DECEMBER 2025

SECOND - THIRD QUARTER

Adversarial Threat Report

Meta

TABLE OF CONTENTS

EXECUTIVE SUMMARY

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

With this new, restructured report, we are building on this effort by covering additional threat areas, including spotlights on Fraud and Scams and AI-enabled threats. We're adding new areas because threat actors continue to evolve, expanding their techniques and approaches. They also increasingly leverage artificial intelligence to enhance their operations and employ novel techniques to obscure their identities and manipulate our systems.

We are also adjusting the publication frequency from a quarterly to a twice-yearly cadence. This new format will enable us to focus on deeper analysis and better inform collective defense against threats that require a whole-of-society approach.

Finally, we are also updating our Inauthentic Behavior Community Standards to simplify and refine our IB and CIB policies and help uninvolved authentic communities, Pages, and Groups that are targeted, managed, or co-opted by CIB operations remain on our services.

Four Pillars of Countering Adversarial Threats

We tackle adversarial threats through four key pillars.

- **Building Platform Defenses** to make our systems as resilient to adversaries as possible. Hardening our core and scaled defenses – from traditional cybersecurity (e.g., access controls) to sophisticated integrity systems (e.g., Facial Recognition technology and AI-enabled detection) and evolving our policies to address new threats.
- **Empowering Users** to help keep themselves safe across our platforms and the broader internet. Providing our users and businesses with the necessary tools, controls, and education to safeguard themselves and their customers.
- **Disrupting and Deterring** the most adversarial threat actors, both on our platforms and more broadly. Deploying threat disruption to identify and remove the entirety of adversarial networks through a combination of deep, expert investigations to detect novel behavior and improvements in scaled detection to keep bad actors off of our services.

- **Cross Society Efforts** across the defender community—including industry peers, governments, and law enforcement to share actionable signals and publicize our findings in order to disrupt adversaries wherever they operate.

Fraud and Scams

Financially-motivated criminal actors are among the most persistent and agile threats we face online. These networks operate across platforms, exploit increased internet dependency, and often operate from jurisdictions with limited rule of law to avoid the risk of arrest or other offline consequences.

The Fraud Attack Chain

The **Fraud Attack Chain** outlines the full life-cycle of actions scammers take to drive scams across the internet. We use it to organize our defense and mitigation strategy against scammers, and to effectively partner with other defenders across society:

- **Build Infrastructure:** Scammers acquire physical assets (e.g., IT equipment, physical locations) to run their operation. This is where law enforcement and governments are uniquely positioned to intervene.
- **Prepare Digital Assets:** Scammers create fake accounts, compromise accounts, and establish a facade of legitimacy online. Online platforms often focus on scaled, automated disruption at this stage.
- **Engage:** Scammers initiate contact and build rapport with potential targets, including through ads or messages. Scammers attempt to move targets to off-platform messaging services.
- **Execute:** Scammers convince the victim to part with money or information. The financial services and banking industries have the clearest visibility here, underscoring the need for intelligence-sharing between platforms and financial partners.
- **Clean Up:** Scammers cover their tracks, moving money and recycling infrastructure for future use.

Ways We Are Countering Scams

Platform Defenses and Empowering Users

We highlight some of our new products to combat scams, including advanced scam detection and warnings on Messenger for suspicious chats, and facial recognition technology to defend against celebrity impersonation in scam advertisements.

- **Facial Recognition Technology:** We use facial recognition technology to defend against celebrity impersonation in scam advertisements, and have onboarded nearly 500,000 public figures to our facial recognition program.
- **Advanced Scam Detection on Messenger:** We are testing more advanced scam detection in chats. When enabled and a new contact sends a potentially scammy message, we warn people and give an option to send recent chat messages for AI scam review. If a potential scam is detected, users get more information on common scams, and we will suggest actions including blocking or reporting the suspicious account.
- **WhatsApp Screen Sharing Warnings:** On WhatsApp, we launched warnings when people attempt to share their screen with an unknown contact during a video call.

Disrupting Scammers

- **Litigation:** We actively combat scammers through litigation by filing lawsuits, sending cease and desist letters and pursuing joint actions with impacted brands.
- **Supporting Law Enforcement:** We support law enforcement efforts to achieve real world actions against scammers such as the disruption actions driven by DOJ's Scam Center Strike Force and recent arrests by the Singapore Police Force.
- **Criminal Scam Syndicates Overview and Case Study:** We provide an overview of Criminal Scam Syndicates, including highlighting Meta's recent contributions to a Criminal Scam Syndicate interdiction in Singapore. We also include a case study detailing removal of a Criminal Scam Syndicate operation originating in Cambodia that impersonated law enforcement and government agencies in a "loss recovery" scam which sought to re-victimize scam victims.

Cross Society Efforts

- **Signals Sharing:** We are partnering with external stakeholders to share and ingest signals to improve scam detection and mitigation efforts throughout the defender community. We detail several of our signals sharing efforts, including those occurring at the bilateral and cross-sector level.
- **Collaborating across defender communities:** Meta is a member of organizations like the Global Anti-Scam Alliance (GASA) and Global Signals Exchange (GSE), which help drive coordinated action against scammers.

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.

Case Studies

As part of our work to counter CIB, we provide detailed case studies on key cases that highlight substantial adversarial shifts, new tactics, or important new insights.

