and Scams. This network targeted older adults seeking government benefits, primarily in the United States.

This network ran ads discussing topics designed to appeal to older adults, primarily government programs and home improvement rebates (i.e., Medicare, ACA benefits, Section 8 extra benefits, gutter replacements). These ads targeted users 50 and older, driving them to third-party scam websites. We discovered the breadth of this network by investigating an initial set of ad accounts that we received through the FIRE program.

This case demonstrates how signals received from partners through the FIRE program can seed investigations that uncover significantly larger interconnected fraud ecosystems operating across our platforms. This case also demonstrates the cyclical nature of signal sharing efforts: we received the initial intelligence from a FIRE partner, investigated and disrupted the network, blocked the fraudulent third-party domains from our platforms, and then shared these domains back with our signal sharing partners. We continue to work alongside our FIRE partners and the broader defender community to identify and disrupt fraud networks targeting our users and the broader internet.

Joint Disruption Week

Working with the US Department of Justice, the Royal Thai Police, and cross-industry partners, we removed more than 1.4 million accounts, Pages, and Groups for violating our policies on Fraud and Scams and shared intelligence that contributed to 63 arrests.

We have reported previously on the coordinated "disruption week" model—windows in which we coordinated on platform action, with real on-the-ground actions leading to both virtual and physical dismantling of these criminal networks, including arrests. In May, we helped convene the third such operation and the largest. Conducted as part of the US Department of Justice's Scam Center Strike Force, it was also the first to bring cross-industry partners into direct operational coordination.

Before and during the operation, we removed more than 1.4 million accounts, Pages, and Groups across Facebook and Instagram, and shared investigative intelligence that each partner could act on. Coinbase froze more than $3 million in cryptocurrency tied to the networks. Microsoft suspended roughly 20,000 accounts. SpaceX terminated thousands of Starlink kits providing the satellite connectivity the compounds relied on. The Royal Thai Police made 63 arrests and identified new scam center locations for follow-up. Each partner did something the others could not.

These criminal enterprises are neither victimless nor bloodless. The compounds behind them, concentrated in Myanmar, Cambodia, and Laos, are enabled by human trafficking and forced labor: real people lured with false job offers and made to run scams under threat of violence. These networks depend on a stack of services that no single company controls: accounts to reach targets, payment rails to move money, connectivity to keep the compounds online, and physical premises to house them. That is why disruption bites hardest when partners act on their own layer at the same time, and it is the model we intend to keep building.

Enforcement disrupts operations already in motion. The interventions below target the Chain's earliest stages—Prepare Digital Assets and Engage—intercepting the moment someone seeks recruitment into a criminal network, or the moment a target is about to hand over money or personal information to one.

AI-Powered Prevention

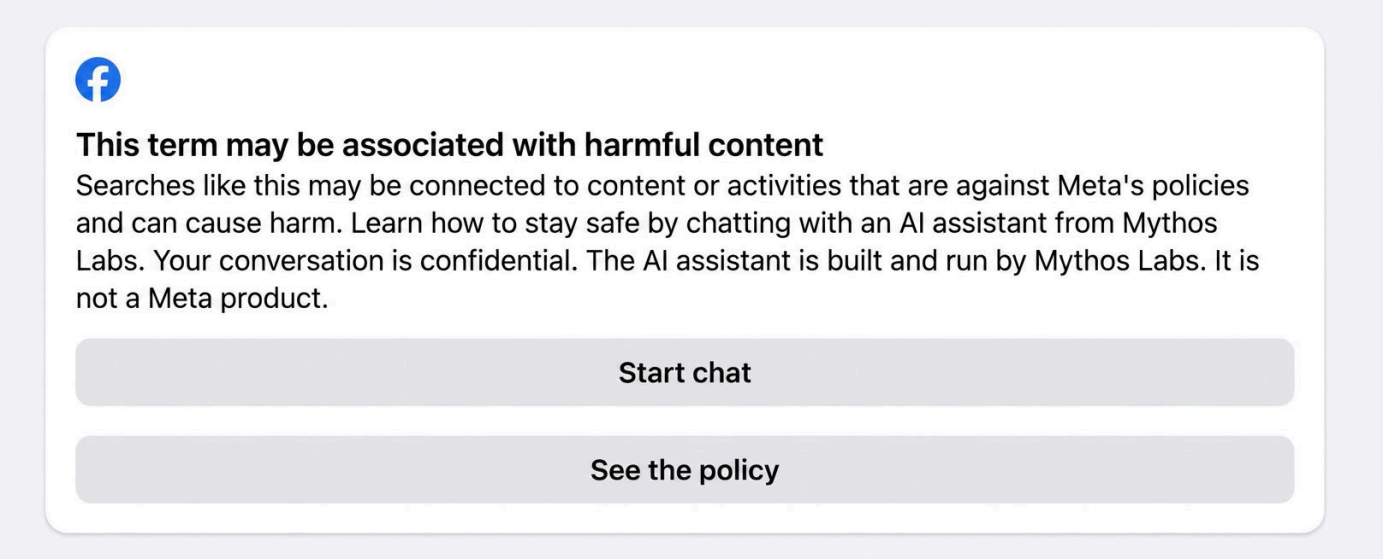
Enforcement removes harm that has begun; prevention aims to stop it before it starts. Over the past half we advanced two preventive interventions, both built and operated by partners on open-source foundations we provide. They represent a defensive application of the same AI capabilities that adversaries are using: where scammers deploy generative models to mass-produce lures and localize them across languages, Meta and its partners use the same technology to equip people to recognize those attacks at the point of contact. Both interventions are built on Llama and operated independently by partners; Meta's role is to provide the model foundation, the platform surface, and the distribution.

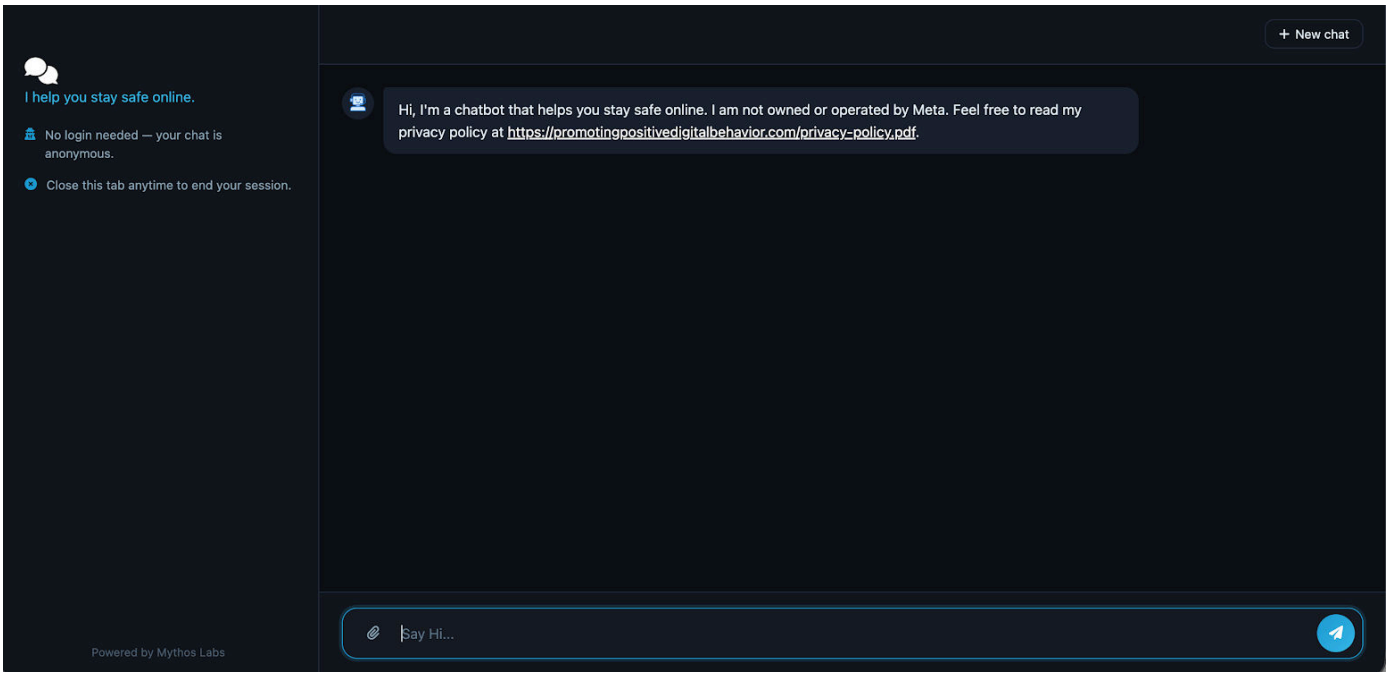
Disrupting Criminal Recruitment: Yahoo Boys Search Redirect

We deployed an AI-powered search intervention that is designed to redirect users searching for Yahoo Boys recruitment material to an educational chatbot, blocking access to violating content and offering an interactive alternative rather than a blank results page.

The Yahoo Boys are a criminal network, designated by Meta as a Dangerous Organization, operating primarily from Nigeria with growing reach across West Africa. The network is responsible for large-scale financial fraud, romance scams, and financial sextortion. Unlike traditional criminal organizations with rigid hierarchies, the Yahoo Boys function through shared culture, coded language, and instructional manuals referred to as "formats" that provide step-by-step guidance for conducting various crimes. Since we designated the Yahoo Boys as a Dangerous Organization in April 2024, we have enforced against associated content across our platforms.