Persistent Iranian Influence Operation ("Endless Mayfly")

We are updating our analysis of a long-running influence campaign, known as "Endless Mayfly," to attribute this operation to **Iran's International Union of Virtual Media (IUVM)**, a US-sanctioned propaganda group with close ties to the Iranian government.

This multi-year operation, spanning from 2013 to activity targeting the 2024 US elections, has shown persistent use of sophisticated technical evasion techniques.

Russian Use of Authentic Operators in Sub-Saharan Africa (SSA)

We have identified a proliferation in Russian use of seemingly unwitting local individuals to execute influence operations in Africa on behalf of Russia-based actors.

Two recent cases **targeting multiple African** countries involved freelancers running Pages that masqueraded as local media, promoting Russian geopolitical interests and denigrating African partnerships with France and the US.

Other Cases

We also provide case summaries of five other CIB operations we have removed, including **operations targeting Poland, India, and Moldova.**

Adversarial Threats in the AI Space

We are proactively monitoring and preparing for the risks associated with the rapid adoption of generative AI. In this section, we focus on three main themes: how adversarial actors use AI to violate our Community Standards, how we use AI to defend our platforms from adversarial abuse, and how we are seeking to secure our AI models from adversarial interference. While AI lowers the barrier for entry for threat actors by increasing their efficiency and helping scale operations, our behavioral and technical defenses against these threats remain robust. Further, AI-enabled defenses are helping platforms harden their defenses against all threat actors, AI-enabled or otherwise.

How Threat Actors Use AI

We have observed threat actors attempting to use AI to violate our platform policies. We discuss efforts by influence operations actors to use AI to create fake news personas and craft tailored articles, as well as Russia-linked actors attempting to use data poisoning to bias AI training data.

We also discuss how scammers are using deepfake technology to impersonate celebrities and using AI to create job scam postings and facilitate romance scams.

AI for Defenders: Defense at Scale

We are integrating highly capable AI into our security stack to defend against threats and empower the broader defender community:

- **AI for Integrity:** We provide an overview of how Meta uses AI to improve our scaled integrity detection systems, including removal of fake accounts, notifying users of suspicious interactions, and improving our behavioral detections.
- **AI for Security:** We use AI-driven tools to secure our internal systems, including technologies to prevent insider threats (e.g., sensitive document classification) and ensure our codebase remains secure.

Model Protections

To defend against threats to our AI models themselves, we have implemented layered guardrails, including:

- **Llama Firewall:** A security guardrail tool that detects and prevents risks like prompt injection and insecure code in AI systems.
- **Rule of Two for AI Agents:** A framework designed to help developers navigate tradeoffs between high risk properties of agentic models (untrustworthy input processing, sensitive data access, or external communication/system changes) to prevent the highest impact consequences of prompt injection.
- **Continuous Red Teaming and Automated Adversarial Testing:** We employ pre-launch threat modeling, automated red teaming, continuous risk monitoring and assessments, and engagement with internal and external specialists for both agentic and non-agentic models.

We are committed to sharing findings, tools, and best practices (e.g., open-source releases of defensive models) with government, industry, and the research community to build a collective defense against these shared adversarial threats.

INTRODUCTION

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

Over the years, our reporting has served as a valuable resource for industry partners, researchers, and security professionals seeking to understand and counter adversarial threats across the internet. We have consistently shared threat indicators, tactical analysis, and detailed case studies to support broader defender community efforts to detect and disrupt influence operations.

However, we recognize that the evolving threat landscape and diverse security challenges across the internet require us to adapt our public reporting. With this new, restructured report, we are building on this effort by covering additional threat areas, including spotlights on Fraud and Scams and AI-enabled threats.

We are also adjusting the publication frequency from a quarterly to a twice-yearly cadence. This new format will enable us to focus on deeper analysis and better inform collective defense against threats that require a whole-of-society approach. We're adding new areas because threat actors continue to evolve, expanding their techniques and approaches. Adversarial threats are more complex and sophisticated than ever before: they operate across the internet, leverage multiple platforms and services, and require coordinated responses that extend beyond any single company's efforts. They also increasingly leverage artificial intelligence to enhance their operations and employ novel techniques to obscure their identities and manipulate our systems.

Finally, we are also updating our Inauthentic Behavior Community Standards to simplify and refine our IB and CIB policies and help uninvolved authentic communities, Pages, and Groups that are targeted, managed, or co-opted by CIB operations remain on our services.

These evolving challenges underscore that tackling adversarial threats requires a whole-of-society approach. While we remain deeply committed to protecting our users and maintaining the integrity of our platforms, we also recognize the importance of working collaboratively to help others strengthen their own defenses and increase awareness of emerging threat trends. We hope this report and our ongoing threat signal sharing will contribute meaningfully to these collective defense efforts.