A distinctive feature of the network is how it recruits. People trying to join the network actively search for instructional resources using in-group terminology that unambiguously signals intent to join or replicate the network's criminal tactics. This search behavior represents a critical intervention point: it is the moment a user is actively seeking to connect with the network, before they have accessed violating content or made contact with an operator. In the Fraud Attack Chain, this sits at the boundary of **Prepare Digital Assets** and **Engage**: the user is actively seeking instructional material, and this intervention targets the search action itself, blocking access to violating content and offering an educational alternative.





The intervention is designed to work as follows: when a user searches for a violating keyword associated with the Yahoo Boys on Facebook, Instagram, or Threads, search results are blocked entirely and the user is shown an inform module explaining that the term is associated with harmful content. The module offers the option to connect with a conversational AI chatbot for educational guidance. If they engage, we direct them to an interactive experience, built on Llama and operated independently by Mythos Labs, that responds dynamically, answers follow-up questions, and sustains engagement over multiple turns. The chatbot has been subject to extensive adversarial testing: it has been trained not to provide operational details or generate content usable for malicious purposes, and to maintain consistent safety standards across multiple languages, including Nigerian Pidgin, Hausa, Yoruba, and Igbo.

This is the first time we have used a conversational AI chatbot as the endpoint of a search redirect. Traditional redirects point to static resources (a webpage, a helpline number, a policy document) which most people click past without reading. A conversational alternative can sustain the kind of interaction that actually shifts behavior. The redirect is available globally but is predominantly encountered by users in Nigeria, where the majority of associated search activity originates. We will continue to iterate on both the scope of keywords covered and the chatbot's capabilities as these criminal networks adapt their language.

The intervention reflects a principle that underpins the rest of this chapter: enforcement alone is not enough. Blocking violating search results removes the supply; redirecting users to educational engagement reduces the demand.

Helping People Recognize Potential Scams in Real Time: Sago

Working with the United Nations Office on Drugs and Crime, the International Justice Mission, and the US Department of State, we supported the deployment of Sago—a free, AI-powered scam-detection chatbot built and operated by Mythos Labs, helping users identify potential scams before engaging with them.

Criminal scam syndicates in Southeast Asia conduct fraud at industrial scale and actively traffic people into compounds where they are forced to perpetrate it. Fraudulent job advertisements promising high-paying remote work or cryptocurrency roles lure victims to locations where they are detained and coerced into running romance, investment, and other scams under threat of violence. The United Nations Office on Drugs and Crime has identified the region as the global hub of these operations, with victims trafficked from over 50 countries. Addressing this threat requires interventions that operate at scale, reach users in local languages, and adapt to rapidly evolving recruitment tactics. In the Chain, these scams reach people at the **Engage** stage: the target has received a lure but has not yet handed over money or information. An intervention that helps a user recognize a scam at that moment disrupts the operation before it reaches **Execute**.

Sago allows anyone who encounters a suspicious message, job offer, or social media post to forward it via WhatsApp or Messenger for immediate analysis. The chatbot, trained on more than one million real-world scams from 35 countries, assesses whether the content is likely a scam, flags the specific risks, and guides the user on how to disengage and report safely. It communicates in more than 40 languages, including Bahasa Indonesia, Javanese, Thai, and Burmese.

Since December of 2025, Sago has been live in Southeast Asia reaching at-risk audiences through targeted ad campaigns on Instagram and Facebook.

Early data from the tool confirms patterns visible across the threat landscape more widely. Employment-related scams (fraudulent job offers, the primary vector for trafficking recruitment) were the most common type users surfaced, followed by investment scams, romance scams, and phishing. Sago helped users identify potential scams they might otherwise have engaged with, and directed people who had already been victimized to local government resources and reporting

channels. The tool's repeat-use rate also appears to have been high, suggesting that scam exposure is not a one-time event for most targets.

Beyond its user-facing function, Sago generates aggregated, non-identifying intelligence through regular Insight Reports, which Mythos Labs shares with local authorities and civil society organizations in Southeast Asia. These reports surface emerging scam typologies, geographic patterns, and signals linked to active operations, creating a feedback loop in which individual interactions with the tool strengthen the ecosystem's ability to detect and respond to scam recruitment. Sago is developed and operated entirely by Mythos Labs; no user data is shared with Meta.

Conclusion

The scam threat did not recede this half so much as it shifted focus, with new sites emerging in some countries and declining in others, all as a result of enforcement pressure being exerted on the syndicates behind them. The threat also shifted technologically, as adversaries reached for AI to build, disguise, and scale what they do. Our response moved with it from the infrastructure scams are built on, to the accounts they prepare, the people they engage, and the proceeds they try to extract.

Three patterns run through the cases above.

First, scam operations increasingly resemble supply chains, not campaigns. Actors are now specializing—one creating accounts, another deploying them; one registering domains, another building audiences on them. That specialization makes the ecosystem more resilient but also more visible: each handoff is a seam we can target, and each actor we remove degrades capability for the others. **Second**, the signals that matter most are often behavioral rather than content-based. Dormant accounts produce no content to classify. Engagement bait looks indistinguishable from ordinary posts. Fake Charity pages solicit donations without ever posting a link that a URL scanner would flag. In each case, it was coordination patterns, structural anomalies, or partner intelligence—not content analysis alone—that exposed the operation. **Third**, no single defender holds enough of the picture to act as effectively alone. The domains we blocked were registered through a handful of registrars who can deny them at the source. The proceeds moved through financial institutions that now hold signals we have shared. The compounds behind these operations sit in jurisdictions where only governments can act, and the people trapped inside them can only be freed by law enforcement willing to enter.

We will continue to share intelligence and to build the mechanisms that make collective action faster and more decisive, because the threats described in this chapter demand nothing less from the defender community.

03

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 are foreign or domestic.

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, any ads associated with CIB networks we remove also come down as part of CIB enforcement.

We monitor efforts to come back by networks we previously removed. Using both automated and manual detection, we remove accounts and Pages connected to networks we took down in the past.

Since we began publicly reporting on CIB disruptions, we have disrupted hundreds of Coordinated Inauthentic Behavior operations worldwide. Our investigations demonstrate a highly global threat landscape; since our public reporting began, at least 135 different countries have been targeted by at least one CIB network. The United States continues to be the most targeted country globally, followed by Ukraine and France. The CIB networks we have disrupted originated from 81 different countries, with Russia remaining the most frequent geographic source of covert influence operations (53 networks), followed by Iran (41 networks), and China (14 networks).

Key Insights

This report details our disruption of multiple covert influence operations that violated our Coordinated Inauthentic Behavior policy. We examine the evolution of campaigns linked to the actors behind Russia's most persistent covert influence operation—Doppelganger. We report on

activity by the Social Design Agency (SDA) and the growing role of the state-linked Autonomous Non-Profit Organization (ANO) “Dialog” in foreign-targeted influence operations. SDA and ANO's tactics have diversified: Doppelganger remains active, but it now sits within a broader portfolio of distinct campaigns.

Additionally, we share case summaries on nine other covert influence operations originating from Russia, France, Iran, Israel, and Ukraine that violated our Coordinated Inauthentic Behavior policy.

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 that our industry partners and the broader security research community can enhance detection and mitigation of similar adversarial activities across platforms and the internet.

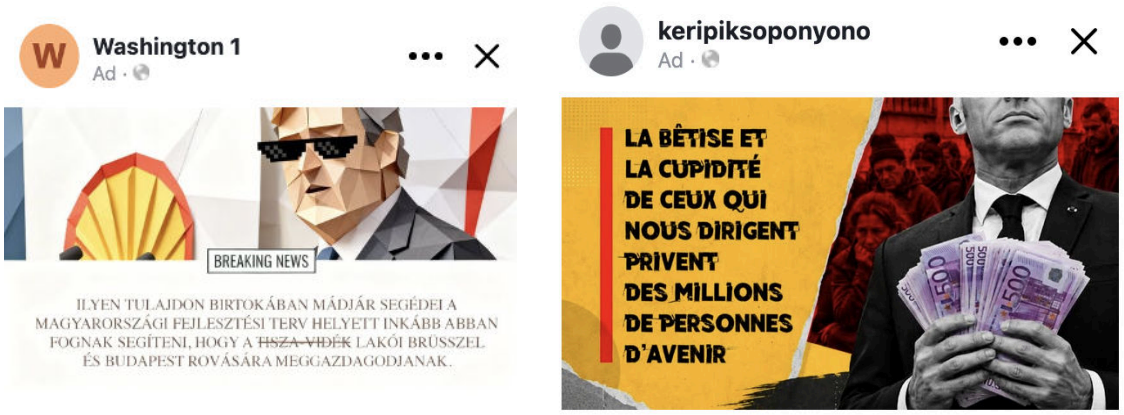
Deep Dive: Beyond Brute Force—Doppelganger’s Changing Operations

We continue to detect and enforce on Russia’s most persistent covert influence campaign, Doppelganger, which we first reported in 2022. Earlier this year, we observed a sudden cessation in Doppelganger activity on our platforms. A month later, Doppelganger returned to our services using the same brute-force tactics characteristic of this operation since late 2024. However, the actors primarily associated with Doppelganger—including ANO Dialog and Social Design Agency (SDA)—sustained these brute-force tactics while simultaneously standing up a number of tactically distinct offshoot campaigns. These campaigns feature more fluid targeting and a more diverse array of creatives than the original Doppelganger network. Throughout, we have continued to refine our automated detection based on the behaviors we observe and to engineer campaign-specific defenses to help block these operators from returning to our platforms. In the sections below, we break down various operations under the Doppelganger umbrella that we have observed since our last update.