The adversaries detailed in the following report constitute advanced persistent threats, which employ constantly evolving adversarial tactics and bespoke methodology to counter both our response, and the response of the Defender Community wholesale. However, you'll notice a consistent trend in our approach to tackling these adversarial actors across the surfaces and harms where they manifest. We've distilled these into four pillars that form the basis of our approach:

- **Building Platform Defenses** to make our systems as resilient to adversaries as possible. Hardening our core and scaled defenses – from traditional cybersecurity (e.g., access controls) to sophisticated integrity systems (e.g., Facial Recognition technology and AI-enabled detection) and evolving our policies to address new threats.
- **Empowering Users** to help keep themselves safe across our platforms and the broader internet. Providing our users and businesses with the necessary tools, controls, and education to safeguard themselves and their customers.
- **Disrupting and Deterring** the most adversarial threat actors, both on our platforms and more broadly. Deploying threat disruption to identify and remove the entirety of adversarial networks through a combination of deep, expert investigations to detect novel behavior and improvements in scaled detection to keep bad actors off of our services.
- **Cross Society Efforts:** Working across the defender community —including industry peers, governments, and law enforcement to share actionable signals and publicize our findings in order to disrupt adversaries wherever they operate.

Each type of adversarial threat presents unique challenges. While the four pillars serve as a consistent frame for our approach, we develop tailored mitigations to address the specific nature of each harm. Throughout this report, readers will see elements of these pillars highlighted—whether discussing our broad efforts to engage external partners to combat Fraud and Scams or our focus on developing strong internal protections against threats to our AI models. We intend to use these pillars to serve as a consistent methodology across future threat reporting and in turn, provide the Defender Community with a novel framework in which other efforts can mirror.

We expect the make-up of these reports to continue to evolve in response to the changes we observe in the threat environment, and as we expand to cover new areas of our security work. This report is not meant to reflect the entirety of our security enforcements, but to share notable trends and investigations to help inform our community's understanding of the evolving threats we see. We welcome ideas from our peers across the defender community to help make these reports more informative.

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

01

FRAUD AND SCAMS

This section covers our insights and approach in combating scams and the adversarial networks behind them. We detail our Fraud Attack Chain, the role that we play as part of the global defender community, address Criminal Scam Syndicates, and recent product developments designed to protect people using our platforms.

The people, networks, and organized criminal groups behind fraud and scam operations are among the most persistent and global agile adversarial threats we and our defender community peers combat. Driven primarily by financial gain rather than ideology or state backing, these networks often operate from jurisdictions with evolving institutional capacity, making them less responsive to traditional deterrents like technical defenses, and legal action. The scammers targeting people online and on our platforms range from relatively unsophisticated, low-level actors to highly sophisticated organized criminal enterprises.

Scammers distribute their activities across multiple platforms and services to increase their resilience to disruption. They also frequently use various online services to both identify targets and carry out the actual theft of money. That's why we are sharing our insights on combating fraud and scam operations with the broader defense community. We hope our findings will foster a unified approach to deterring and defending against scams that span multiple platforms.

Criminal enterprises are exploiting internet dependency at an increased rate. This practice accelerated during the COVID-19 pandemic when traditional in-person criminal activities decreased due to lockdowns. As cybercrime surged, it became a more lucrative and less risky venture for criminal networks.

Identifying those responsible for online crimes and coordinating international efforts to hold them accountable can be complex and time-consuming. As a result, immediate consequences such as arrest and public exposure- which are common in physical crimes- are rarely a primary concern for individuals engaged in online scams. This emboldens criminal organizations and individuals involved in online scams.

This is why we actively support law enforcement agencies worldwide to identify and prosecute the people and organizations responsible for scam-related crimes, and take legal action to hold scammers accountable. Just this month, the Department of Justice recognized Meta's support for its new Scam Center Strike Force launched to go after crypto investment fraudsters. We also joined the National Elder Fraud Coordination Center (NEFCC) which aims to gather intelligence from its members to build referrals for law enforcement agencies to investigate. Over the last few years, we

have filed more than 60 lawsuits and issued over 2,000 cease and desist letters to individuals and entities that target our platforms and users with abusive schemes. These include celeb-bait ads, brand impersonation, account takeovers, subscription scams, bait-and-switch tactics, fake engagement, unauthorized surveillance, and other unlawful conduct. We have pursued joint litigation with companies such as Gucci, and recently prevailed in separate cases involving ad abuse and attempts to circumvent our enforcement measures.

Transnational Criminal Scam Syndicates are among the most sophisticated scam networks. Criminal Scam Syndicates leverage cutting-edge technology to evade detection and often operate scams across multiple platforms for extended periods, even longer than what we've observed in some state-backed influence operations. But scammers don’t just exist in these organized compounds. Sophisticated individuals and groups often conceal themselves as legitimate businesses, mimicking the very users and advertisers we aim to protect. These groups exploit our advertising policies through methods such as AI-enabled translation services, false product or celebrity endorsements, scripting, and other techniques akin to those used by influence operations and espionage networks.

We organize our efforts to combat scams around the Fraud Attack Chain - an internal model we first developed two years ago, and shared publicly early last year to explain the full life-cycle of fraud and scams. The Fraud Attack Chain has been integral to helping us identify where we may be able to take action against scammers, support the work of other stakeholders, and where we may need to rely on broader community efforts.