Brute Force Tactics Fluctuate But Persist

By late 2024, Doppelganger appeared on our platforms as a high-volume, low-quality operation that prioritizes sheer output over persona building. It runs ads with minimal text overlaid on images with no links to off-platform websites, builds no audience, and primarily targets Germany, France, and Israel. Utilizing a "brute force" approach across the internet, the network relies on a vast, constantly regenerating infrastructure of spoofed websites and attempts to seed its content across as many social media platforms and services as possible. Technically sophisticated adversaries, such as Doppelganger, employ advanced operational tradecraft and inauthentic accounts to evade detection by our automated systems and obfuscate their identities. However, as we have previously reported, Doppelganger’s attempts to run ads are quickly detected and blocked by our automated systems, typically before anyone sees them.

However, in late January 2026, the campaign's daily activity suddenly ceased for more than a month. The operation re-emerged in March with the same tactics, but this time targeting a new country, Hungary, ahead of its election. The narratives shared by the new offshoot were critical of EU-Hungary relations and suggested the Tisza party would undermine Hungarian interests through shadowy deals with Brussels. Our proactive defenses detected this new, Hungary-targeting activity before the election and we removed the violating assets from our platforms. After this effort, in late May, the operation resumed its typical targeting of audiences in France and Germany, but notably dropped Israel from its targeting.



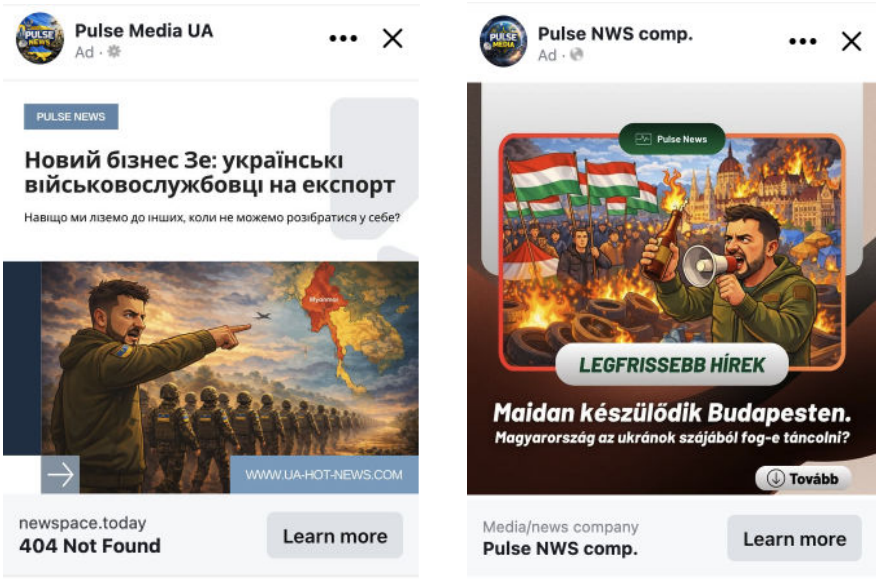
Images: Examples of Doppelganger activity. On the left: [non-literal machine translation] “With assets like these, Magyar's people won't bother with Hungary's development plan—instead, they'll help Tisza's inner circle get rich on EU and Hungarian taxpayer money.” On the right: [machine translation] “The stupidity and greed of those who rule us deprive millions of people of a future”.

Social Design Agency-Linked Offshoot Campaigns

Following the changes to Doppelganger's traditional operating mode, we observed several campaigns linked to the Social Design Agency (SDA), a Russian IT and public relations company to whom we publicly attributed to Doppelganger in 2022. In this reporting period, we removed multiple clusters of inauthentic activity linked to SDA actors under our protocols to enforce on CIB recidivism.

An example of an SDA campaign we removed was a network of 10 Facebook accounts, 7 Pages, and 4 Instagram accounts targeting the EU and Ukraine, which we removed for violating our policy against Coordinated Inauthentic Behavior. About 1,300 accounts followed one or more of these Pages. The operators created fake accounts to amplify content from news brands—“Pulse NWS Media”, “Pulse Media UA”, for example—and actor-controlled domains eu-hot-news[.]com and ua-hot-news[.]com, that branded themselves as local media outlets while under the control of Russian actors. Advertisements directing audiences to these domains used hallmarks of earlier Doppelganger efforts, such as complex redirects—in this instance, using newspace[.]today—and geofencing, tactics we observed and reported on in 2023 and 2024. The operation distributed

tailored narratives focused on Ukraine war fatigue, the rising cost of living, and delegitimizing President Zelenskyy.



Images: Examples of ads imitating Doppelganger tactics. On the right: [non-literal machine translation] “BREAKING NEWS: A Maidan-style uprising is being planned in Budapest. Will Hungary be made to dance to Ukraine's tune?”. On the left: [non-literal machine translation] “Zelenskyy's new business: exporting Ukrainian soldiers. Why do we meddle in other countries' affairs when we can't get our own house in order?”.

ANO Dialog’s Growing Role

We have observed the Autonomous Non-Profit Organization (ANO) “Dialog” take a more active role in foreign-targeted covert influence campaigns. ANO Dialog is a state-linked organization founded in 2019 by the Moscow city government. The organization has previously been linked to the Doppelganger campaign by the US Department of Justice, which attributed early Doppelganger infrastructure to ANO Dialog alongside the Social Design Agency (SDA) and Structura National Technology. We have also identified operational overlaps—most notably, both the Doppelganger campaign and a recent ANO Dialog network have outsourced work to the same contractor. We also observed tactical similarities, such as the common use of URL redirect techniques.

Below, we provide summaries of two covert influence campaigns we have attributed to ANO Dialog:

We removed 25 Facebook accounts, 13 Pages, and 13 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 9,600 accounts followed one or more of these Pages and about 10,400 accounts followed one or more of these Instagram accounts. The network targeted the June 2026 Armenian parliamentary elections attempting to boost the "Strong Armenia" party and Russian-Armenian billionaire Samvel Karapetyan. Apart from promoting their

page in Armenia directly, this network also targeted the Armenian diaspora in Germany, France, and the US. To lend credibility to their operation, networks linked to this campaign repeatedly boosted authentic news articles aligned with their strategic objectives. To disseminate its messaging, the operation maintained a cross-platform presence, operating on our platforms as well as on TikTok and YouTube. Furthermore, the network redirected users to authentic media articles to lend credibility to its assets, and created distinct local media brands such as “Ar Media” to distribute its content.

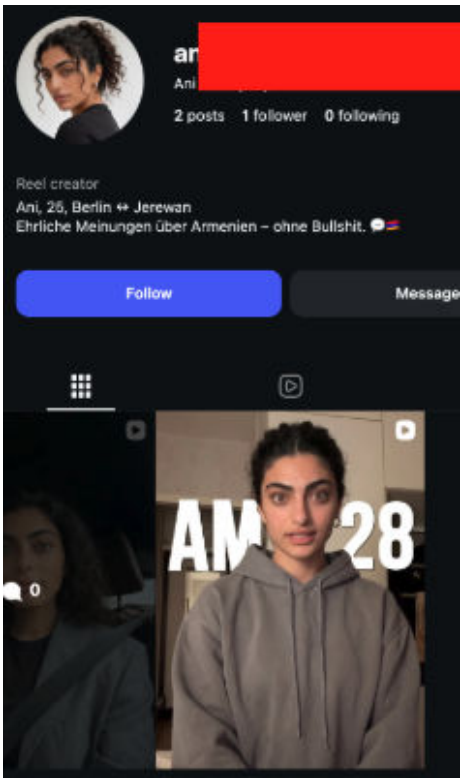
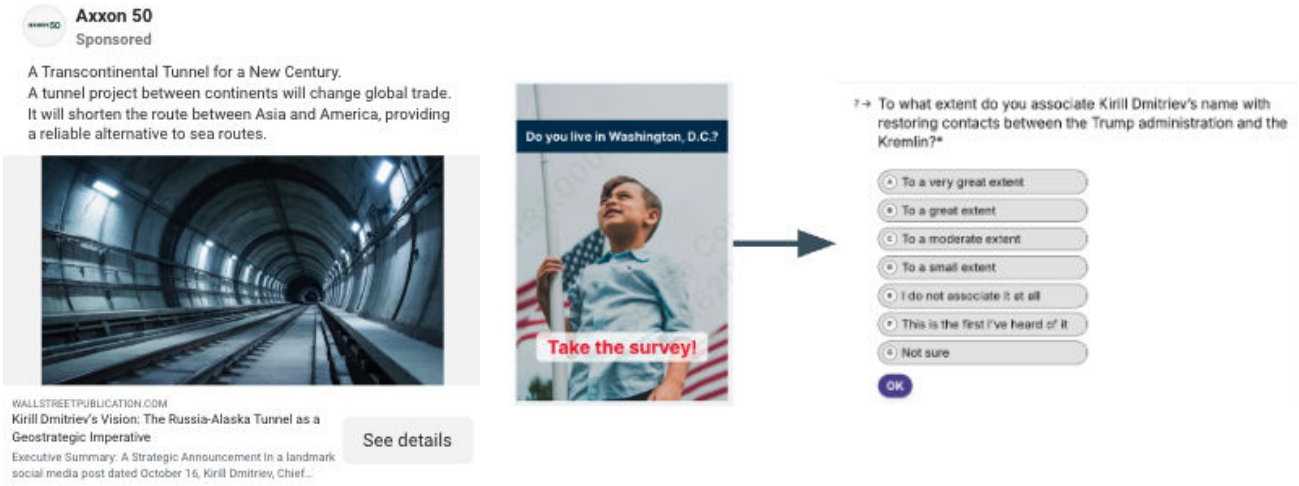


Image: Example of a likely AI-generated persona targeting the Armenian diaspora in Germany. Translation of Instagram bio: “Ani, 25, Berlin ↔ Jerewan Honest opinions about Armenia - without the bullshit. 💬🇦🇲”

We removed 69 Facebook accounts, 58 Pages, 3 Groups, and 14 Instagram accounts targeting Moldova, the US, Ukraine, and several African nations. About 8,000 accounts followed one or more of these Pages and about 900 accounts followed one or more of these Instagram accounts. Coordinated by ANO Dialog alongside the Russian PR firm ООО «ЮС-МЕДИА» (us-media[.]ru), this campaign closely integrated public opinion polling and classic influence campaigns. The surveys often used emotionally charged, leading questions to manipulate sentiment, and the resulting "data" was then disseminated via ads to make their narratives appear as trustworthy statistics. The fictitious “vocepentru / rvm” research institute, which was used to disguise public opinion surveys targeting Moldova, was first highlighted in a report by GLOBSEC, a global think tank headquartered in Slovakia. The campaign also boosted authentic news articles to advance its goals; in the US, the network seeded positive coverage of Russian presidential representative Kirill Dmitriev into

financially motivated "pay-to-play" news sites and promoted authentic Reuters and Politico articles about him. The operation at the same time distributed surveys to measure whether Americans viewed Dmitriev positively.



Images: Examples of influence campaigns targeting the United States. On the left: Article promoting Kirill Dmitriev was seeded by this network on the pay-to-play "Wall Street Publication" website and then promoted via ads. On the right: Opinion survey distributed at the same time as the influence campaign.

Countering Doppelganger's Evolving Threat

We have remained vigilant through Doppelganger's evolution from a single, centralized operation into a series of smaller, distinct campaigns over the last year. Our team is improving automated detection methods based on behaviors we observe and engineering campaign-specific defenses to block operators from returning to our platforms. This steady defensive pressure forces operators to evolve constantly, which degrades the quality of their campaigns and reduces the reach and impact of influence operations on our platforms and the broader internet.

Deep Dive: AI-Enabled Influence Operations

As we increasingly see influence operations we disrupt using AI in some form, we are providing a detailed assessment below of how this integration is evolving and what it means for defenders. In our H2 2025 Adversarial Threat Report, we assessed that while artificial intelligence was lowering barriers to entry for influence operation actors and increasing their operational efficiency, behavioral and technical defenses remained robust against AI-enabled threats. Nearly a year later, we observe that the underlying dynamics have shifted in ways that warrant close attention.

AI-enabled content generation has moved from an occasional tactic employed by the most resourced actors to a near-standard component of influence operation tradecraft. Every one of the CIB networks we disrupted and shared in this report incorporated some form of generative AI that we could identify—ranging from basic profile photo generation to fully automated, multi-modal

content production pipelines. This near-ubiquity represents a meaningful shift from even twelve months ago.

Below, we outline three key trends that characterize this evolution.

Trend 1: Full-Stack Synthetic Personas

The earliest and most common AI application in influence operations, synthetic profile photos, remains prevalent. However, we now routinely observe actors deploying AI across multiple content modalities simultaneously: text, static imagery, and video within a single brand, persona, or campaign.

One Russia-origin network (see in Other Case Studies below - 5. Russia) we disrupted exemplified this shift. The operation deployed what we term "full-stack synthetic personas": accounts featuring AI-generated profile photos and videos of the same fictitious personas, as well as AI-generated backstopping, including bios, research articles, and/or graphics.

This integration of AI-generated text and media across various facets of online profiles provides a greater semblance of authenticity for these inauthentic accounts.

Image: Example of a “full stack synthetic persona” on Instagram from a Russia-origin network.

Trend 2: Industrialized AI Workflows

Perhaps the most technically significant development we observed was threat actors experimenting with advanced AI- and software-based tools to automate AI-generated content creation and distribution at scale. We disrupted multiple networks engaging in this kind of activity.

For example, one commercial operation based in Israel demonstrated a fairly sophisticated setup, integrating one AI-platform's API to automate commenting at scale. The operation deployed clusters of inauthentic, automated accounts that used AI to mass-produce contextually relevant comments designed to artificially boost engagement on the network's primary assets, a tactic to boost organic reach.

These developments suggest a transition from AI as a content generation tool to AI as infrastructure, embedded within automated systems that can produce, adapt, and distribute content with minimal human intervention per unit of output.

Trend 3: Video Generation Reaches Operational Maturity

We observed video generation capabilities across actors of varying sophistication and origin, including networks linked to Russia, Iran, and IO-for-hire companies.

Applications varied: some actors generated talking-head videos of synthetic personas; others produced stylized animated content; still others created video advertisements. The Armenia-targeting ANO Dialog network we disrupted produced AI-generated videos that maintained visual consistency with their operators' generated profile photos—a level of persona management that would have required significant manual effort previously.

Notably, video generation appeared across both state-linked and for-hire influence operations, suggesting the capability is diffusing. While AI-generated video remains distinguishable through forensic analysis in most cases we reviewed, the volume and variety of deployment indicates actors are finding operational value despite current limitations.

Assessment and Outlook

Despite the capabilities expansion documented above, our core assessment from 2025 holds: behavioral detection approaches remain effective against AI-enabled influence operations. The networks disrupted during this period were identified through behavioral signals, technical indicators, and network analysis—signals that are difficult to mask entirely, even with AI.

However, the velocity of adoption and integration we observed warrants continued investment in AI-enabled defensive capabilities. The industrialization of AI workflows, in particular, suggests that the next phase of this competition may be defined less by content quality than by operational scale and adaptation speed.

Other Case Studies

1. France & Spain

We disrupted a Coordinated Inauthentic Behavior network originating in France and Spain and targeting audiences in France, Mali, Chad, DRC, Senegal, Cameroon, Burkina Faso, and Côte d'Ivoire. We removed 8 Facebook accounts, 24 Pages, and 6 Instagram accounts for violating our

policy against Coordinated Inauthentic Behavior. About 184,200 accounts followed one or more of these Pages, and about 5,400 accounts followed one or more of these Instagram accounts. Through these enforcement actions, we determined that network operators engaged in around $21,400 in spending for ads, paid for mostly in US Dollars, on Facebook and Instagram to amplify their content. Meta's investigation revealed links to foreign freelancers, likely working on behalf of Russian individuals to obfuscate the network's origin.

The actors behind this operation used fake profiles, some of which appear to have been procured from account farms, to admin pages, place ads, and generate content. The operation demonstrated an ability to adapt to enforcement actions; in an attempt to avoid our political advertising policies, it ran non-civic patriotic advertisements to build authentic audiences, then organically seeded political content into their feeds. To execute their campaign, operators used inauthentic profiles to administer a collection of local-branded Pages featuring regional nuance, and coordinated across multiple cells, even sharing ad creatives concurrently with other known Russian operations. We observed that network operators consistently produced and amplified narratives supporting Russian political interests, including anti-Ukraine sentiment in France and the promotion of military-technical collaboration between Moscow and the Sahel region.

2. Israel

We disrupted a Coordinated Inauthentic Behavior network originating in Israel and targeting audiences in France, the United Kingdom, Australia, Iran, Togo, Gabon, Angola, and the United States. The operation maintained a cross-platform presence, including on TikTok, X, and their own websites, and we removed 215 Facebook accounts, 4 Pages, and 1,044 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 32,500 accounts followed one or more of these Pages, and about 248,000 accounts followed one or more of these Instagram accounts. Network operators engaged in around $100 in spending for ads. This network maintained clusters of lower-sophistication, automated fake accounts specifically designed to artificially boost engagement to its main accounts. Meta's investigation revealed links to individuals based in Israel. We assess this is an influence-for-hire operation, likely run on behalf of a number of distinct clients.

Although the operation targeted a number of countries with targeted narratives, operators followed the same tactics in developing these distinct manipulation campaigns; the actors behind this activity established Pages and Instagram accounts posing as local, civically-engaged brands, news sources, religious organizations, and political activists. The operation then deployed clusters of fake, automated accounts that used AI to mass-produce contextually relevant comments designed to artificially boost engagement on the network's primary assets, a tactic to boost reach that the network's own operators referred to as 'Mother-Child'. Ahead of France's 2026 municipal elections, the network created brands to criticize political candidates and spread narratives about the "Islamization" of France. Network operators disseminated varied messaging tailored to different regions, including partisan content in Togo and Gabon, anti-government messaging in Iran and Angola, and anti-immigrant and anti-Muslim campaigns in France, the United Kingdom, and Australia. The network also aimed a small cluster of these bots at the NYC mayoral campaign.

These accounts posted into local Groups and made comments on Facebook. We found this network as part of Meta's investigation into suspected Coordinated Inauthentic Behavior in the region.

Image: Example of comments by inauthentic accounts on an Instagram account controlled by the network.

3. Iran

We disrupted a Coordinated Inauthentic Behavior network originating in Iran and targeting audiences in Nigeria and South Africa, as well as Kenya, Ethiopia, and Sudan. The operation maintained a cross-platform presence, including on X. We removed 23 Facebook accounts, 13 Pages, 1 Group, and 11 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 12,800 accounts followed one or more of these Pages, about 400 accounts followed this Group, and about 4,200 accounts followed one or more of these Instagram accounts. Meta's investigation revealed links to actors in Iran.

The network employed sophisticated persona development, creating inauthentic accounts with specific activist identities, such as researchers, nurses, and women's rights advocates. Operators used profile pictures of local individuals overlaid with activist symbols to gain credibility with target audiences and hide the Iranian origin of the accounts. The operation also invested in generating

high-production-value content using artificial intelligence to create polished graphics aligned with the network's narratives. To disseminate their content, operators posted into local activist Facebook Groups, managed localized Pages, and utilized inauthentic accounts to comment on the network's posts to simulate popularity. The network employed advanced operational security, using proxy IP addresses and clean devices to minimize infrastructure-level overlaps. We observed that network operators systematically amplified narratives promoting anti-Western and anti-imperialist themes, supporting pro-Hamas activism, and advocating against African states normalizing diplomatic relations with Israel.