Introducing the Fraud Attack Chain

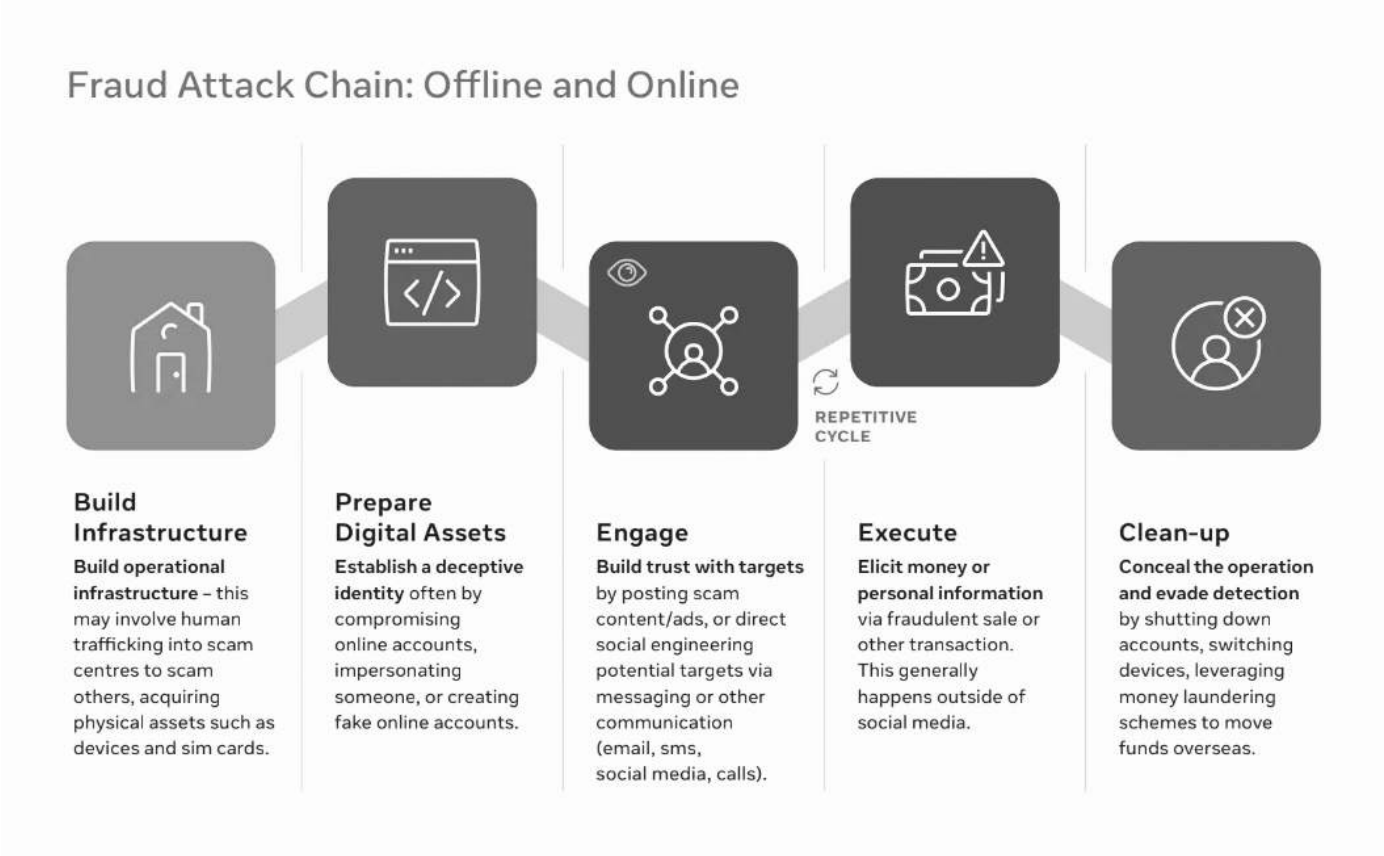


Image: An overview of the Fraud Attack Chain.

The Fraud Attack Chain maps out the tactics scammers rely on - both online and offline - and categorises them according to theme and sequence. This Attack Chain allows defenders to identify effective points to successfully neutralize adversarial, sophisticated scammers, either unilaterally or in collaboration with the wider defender community. Understanding and disrupting the Attack Chain helps defenders substantially increase the cost of running successful scam operations, reduce scammers' margins, and in the best case scenarios completely shut down the abuse.

The Fraud Attack Chain consists of five distinct phases:

1. **Build Infrastructure**
2. **Prepare Digital Assets**
3. **Engage**
4. **Execute**
5. **Clean-up**

The most effective way to combat scams and protect potential targets is to disrupt this activity early in the Attack Chain. This approach also prioritizes the safety of the target, as the scam's

payload is typically delivered at later stages in the chain. Each step in the Attack Chain builds upon the previous one, which means that early interventions increase operational costs for threat actors and reduce potential harm to individuals and communities.

Detailing the chain

1. **Build Infrastructure:** The adversary establishes the foundational physical and scalable operations necessary for their scams. This involves acquiring IT equipment, SIM cards, securing hosting services for websites, obtaining physical locations for conducting scam activity, and in certain cases building criminal networks for scam-adjacent harms such as human trafficking. Due to the upfront cost, this infrastructure is often designed for reuse. At this stage, the online defender community has very limited visibility into a potential attack. However, targeted enforcement at this stage significantly increases the operational costs for scammers, as these physical assets are much more expensive than digital ones. Subsequently, taking action against physical infrastructure often simplifies attribution, heightening the risk of real-world consequences for the scammer or the criminal network. Law enforcement and governments have the most opportunity to intervene during this phase. The online defender community can support through information sharing when platforms identify these early-stage scam networks. We've seen increases in government-led actions against such operations – from sanctions imposed on the operators, to cutting power and access to the internet for these sites.