4. Russia

We disrupted a Coordinated Inauthentic Behavior network originating in Russia and targeting audiences in Hungary. We removed 6 Facebook accounts and 4 Pages for violating our policy against Coordinated Inauthentic Behavior. About 15,200 accounts followed one or more of these Pages, and we determined through these enforcement actions that network operators engaged in around $1,400 in spending, paid for mostly in US Dollars, for ads on Facebook. We found this network as a result of Meta's investigation and took action ahead of the election in Hungary. Our investigation found links to individuals in Russia.

The operation employed sophisticated operational security and impersonation tactics, running Pages that posed as local Hungarian and Ukrainian media. These assets, managed by inauthentic accounts obscuring their origin through proxy servers, amplified links to a set of external domains, which included both fabricated news sites and fictitious investigative journalism outlets. Operators activated dormant Pages—some created months in advance to post legitimate news to appear authentic—and renamed them simultaneously just prior to launching their campaigns. They then utilized undeclared political advertisements to launder their fabricated, AI-generated content from the deceptive websites onto our platforms to target domestic audiences in Hungary. We observed the network systematically disseminate messaging focused on the Hungarian elections, utilizing narratives designed to discredit Fidesz's political opposition—most notably the Tisza party and its candidates.

5. Russia

We disrupted a Coordinated Inauthentic Behavior network originating in Russia and targeting Western audiences, with France as a primary focus. The operation maintained a cross-platform presence including on Telegram, TikTok, YouTube and X (threat indicators can be found in our GitHub repository). We removed 7 Facebook accounts, 9 Pages, and 8 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 200 accounts followed one or more of these Pages and about 200 accounts followed one or more of these Instagram accounts. We determined through these enforcement actions that network operators engaged in around $6,400 in spending for ads on Facebook and Instagram, paid for mostly in US Dollars, Euros, and Guatemalan Quetzales. Meta's investigation revealed links to actors in Russia.

The operation's central tactic was the creation and backstopping of a fabricated think tank called the "International Burke Institute," which posed as a legitimate academic institution based in Israel. Operators created a pseudo-scientific "Burke Sovereignty Index" that claimed it could measure the real independence of states. The index was invoked to support the operation's narrative that certain countries were losing sovereignty to the US and institutions such as NATO and EU. The network used AI to create fake personas on Instagram—posing as analysts, investigative journalists, and former government employees—complete with AI-generated profile photos and face-to-camera video Reels of the same nonexistent individuals. The individuals behind this network attacked Western institutions within the United States and European governments, while portraying the latter as lacking true sovereignty. Furthermore, the operators attempted to launder their narratives into the legitimate academic realm by seeding fabricated research papers featuring fictitious authors into authentic academic research hosting platforms. We found this network after reviewing information shared with us by our peers at OpenAI.

6. Ukraine & Bulgaria

We disrupted a Coordinated Inauthentic Behavior network originating in Ukraine and Bulgaria and targeting audiences in the United States, Germany, Italy, the United Kingdom, the Netherlands, France, Belgium, Latvia, Ukraine, and Moldova. The operation had a cross-platform presence, including accounts on TikTok, X, and Telegram as well as maintaining their own websites. We removed 72 Facebook accounts and 16 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 500 accounts followed one or more of these Instagram accounts. Meta's investigation revealed links to a Ukrainian expat in Bulgaria.

The network employed sophisticated grassroots impersonation tactics, creating a series of localized news Pages and standalone websites whose names and languages aligned with their target countries. The operation attempted to make these websites appear authentic by investing in distinct branding, utilizing Large Language Models to generate unique articles interspersed with influence content, and employing fictitious bylines backstopped by fake journalist accounts on our platforms. The network utilized a multi-layered distribution strategy incorporating two distinct amplification clusters to distribute narratives from the core Pages. The first was a traditional network of inauthentic amplification accounts that posted URLs directly to target Pages, while the second cluster exhibited scripting signals by posting identical narratives across different accounts in very short succession. We observed that network operators produced and amplified narratives advocating support for Ukraine's military, raising fears of Russian influence in the European Union, and undercutting politicians deemed pro-Russian.

7. Russia

We disrupted a Coordinated Inauthentic Behavior network originating in Russia and targeting audiences in Europe (Ukraine, Moldova), the Middle East (the United Arab Emirates and Syria), and Africa (Libya and the Central African Republic). The operation maintained a cross-platform presence, including on Telegram, and we removed 16 Facebook accounts, 24 Pages, and 1 Group for violating our policy against Coordinated Inauthentic Behavior. About 64,600 accounts followed

one or more of these Pages. We determined through these enforcement actions that the network engaged in around $27,600 in spending for ads on Facebook, paid for mostly in US Dollars. Although the actors behind this network took sophisticated measures to conceal their identities and location, Meta's investigation attributed this activity to Russian individuals. Some of these individuals were previously suspected of ties to the Internet Research Agency, which was shuttered in 2023. We also observed overlaps with a Russian network that we previously linked to false flag real-world stunts in Germany and France, which we reported on in August 2024.

The operators behind this network engaged in complex deception by masquerading as legitimate local entities and creating fictitious opinion research pages, utilizing fake and compromised accounts to administer Pages and developing localized fictitious personas. To obfuscate their location, the network registered with phone numbers aligned with their target personas locations and utilized proxy IP infrastructure. The network also operated Telegram channels that impersonated prominent Ukrainian politicians. This network criticized the governments of Moldova and Ukraine and advanced narratives supporting the governments of Azerbaijan and the UAE. It also produced original video content designed to influence judicial proceedings related to an Interpol extradition request in the UAE and Azerbaijan. The network promoted the incumbent government in the Central African Republic. In Libya, the network posed as local news agencies, which shared narratives that were critical of Italy and the International Criminal Court.

8. Iran

We disrupted a Coordinated Inauthentic Behavior network originating in Iran and primarily targeting audiences in Azerbaijan, alongside minor targeting of neighboring Georgia and Armenia. The operation maintained a cross-platform presence, including on TikTok and Telegram. We removed 15 Facebook accounts, 7 Pages, and 96 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 23,600 accounts followed one or more of these Pages, and about 77,200 accounts followed one or more of these Instagram accounts. Although the people behind it attempted to conceal their identity, our investigation found links to Azerbaijani individuals in Iran.

The network engaged in sophisticated brand development, using inauthentic accounts to manage media entities that masqueraded as authentic local Azerbaijani news outlets. To tailor their messaging, operators employed a highly segregated content strategy, building distinct brands focused exclusively on single geopolitical topics—such as criticizing the EU, domestic policy, or promoting religious content—without using overt pro-Iran framing. The operation also demonstrated a notable investment in original content production, leveraging AI-generated media imagery and videos featuring AI-generated characters to criticize Azerbaijan's domestic policies. To amplify these narratives, the network utilized clusters of inauthentic accounts to artificially boost engagement on their own media accounts, while also commenting directly on the posts of authentic Azerbaijani media entities. The operation demonstrated moderate operational security through strict operator and device segregation between brands, proxy IP usage to obfuscate their origins, and the creation of reserve accounts in an attempt to increase the longevity of their brands

against enforcement actions. We observed that network operators promoted pro-Iran and pro-Shia narratives, criticized Azerbaijani internal affairs, and attempted to undermine confidence in the Azerbaijani government.

9. Iran

We disrupted 4 Facebook accounts and 31 Instagram accounts originating in Iran and targeting audiences in the United States for violating our policy against Coordinated Inauthentic Behavior. About 79,400 accounts followed one or more of these Instagram accounts. Although the people behind it attempted to conceal their identity, our investigation found links to actors in Iran.

We observed that network operators relied on sophisticated operational security and evasion tactics to mask their Iranian origins, exclusively using United States and Canadian proxy or hosting IPs. The operation used inauthentic accounts to pose as US-based individuals—such as activists, students, and graphic designers—who claimed to be based in American cities like Washington D.C., San Diego, and Atlanta. These fake personas were used to manage civic meme accounts and engage with authentic American users by tagging real journalists and politicians in their posts to solicit content collaboration. Once their infrastructure was established, the operators produced and amplified content (some of which was AI-generated) focusing on American politics, including anti-Republican views, alongside messaging on the Israel-Palestine conflict and immigration. We found this activity as a result of an internal investigation.

04

TERRORIST ORGANIZATIONS

We address the risks posed by terrorist organizations through the lens of our Dangerous Organizations and Individuals policy (DOI). Terrorist organizations and their members seek to exploit our platforms to achieve a range of objectives. They seek to abuse our platform to distribute propaganda, both official and produced by their membership; to build their social networks and to recruit new members; to intimidate our users, and the members of political, social, ethnic or religious groups that they regard as adversaries; to share terrorist materials and resources, to issue calls to violence against specific targets; and to raise funds, either directly or through the misrepresentation of charitable causes.

Over the years, the purposes for which terrorist organizations seek to exploit our platforms have remained remarkably consistent. However, the methods they deploy, and the adversarial behaviors they exhibit as they try to evade enforcement, are always changing. Most of our 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. Guided by the insights of our subject matter experts, we continually change and develop the detection and enforcement methods we deploy against malicious actors and the content they produce.

Actor enforcement—enforcement against accounts rather than against content—is central to our enforcement against Dangerous Organizations and Individuals. Prominent among the actor enforcement tools we deploy are Strategic Network Disruptions: expert-led, targeted actions that identify and disable whole networks of bad actors, while gathering insights that can inform our wider integrity efforts, from prevention to scaled enforcement. We use Strategic Network Disruptions to remove networked bad actors who seek to abuse our policies across a range of harm areas, including those mentioned elsewhere in this report.

Recently, we have started to combine high-impact Strategic Network Disruptions against Dangerous Organizations and Individuals with another area of work: targeted, message-based prevention campaigns which we call Custom Messaging. Custom Messaging campaigns conducted on Facebook and Instagram are designed to safeguard, educate and inform users that we believe

may be particularly vulnerable to exploitation by Dangerous Organizations. For example, they may be vulnerable because they have engaged with a violating user or content we have subsequently disabled, or because of other factors, such as an ongoing civil conflict in their country. Often these campaigns are delivered in partnership with or following consultation with Non-Governmental and Civil Society Organizations that combine counter-terrorism, counter-organized-crime or youth safety expertise with in-country experience. By pairing Strategic Network Disruptions against Dangerous Organizations with Custom Messaging campaigns that reach the people who have potentially interacted with those networks, we can direct our prevention efforts toward those who are most likely to benefit from them.

In our previous Adversarial Threat Report, we shared an update on our new AI image comprehension tools, a type of computer vision which allowed us to sift through huge volumes of data to identify important signals. This helps us more efficiently disrupt cartel and illegal narcotics dealing networks, and it has also helped us make significant gains in disrupting terrorist organizations. Over the past year, AI image comprehension has become a key component in amplifying human investigators' expertise at scale.

In this chapter, we detail a number of case studies of Strategic Network Disruptions we have undertaken against terrorist groups, the novel, often AI-based actor enforcement methods we have applied to them, and the associated Custom Messaging campaigns we have conducted in support of younger or vulnerable people they have sought to exploit. These case studies include actions taken targeting an Afghanistan & Pakistan-based terrorist group and a Pakistan-based ethno-nationalist terrorist group, a terrorist organization, and an extreme far-left terrorist organization in Colombia. While these organizations vary widely in their ideology and operating context, they overlap in their use of social media to attempt to recruit and exploit our users—and particularly younger people.

Terrorist Networks in Pakistan and Afghanistan

Pakistan and Afghanistan-based Terrorist Group

In the second half of 2025, we identified a network operating on Meta platforms that was maintained by a terrorist group operating in the border region of Afghanistan and Pakistan. This terrorist group aims to overthrow the Pakistani state through violence. It maintained the network in an effort to engage with sympathetic audiences, disseminate propaganda, and coordinate among members through messaging features. Our investigation identified accounts that appeared to be actively seeking to recruit fighters, including through the targeting of youth directly on Facebook Messenger. These accounts also sought to glorify child soldiers and conduct intra-group coordination activity, including the exchange of operational updates and tactical guidance within group messaging threads.

The network relied in part on accounts representing official terrorist media outlets, which sought to disseminate propaganda across both Facebook and Instagram. Operators also engaged in efforts

to use messaging apps for deeper coordination, while directing audiences to content on other social media platforms. Operators unsuccessfully sought to evade enforcement through the creation of multiple accounts across numerous different devices.

In September 2025, we disabled the accounts that comprised this network for violating our Dangerous Organizations and Individuals policy. This included a total of about 1,600 assets, largely Facebook accounts, with some Facebook pages, groups, and Instagram accounts. As part of this enforcement, we also blocked a series of URLs associated with the network's off-platform propaganda infrastructure from being shared on our platforms. These URLs were mainly links to specific assets on other social media platforms.

We coordinate with the Pakistani authorities in line with applicable law and our platform policies through information exchange. This information exchange strengthens our counter-terrorism enforcement.

Pakistan-based Ethno-Nationalist Terrorist Group

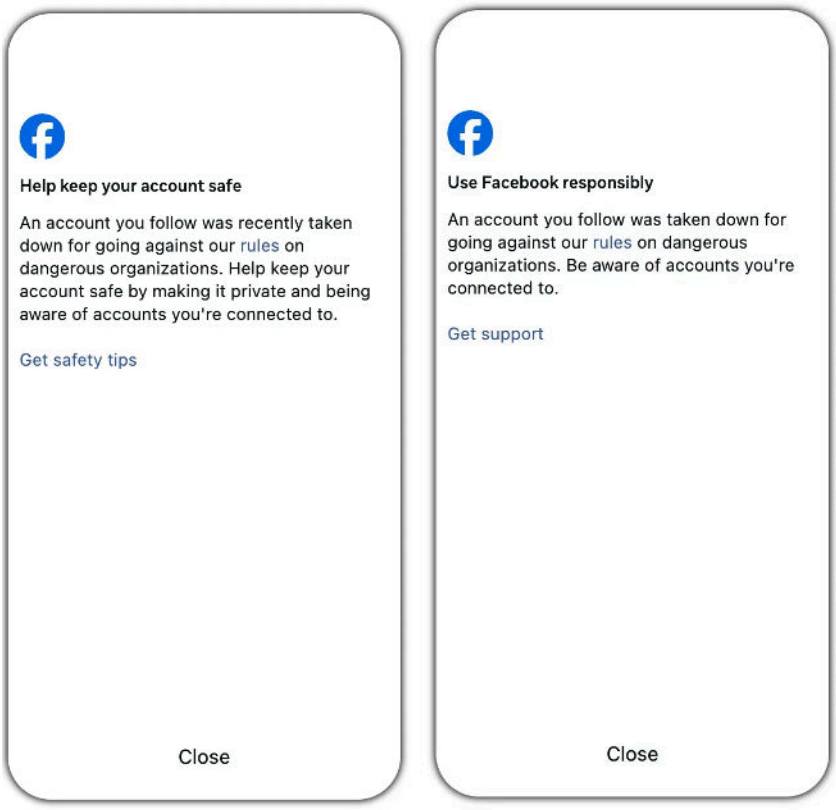
We also conducted Strategic Network Disruptions against an ethno-nationalist terrorist group operating in Pakistan. This organization, which seeks to establish an independent state through violence, used our platforms to attempt to disseminate propaganda and recruit supporters.

In Strategic Network Disruptions conducted across 2025, we disabled approximately 450 assets associated with this network. A majority of the disabled assets were initially surfaced by our AI image comprehension systems.

This combination of manual investigation, AI detection, and rapidly built, targeted automated enforcement is increasingly characteristic of how we conduct Strategic Network Disruptions targeting terrorist groups. The use of subject matter experts to build these detection and enforcement mechanisms ensures high precision, while the use of automated enforcement functions at the sub-scale level—targeting a narrow threat type or category of bad actor—allows us to maintain continuous pressure on networks between disruptions. This in turn makes it significantly more difficult for terrorist organizations to reconstitute their presence.

Custom Messaging in Pakistan

Following these two sets of Strategic Network Disruptions, we launched a Custom Messaging campaign targeting younger users in Pakistan who had been connected to accounts associated with these terrorist networks. This campaign exemplifies our increasingly typical approach of pairing enforcement with prevention. By identifying the younger people who had interacted with or followed accounts associated with terrorist organizations, we were able to deliver targeted educational messages designed to inform them of the risks of engaging with such groups, and to empower them to protect their online experience. These messages were delivered in-app, on Instagram and Facebook. The screenshot below shows the kind of messaging that is used in Custom Messaging campaigns, though the specific wording, and associated links, changes from campaign to campaign.



The campaign reached approximately 130,000 users in Pakistan, who had connections to previously disabled terrorist accounts. Targeting thresholds were calibrated to prioritize early intervention, ensuring that even the most tangentially connected youth would receive supportive messaging.

The messaging language explained their connection to dangerous organizations, cautioned about the risk of further enforcement, and provided educational links—including links to our Community Standards and safety tools—to help those younger people improve their digital experience.

For this campaign, we partnered with Shehri Pakistan, a civic and digital literacy initiative, that produced a custom animation video titled "The Choice of Love for a Changing World." This video serves as a motivational appeal encouraging viewers to choose compassion over resentment, and is embedded on Shehri Pakistan's landing page, as a linked resource within the custom messages.

Terrorist Organization in Syria, Iraq, and Other Countries

Resurgence and Cross-Platform Displacement

Across 2024 and early 2025, we identified significant cross-platform migration of terrorist groups, as well as sympathetic groups. In some cases, this led some terrorist groups to seek to return to

Facebook and Instagram. This resurgence was driven by a number of factors. Changing patterns of enforcement across some other social media platforms displaced propaganda networks that previously had an established presence on those platforms, and led them to seek other platforms to try to exploit. Simultaneously, geopolitical events created conditions that made terrorist groups more able to exploit violent imagery as a recruitment tool, and to exert messaging reach. We shared these observations with industry partners through the Global Internet Forum to Counter Terrorism (GIFCT).

We observed an increase in accounts associated with a specific terrorist group during this period. The organization leveraged established cross-platform coordination techniques: members maintained command-and-control channels on other social media messaging platforms, while directing followers to newly created Facebook and Instagram accounts. By exploiting cross-platform links, these actors sought to more rapidly build engagement on Meta's services.

This pattern underscores the need for continual vigilance to prevent adversarial networks from reconstituting across platforms, and reinforces that no single company can solve terrorist exploitation of the internet alone. That is why Meta has invested in cross-industry collaboration through the Global Internet Forum to Counter Terrorism (GIFCT). As a founding member, Meta helps build and maintain GIFCT's shared technical infrastructure that enables companies of all sizes to detect and disrupt terrorist actors at scale. GIFCT also helps level up capacity across the industry, equipping newer and smaller platforms with tools and expertise to rapidly disrupt adversarial actors online. Through shared insights on emerging trends, threat actor tactics, and cross-platform migration patterns, GIFCT enables a more coordinated industry response, making it harder for terrorist networks to operate online.