2. **Prepare Digital Assets:** With the physical infrastructure prepared, scammers move to create or acquire the necessary digital assets to interact with targets across various online platforms and services. This phase involves creating or purchasing fake accounts, compromising legitimate user accounts, and using specialized software to identify and target individuals.

   A key tactic during this stage is the effort to establish legitimacy, often by compromising popular, verified accounts or purchasing fake accounts and engagement.

   As detailed in our public reporting, our detection technology helps us block millions of attempts to create fake accounts every day, often within minutes of creation. These fake accounts are commonly used in scams or to facilitate other violations of our Community Standards. Automated and scaled platform efforts focused on disrupting these digital assets are particularly effective at this stage.

3. **Engage:** In the "Engage" phase, threat actors initiate contact and build rapport with potential victims. This often involves presenting targets with enticing opportunities through various communication channels like posts, advertisements, direct messages, emails or calls, or misrepresenting legitimacy and enticing the target to reach out themselves.

Across the Engage phase we have often seen Adversarial actors try to convince their potential victim to move to another platform or communications provider, making it harder for individual platforms to track the scam.

Whilst online platforms do have the most visibility into the developing scams at this early stage, we know scammers deliberately mirror "authentic" behavior in an attempt to avoid detection and enforcement by the defender community. Without clear signs of adversarial intent, defenders may lack the necessary level of confidence to act against a developing scam. To address this, we've partnered with stakeholders across the Attack Chain- for greater visibility into the whole lifecycle of the scam (which we discuss below). Additionally we have partnered with NGOs, and creators to run PSA-style awareness campaigns, rolled out new counter-scam products (discussed in the New Product Developments section below), and increased the cost of running low-quality ads to deter bad actors.
When we detect scammers engaged in violating behavior, we take appropriate measures, which may include removing content that violates our Community Standards and disabling the scammers' accounts.

4. **Execution:** The scammer then convinces their target to part with their money, information or something of value. This may involve selling counterfeit goods, failing to deliver purchased items, manipulating fund transfers, or stealing account credentials to facilitate further fraudulent activities. Online platforms typically lose any visibility they may have had into the engagement at this phase of the scam, which most often occurs on banking and fintech platforms. That is why we partner across the financial industry through our Fraud Intelligence Reciprocal Exchange (FIRE) program, which broadens the defense community collaboration in combating scams by providing a platform for bilateral information sharing between us and the financial industry. We are also proud partners of the Global Signal Exchange (GSE), the world's first global, multistakeholder and cross-sector clearing house for threat signals.
5. **Clean Up:** In the final phase, the scammer focuses on evading detection and enforcement. This may include scammers blocking victims' accounts or phone numbers, moving money through various accounts, and employing different tactics to conceal their activities and recycle their methods and infrastructure for future use. The defender community typically has minimal visibility into a scammer's activity at this phase.

Connecting Dots Across the Fraud Attack Chain

Different defenders have varied visibility into scam operations across the Attack Chain. For example, financial institutions often have excellent insight at the Execution stage (i.e. when a financial transaction occurs). Additionally, in many countries, financial institutions are legally obligated to require prospective accountholders to provide personal and financial information to open an account, information most online services lack.

Telecommunications companies and domain hosting providers may have unique insight into the Build Infrastructure phase, as SIM cards have to be activated, domains need to be registered, and devices have to come online before a scam can begin. When these signals are shared between service providers and online platforms, companies can use those indicators in order to enable early stage interventions before a scam begins.

Ongoing collaboration with a broad community of stakeholders, including law enforcement, hosting providers, telecommunications companies, financial institutions, and other online platforms is essential to disrupt the physical and financial infrastructure scammers rely on, such as account farms, scam centers, hosting services and providers, bank accounts and credit cards. We have begun working to share scam signals with partners and stakeholders at multiple levels, so we can collectively identify and mitigate abuse across the internet earlier in the Attack Chain.

- **Bilateral:** We exchange signals directly with partners to increase speed and precision of scam mitigation efforts. We partnered with Microsoft (through the Global Signal Exchange) and Google to exchange early signals of potential account compromise and help prevent takeovers of Meta accounts from compromised email addresses. This resulted in earlier detection of potential fraudulent activity, fewer successful account takeovers, and faster recovery for victims. We plan to continue scaling this model, with a goal of cutting off scam infrastructure on both our services and others.
- **Industry and Cross Sector:** At the industry and cross-sector level, we are committed to public-private coordination to help close gaps across the entire Attack Chain. We work with the GSE to exchange threat signals at scale with a coalition of government and industry stakeholders, helping to close the loop between public and private sector defenses. This collaboration serves as a blueprint for cross-sectoral intelligence sharing in the fight against online fraud. We are also a foundation member of the Global Anti-Scam Alliance (GASA) working to protect consumers worldwide from scams. Separately, recent intelligence sharing led to Meta's removal of about 29,000 accounts engaged in job scams in Australian Facebook groups, and 1,850 scam enablers such as websites and scam job advertisements were referred for removal.

Across the online space we see adversarial actors also using advertising tools, including our own, to perpetuate scams. We have stringent policies and detection in place to prevent the misuse of our advertising tools and as of October, we've removed 134 million pieces of ad content this year across Facebook and Instagram due to violations of our Fraud, Scams and Deceptive Practices policies, the majority of which we proactively removed before they were reported to us.