On Meta platforms, these accounts largely sought to distribute propaganda, and to direct new followers toward their presence on other social media platforms. The network also sought to evade automated detection and human reviewers through text and image-based obfuscation.

Strategic Network Disruptions

Following the detection of these networks, we conducted a series of large-scale enforcement operations. These disruptions were designed to dismantle the coordinating infrastructure of the network, not just remove individual pieces of content, and in doing so to undermine their ability to reconstitute on our platforms.

In December 2024, we removed about 1,100 assets in the first coordinated disruption, spanning multiple countries, with the highest concentrations in Syria, Iraq, Türkiye, Jordan, and Afghanistan.

In February 2025, we conducted a dedicated two-week intensive enforcement effort, a surge effort involving multiple specialist teams. This resulted in the removal of about 4,300 more assets. This concentrated approach proved highly disruptive to the network's ability to operate. In order to prevent it from reconstituting, we deployed sub-scale automated enforcement, in addition to regular counter-recidivism methods. As a result, numerous recidivist accounts associated with the

network were detected and disabled, and newly produced violating media content was added to our AI-assisted image and video detection systems.

Automated Enforcement Pipelines

Recognizing that terrorist networks rapidly attempt to reconstitute after disruptions, we deploy continuous enforcement capabilities to maintain sustained pressure, in addition to augmenting the scaled, classifier-based enforcement that results in a significant portion of DOI enforcement. These automated systems are a key way in which we maintain enforcement between manual disruptions. In our enforcement against this terrorist organization, we produced a number of innovative automated enforcement pipelines.

We built an automated system designed to detect and block off-platform URLs that the organization used to host and distribute propaganda publications on file-sharing websites. Members of the group shared these URLs, primarily hosted on disposable file-sharing platforms, in the comments sections of high-visibility posts in an effort to evade enforcement. The pipeline, which uses a Large Language Model (LLM) to conduct initial analysis of the links' features, disables violating URLs with high confidence to prevent their further distribution on our platforms. Since its launch, the pipeline has detected more than 100 URLs across more than 10,000 comments and shares, operating at 100% precision. This approach targets the distribution infrastructure rather than individual posts, providing a more efficient enforcement method.

We also developed and deployed a new automated system that detects accounts based on the ways they signal their group membership to potentially sympathetic audiences. By targeting the specific methods these accounts tried to use to advertise their affiliation, this system disrupts a core element of their attempted on-platform strategy.

Following the deployment of these enforcement efforts, from Strategic Network Disruptions to automated pipelines, we observed high levels of frustration among the organization's operatives. Members of the group were observed complaining about the increasing difficulty of day-to-day operations. Some accounts within the network expressed their desire to move their network off Meta platforms even before their accounts were also disabled.

We continued this pattern of actor enforcement, driven by Strategic Network Disruptions and automated sub-scale enforcement, through multiple subsequent rounds throughout 2025. Across all enforcement rounds in 2025 and early 2026, we removed about 8,900 assets associated with this organization through strategic network disruptions. During this period, we also referred credible threats of violence to law enforcement authorities in the United States and Europe, leading to law enforcement action. Additional rounds of enforcement addressed affiliated organizations operating in Southeast Asia and the Middle East.

AI image comprehension helped investigators speedily detect adversarial actors and obfuscated terrorist content. Unlike traditional classifiers, LLMs are able to understand and reason through content they have not been exposed to before, making them well suited to deployment against

adversarial actors. Leveraging these new tools, investigators can identify certain categories of violating terrorist content an order of magnitude more quickly.

Custom Messaging in Germany

Among the most concerning trends identified during this period was the organization's deliberate and systematic targeting of younger people. Its propaganda apparatus shifted toward content specifically designed to appeal to younger adult audiences, particularly in Europe. Rather than leading exclusively with graphic violence, these materials also emphasized themes of identity, belonging, community, and purpose—psychological needs that research has identified as particularly salient among younger people. The organization's content featured images of child soldiers, and explicitly addressed messaging "to the youth," seeking to inspire younger people to align with its agenda.

On our platforms, we observed this activity concentrated in Western European countries, particularly Germany, where affiliated accounts used local languages and culturally relevant messaging in an effort to connect with younger users. Accounts operated as coordinated networks, seeking to amplify one another's content and direct followers to other social media messaging platforms for further engagement. The network particularly sought to exploit Instagram Stories and Reels for these recruitment efforts.

Following enforcement against these accounts, we launched our first counter-terrorism Custom Messaging program in Europe. This campaign delivered tailored educational content to younger users identified as potentially exposed to the organization's propaganda materials.

The campaign targeted approximately 8,000 users aged 13–34, primarily based in Germany, who had followed one or more accounts previously disabled for terrorism. The audience was divided into two segments based on how they had engaged with the network or its content, with each segment receiving differentiated messaging appropriate to its level of involvement. The messaging clearly articulated to users why they were receiving the communication, enhancing transparency and deterring future violating behavior. The message content further sought to empower users through links to educational, safety, and support resources, including updating Meta's Safety Center with resources for youth involved in violating activity.

Terrorist Organization in Colombia

An extreme far-left terrorist organization operating across the Colombia-Venezuela border was identified as maintaining a propaganda network on Meta's platforms. This terrorist organization has prospered in recent years by expanding into territories vacated by a rival group, taking over illegal narcotics trafficking routes and contraband operations. It maintains close ties to transnational drug trafficking organizations.

On-platform, the organization sought to use Meta's services to disseminate ideological material and recruit new members. This organization is known to specifically try to target younger people

over social media. It also sought to highlight its capacity to challenge state authority, by posting images and videos of its presence in rebel-controlled regions, as well as footage of military-style training and weaponry. Propaganda accounts further attempted to direct audiences to off-platform propaganda radio channels, hosted on encrypted messaging services. Moreover, our investigation found evidence that members of this network might attempt to use the platform to coordinate offline harm, including threatening their rivals.

Strategic Network Disruptions

We conducted a series of Strategic Network Disruptions targeting this organization's infrastructure over 2025. In the first enforcement round in early 2025, we disabled more than 250 assets associated with this network, largely Facebook accounts. We also blocked a number of URLs representative of the group's propaganda radio infrastructure, predominantly links to channels on encrypted messaging platforms. In a subsequent enforcement round in May 2025, we disabled more than 800 assets associated with this organization and an allied terrorist group, also operating in the region, and blocked additional URLs.

A central focus of this disruption was the organization's propaganda radio network, a decades-old broadcast apparatus that has long served as a primary vehicle for recruitment and ideological messaging. As traditional broadcast reach has declined, the network migrated online, redistributing its audio programming across encrypted messaging channels and inauthentic social media accounts presented as independent news outlets. We targeted this infrastructure directly, by removing the accounts that amplified and redistributed the radio content, and blocking the URLs used to direct users to its off-platform channels. Because the network distributed the same programming redundantly across many accounts and web domains, so that individual account disables would not stop the network’s operation, AI-assisted detection and automated recidivism systems were central to our enforcement. These methods enabled us to surface the dispersed and frequently re-created assets quickly and to sustain enforcement pressure as the network attempted to reconstitute.

Proactive Automated Enforcement

Following these manual disruptions, we deployed an AI-based pipeline specifically targeting terrorist organizations in Colombia. This system uses an LLM to identify high-confidence accounts representative of the organization, and surfaces them for a human’s final decision. This represents a significant advance in our counter-terrorism enforcement capability. By combining the insight of subject-matter experts with the scalability of automated systems, we can maintain continuous enforcement pressure on terrorist networks, making it substantially more difficult for them to reconstitute after disruption, and representing a model that can be extended to other terrorist organizations globally.

Complementing this LLM-based system, our subject-matter experts also operate a set of automated pipelines driven by expert-defined diagnostic signals–including characteristic keywords, naming conventions, and network markers. These signal-driven pipelines are a foundational

component of our proactive enforcement: in 2026, they detected and disabled more than 350 assets associated with these terrorist organizations in Colombia, and continue to surface accounts as the network attempts to re-create them.

Custom Messaging in Colombia

Following these enforcement operations in Colombia, we launched a Youth Custom Messaging campaign—our first DOI campaign in the country—designed to safeguard younger users from recruitment and influence by this terrorist organization.

The campaign targeted approximately 20,000 users in Colombia, aged 13–25, who had interacted with networks associated with terrorist organizations operating in the region. The audience was divided into segments based on the severity of observed interactions: a primary audience comprising users with friend connections to previously disabled members of terrorist organizations, and a secondary audience comprising users with both friend connections and behavioral signals indicating increased vulnerability to exploitation. For this campaign, we developed new audience identification methods, including the innovative use of AI image comprehension to identify at-risk users in the absence of reliable traditional signals—a technique that achieved high precision in identifying potentially vulnerable youth.

We partnered with Tiempo de Juego, a Colombian NGO, to develop tailored digital resources and educational content for vulnerable younger people. 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. This collaboration ensured that our interventions were both effective and deeply relevant to the local community.