Next, we provide an overview of our efforts to combat the threat posed by Criminal Scam Syndicates, including highlighting recent work tackling those seeking to impersonate government agencies in an effort to "double tap" scam victims. While this case study highlights one specific instance, it represents just a fraction of our comprehensive efforts to counter these adversarial actors.

Criminal Scam Syndicates

One of the most persistent and challenging threats the defender community faces today is posed by Criminal Scam Syndicates. Over the past decade, we have seen the emergence of industrial-scale sophisticated scam compounds and criminal syndicates which perpetrate global fraud. However, these entities have their roots in earlier forms of telecommunication fraud. Research by academic institutions, international bodies like the United Nations Office on Drugs and Crime (UNODC), and investigative journalism pinpoints a significant escalation and consolidation of these operations, particularly in Southeast Asia, beginning in the mid-2010s and accelerating dramatically during the COVID-19 pandemic.

We've previously reported on enforcement actions against these global Criminal Scam Syndicates. But the scale and sophistication of the Criminal Scam Syndicate threat is unprecedented, with the US Institute of Peace estimating that up to 300,000 people are forced into online scam work by these criminal organizations, and with up to USD 64 billion stolen annually from victims as of the end of 2023. These Criminal Syndicates have the means and motive to iterate on their approach, and are adapting to the mitigations put in place by the defender community.

In October, the Singapore Police Force arrested seven members of a transnational criminal scam network, based in part on information Meta shared through a partnership with the FBI. According to the Singapore Police Force (SPF), victims were enticed to transfer money to anonymous bank accounts to purchase gambling credits for an online gaming site, and were eventually informed they would have to purchase additional credits to be able to redeem any winnings they had earned through the site. The SPF reported that victims lost in excess of USD 175,000. This type of partnership between Meta and law enforcement agencies is key to successfully pushing back on Criminal Scam Syndicate operations.

We continue to look for and block attempts by criminal syndicate-run scam centers to create accounts on our platforms.In the first half of 2025, our teams detected and disrupted nearly 12 million accounts — across Facebook, Instagram, and WhatsApp — associated with criminal scam centers, and we are working with law enforcement agencies around the globe to share our insights.

In response to these efforts, criminal networks are evolving and adapting. Facing increased pressure from authorities in Southeast Asia, we've observed these operations expanding to other regions, including Africa and Latin America. They continue to refine their tactics, leveraging advancements in artificial intelligence and deepfake technology to enhance the credibility of their scams. Through our investigations, we've observed Criminal Scam Syndicates expand beyond "pig butchering" romance scams to investment scams, commerce scams, illicit gambling, and impersonation scams, as they adapt to try to evade our detection and continue defrauding victims.

While online platforms such as Meta play a pivotal role in disrupting scam operations and strengthening the cyber resilience of our users, achieving meaningful progress in combating this threat ultimately depends on driving impactful consequences for the criminal organizations behind these schemes.

Scams are a global problem, and scam activity comes from many different sources around the world. Many countries are actively seeking to drive positive change. In these cases, additional support for intelligence sharing, capacity building, and technical assistance are essential to empower local authorities and foster sustainable improvements.

Drawing inspiration from the Financial Action Task Force (FATF) approach to money laundering, which involves publishing lists of countries failing to take adequate action and prompting international penalties, we advocate for the development of a similar framework to address scam proliferation. Such a list would serve as a focal point for coordinated counter-scam strategies and facilitate the implementation of appropriate, targeted interventions.

Importantly, any penalties or enforcement actions must be carefully calibrated to avoid unintended consequences - particularly for individuals who have been coerced or trafficked into scam operations. Human rights considerations must remain at the forefront of any response to ensure that efforts to combat scams do not inadvertently harm vulnerable populations.

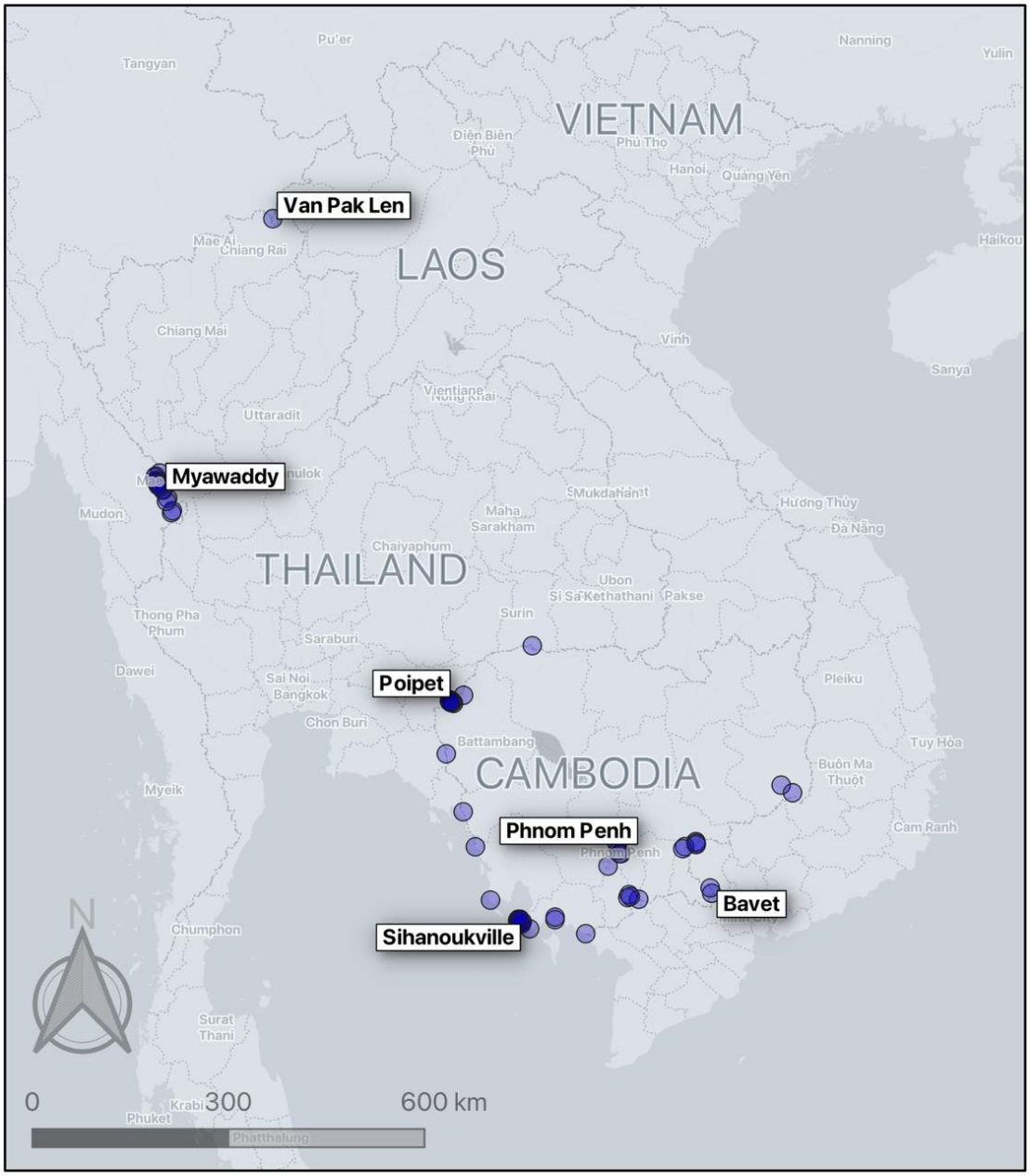


Image: Known Criminal Scam Syndicate locations.

Criminal Scam Syndicates Case Study

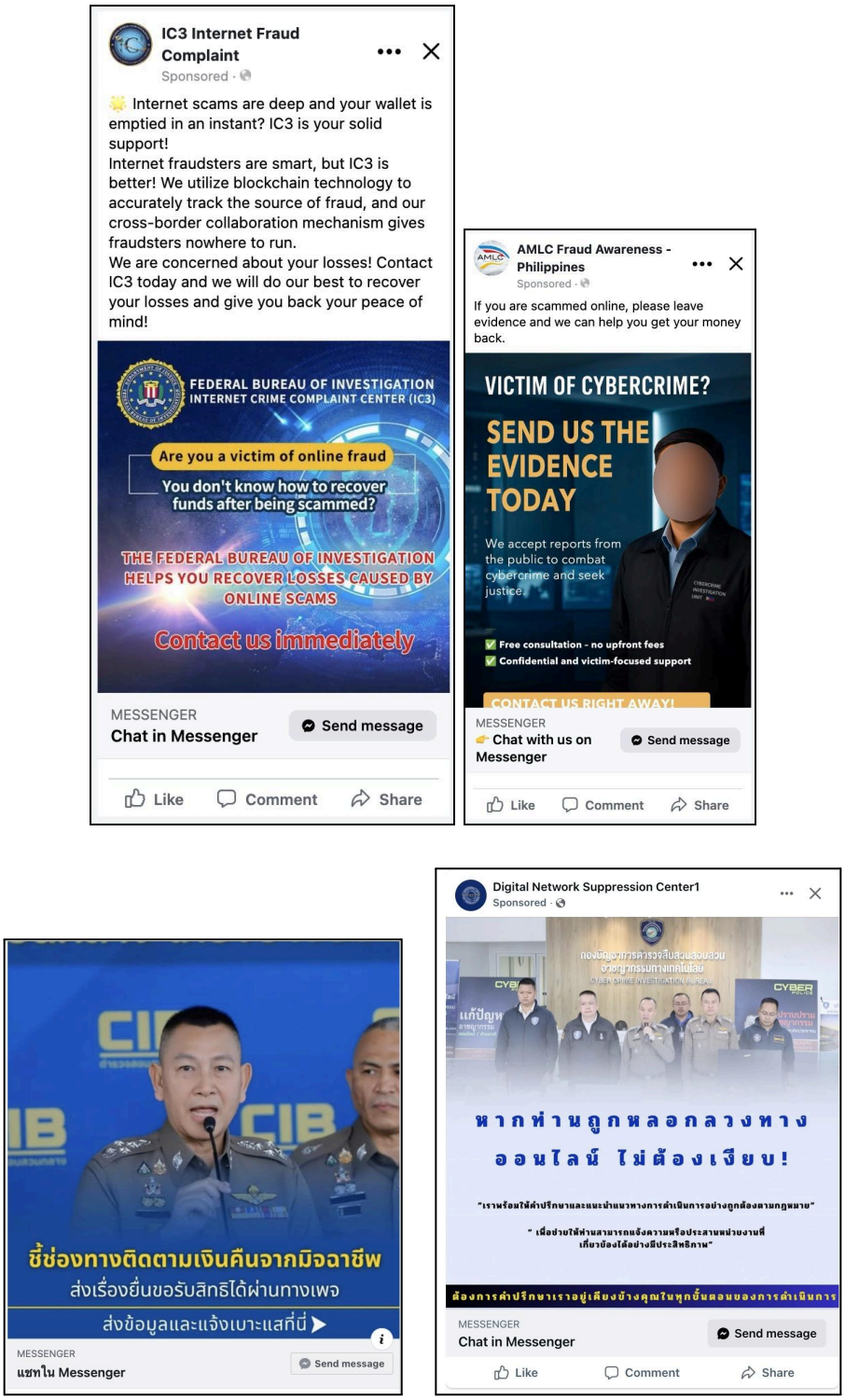
We removed over 6,400 Facebook accounts and Pages between January and October 2025 for violating our policies against Fraud, Scams and Deceptive Practices and Dangerous Organizations and Individuals.[1] This network originated in Cambodia, and impersonated American, Australian, Indonesian, Philippine, Thai, and Vietnamese law enforcement and government officials.