Looking Ahead

Terrorist organizations remain adaptive and persistent adversaries. As enforcement improves across the technology industry, we anticipate a continued evolution in their tactics, including greater use of cross-platform coordination, more extensive obfuscation techniques, and further deliberate targeting of younger people. The AI-driven detection and enforcement systems described in this section represent a new generation of counter-terrorism tools. These new capabilities allow us to maintain continuous enforcement pressure on terrorist networks, closing the gap between adversary adaptation and our detection. Counter-terrorism enforcement has historically been constrained by the speed at which content mutates, with new propaganda themes, rapidly changing coded language and new visual styles. AI partially overcomes that constraint by enabling automated systems that don’t purely rely on memorized features, recognizing harmful content even as it changes. It also helps us build more accurate audiences for more effective prevention campaigns. Combined with expert-led analysis and cross-industry collaboration, an AI-augmented approach will be fundamental to Meta’s future counter-terrorism efforts.

05

SURVEILLANCE-FOR-HIRE

Introduction

The surveillance-for-hire industry continues to target the users of our platforms and people across the internet. Over the past year, we have seen established vendors attempt to come back onto our platforms, while newer companies have entered the market with their own tooling and customer bases. These actors continue to target journalists, activists, political figures, and others around the world on behalf of their clients, as evidenced by recent public reporting.

We have been investigating and taking action against the surveillance-for-hire industry for years. We continue to investigate, track, and take action against this activity. Our response includes removing attacker and test accounts from our apps, blocking malicious infrastructure and links from being shared on our platforms, hardening product security against potential vulnerabilities these vendors seek to exploit, pursuing legal action against vendors where appropriate, notifying users we believe have been targeted, and sharing threat indicators with industry peers and the broader security community. This work spans multiple teams across the company and reflects a commitment to disrupting the full attack chain, rather than addressing any single stage in isolation.

The Surveillance-for-Hire Threat Landscape

Surveillance-for-hire vendors provide the capabilities to deploy complex, strategic campaigns against their targets, which have frequently included civic actors, journalists, and activists. These vendors leverage a variety of techniques from **zero to multi-click exploits** that range from no user engagement to some user engagement based on the scenario, expertise of the vendor, funding, and targets involved. These tools challenge traditional defense mechanisms and surreptitiously work to compromise people's devices. These tools and capabilities are sold to governments and agencies around the world, democratizing intrusive surveillance capabilities to anyone willing to pay.

Campaigns & Cases from 2025 and 2026

In 2025 and 2026, we continued to see both well-known and newer surveillance-for-hire operators targeting the users of our apps with zero-click and multi-click attacks. We disrupted recidivist activity from companies such as Intellexa/Cytrox (Predator) and NSO, while also detecting up-and-coming players on the market without clear attribution at this time. We hardened our own products and worked with industry partners to help secure their systems where we observed those systems forming parts of exploit chains across multiple threat actors. The following cases exemplify some of the trends and tactics we investigated during this time period.

Intellexa/Cytrox

Despite international sanctions and repeated enforcement actions, Intellexa and its subsidiary Cytrox remain a persistent adversary. Cytrox produces the spyware product known as Predator, and we have tracked their attempts to target our platforms since 2021. During this reporting period, we identified multiple clusters of activity linked to Cytrox. On Facebook and Instagram, Cytrox created inauthentic accounts and used these for testing on Facebook Messenger and IG Direct Messages. We took action by disabling attacker accounts and sharing indicators with industry partners.

ASIGINT

WhatsApp identified and disrupted an Italy-based surveillance-for-hire company, ASIGINT (a subsidiary of SIO SPA), that created a fake iOS version of WhatsApp. ASIGINT is an Italian commercial spyware company that has been publicly reported to target people with malicious software posing as messaging or telecom apps. Our security team proactively identified around 200 users, primarily in Italy, who we believe may have downloaded this malicious unofficial client. We logged them out, alerted them to the risks to their privacy and security, and encouraged them to remove it and download the official WhatsApp app. We also disabled associated Facebook and Instagram accounts linked to the actor and shared our findings with our industry peers.

This was not a WhatsApp vulnerability; end-to-end encryption continues to protect communications of people who are using WhatsApp's official apps. The threat actors behind the malicious client used social engineering tactics to trick people outside of our app into downloading their malicious software masquerading as WhatsApp. ASIGINT created a modified version of WhatsApp's iOS client. We believe the actor distributed the fake application by tricking users into downloading it on third-party websites and developer testing channels. We sent a formal legal demand to ASIGINT to stop any such malicious activity.

Surfilter Network Technology and Sinovatio

Outside the traditional surveillance-for-hire space, we also disrupted Facebook and Instagram activity associated with two China-based companies that claim to specialize in network traffic interception and analysis—collecting and viewing certain types of data available via internet and

messaging traffic—rather than endpoint compromise. Even when this traffic is encrypted, surveillance vendors may be able to obtain or infer some information from interception, such as timing and volume of messages, and at times, information that could help identify the sender.

Surfilter Network Technology is a Chinese information security company with public offerings that include "Social Network Public Opinion Management." We found them creating inauthentic accounts and engaging in testing activity on Facebook and Instagram in an attempt to intercept messaging traffic.

The second company we found, Sinovatio—formerly known as ZTESec and originally spun off from ZTE due to its surveillance technology—also specializes in network traffic analysis. Sinovatio's publicly stated capabilities include intercepting traffic from Skype, Telegram, Viber, Facebook Messenger, Line, WeChat, and numerous other applications. Similar to Surfilter, we found them engaging in testing of our platforms via inauthentic accounts and took action to remove this activity.

These companies represent a different segment of the surveillance-for-hire industry: rather than targeting individual devices with spyware, they work at the network layer to try to intercept, inspect, and analyze communications traffic at scale. Their tools are likely used for reconnaissance, tracking, and bulk surveillance by their customers. We took action against their presence on our platforms and continue to monitor for recidivism.

Meta deploys a range of tactics to help protect our users' messages and data, including end-to-end encryption on many of our messaging services. There is no simple solution to network traffic interception, but users can also take steps to protect themselves, including using reputable Virtual Private Networks (VPNs), turning on encrypted DNS, keeping OS, browsers, and apps updated, and using built-in privacy features offered by apps and browsers.

Bindecy

We also investigated and enforced against an Israel-based surveillance-for-hire vendor called Bindecy that maintains a lower profile than many larger vendors. Bindecy provides a range of digital surveillance tools, one of which is a mobile spyware implant. Bindecy has connections to other Israeli surveillance companies, including having acquired Merlinx/Equus Technologies.

Bindecy set up a cross-platform network of inauthentic Facebook and Instagram accounts operated from Israel. These accounts exhibited behavior consistent with surveillance-for-hire vendors, including testing against Messenger and Instagram to confirm their implant functionality.

Empowering Users: User Reporting and Security Features

Much of this report focuses on our investigative and enforcement work: identifying threat actors, disrupting their operations, and removing their presence from our platforms. But defense against surveillance-for-hire is not solely a back-end effort. We have also invested in building user-facing features that give people direct control over their own security posture, particularly those who may

be at heightened risk of targeting. Some of these features are on by default; others are opt-in tools designed for users who need additional protection beyond what we provide to everyone. The following section covers what users can do to strengthen their own defenses, and how reporting suspicious activity back to us makes our investigative work possible.

User Reporting

WhatsApp and Messenger use end-to-end encryption to protect people's personal messages so that no one outside the chat can read them, including Meta, unless a user chooses to share them with us. For example, when you report someone on WhatsApp, the app receives up to the last five messages sent to you by the reported user or group so we can review and take appropriate action. We encourage anyone who encounters suspicious behavior on our messaging services to report it through the in-app reporting flows.

Security Features and Hardening

Beyond detection and enforcement, we continue to ship product features across our apps that reduce the attack surface and make it harder to target our apps.

WhatsApp

Strict Account Settings. WhatsApp's Strict Account Settings, a lockdown-style security mode, is designed for a few of our users, like journalists or public-facing figures, who may need extreme safeguards against rare and highly-sophisticated cyber attacks. When enabled, it bundles multiple protections against known spyware tactics into a single toggle: silencing calls from unknown numbers, restricting who can add the user to groups, blocking attachments and media from non-contacts, and disabling link previews. The feature also requires two-step verification and is located in Settings > Privacy > Advanced.

Device linking protections. WhatsApp protects users against attempts at account compromise through unauthorized device linking, a tactic where attackers may trick users into scanning a QR code that links their account to the attacker's device. WhatsApp will alert you when behavioral signals suggest a linking request might be suspicious. These alerts show you where the request is coming from and warn that it could be malicious, giving you the chance to pause and reconsider.

Protecting IP address on calls. This setting relays voice and video calls through WhatsApp's servers rather than establishing direct peer-to-peer connections, making it harder for an attacker to infer a target's IP address from a call. When using call relay, the call quality may be reduced.

Facebook

Security Checkup. Facebook offers a Security Checkup tool that reviews a user's account and recommends actions to strengthen it, including updating passwords, enabling two-factor authentication, and turning on login alerts.

Advanced Protection. For users who may be higher-profile targets, Facebook provides Advanced Protection, which simplifies the adoption of stronger security measures like two-factor authentication while adding additional protections such as monitoring for potential hacking threats against their accounts and Pages.

Message delivery controls. Users can choose who they receive messages and message requests from, restricting their inbox to a closer circle of contacts. Limiting who can reach you directly reduces the surface for phishing attempts or unwanted outreach from unknown accounts.

Instagram

Security Checkup. Instagram's Security Checkup guides users through steps to secure their account, including reviewing login activity, confirming profile information, auditing accounts that share login credentials, and updating recovery contact information.

Advanced Protection Mode. Some high-profile accounts may be automatically enrolled in Advanced Protection Mode, which applies stronger security defaults including mandatory two-factor authentication.

Limits. Instagram offers a Limits feature that lets users temporarily restrict interactions from certain groups without notifying those accounts. Users can limit interactions from everyone outside their Close Friends list or from accounts that started following them within the past week, which can be useful during periods of heightened targeting or public exposure.