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

This specific case study is a subset of our efforts to disrupt Criminal Scam Syndicates – with a focus on addressing the “loss recovery” or “double-tap” government impersonation scam archetype. This type of scam targets users who have already been scammed once with false promises that they can recover their losses. To enable this scam, Criminal Scam Syndicates create inauthentic profiles that purportedly represent law enforcement, regulators, other government agencies and officials, and advocacy organizations. These profiles claim to investigate scam activity and — for a “fee” — falsely offer to recover losses for victims.

The network frequently impersonated law enforcement agencies from the U.S. or Australia, or regional agencies in East Asia or Southeast Asia (e.g., Indonesia, Philippines, Thailand, Vietnam). These accounts use logos and media that attempt to represent official government or law enforcement branding, including seals, insignia, badges, and personnel imagery. In some instances, they use official-sounding but fictitious names like "Cyber Crimes Investigations Service" or "Counter Fraud Alliance."

They solicit scam victims by posting publicly or by targeting Facebook Groups, and online forums where victims are looking for support. Ultimately, the network directs victims to engage with their "investigators" across messaging services. There, they request various fees, such as application fees, expense reimbursements, or transfer fees, under the pretense of conducting an investigation or returning recovered funds, none of which are ever delivered.



Images: Examples of specific government impersonation scams.

Images: Examples of non-descript or fake government "counter-scam" content.

We found this activity as a result of our investigations into Criminal Scam Syndicate activity, government reports, and user reports. We used a combination of technical and behavioral signals to identify commonalities used by these networks and known Criminal Scam Syndicates, particularly scam compounds located in Cambodia. We remove this scam content and violating accounts on an ongoing basis as part of our work to counter Criminal Scam Syndicates. We also work with governments and other authorities to protect their brands.

New Product Developments

In addition to deep dive investigations to counter the most sophisticated scammers, we use the insights from our investigations to develop specialized detection tools, automated mitigation systems, and behavioral analysis frameworks designed to identify and disrupt fraudulent activity

before it can reach our users. By combining strong internal safeguards with increased cross-industry information sharing, we aim to create a robust ecosystem that not only protects our platforms but contributes to the broader fight against adversarial fraud and scam operations targeting users across the internet. Some of our key product innovations to counter scams are:

1. **Scam Detection on Messenger:** We are testing more advanced scam detection in chats. When this is enabled and a new contact sends a potentially scammy message, we warn users and give them an option to send recent chat messages for AI scam review. If a potential scam is detected, we will provide the user with additional information on common scams, and provide suggestions for blocking or reporting the suspicious account. As mentioned above, a common tactic amongst scammers involves building rapport with their potential victims to maximize their ability to extract money. Across the internet, scammers work to establish initial contact with a potential victim, before quickly attempting to move these individuals to messaging platforms, where rapport-building can continue. We see this across multiple scam types, including romance scams, customer service scams, or long-tail investment scams.

2. **Facial Recognition:** Adversarial actors seeking to scam our users are attempting to impersonate real-world celebrities, political figures or influencers in scam advertisements. We have introduced facial recognition technology to increase our ability to protect our users from these fraudulent product endorsements. Today, there are nearly 500,000 public figures that are being protected from having their likeness misused in these scams. The expansion of facial recognition technology in particular more than doubled the volume of celebrity-bait scam ads we were able to detect and remove in testing. In the first half of 2025, user reports of celebrity-bait ad scams, out of total ad impressions, dropped by 22% globally.

3. **WhatsApp Screenshare Warnings:** On WhatsApp, we launched warnings when people attempt to share their screen with an unknown contact during a video call. We know scammers may pressure their targets to share their screen to trick people into giving away sensitive information including bank details or verification codes. With this new tool we give our users more context to spot and avoid scams.

02

COORDINATED INAUTHENTIC BEHAVIOR

What is Coordinated Inauthentic Behavior (CIB)

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

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

Key Insights

This report details our disruption of multiple covert influence operations that violated our Coordinated Inauthentic Behavior policy. We provide updated attribution analysis connecting a series of persistent Iranian government-linked influence operations that have repeatedly attempted to establish networks on our platforms using advanced evasion techniques. We also examine the evolving tactics, techniques, and procedures (TTPs) employed by Russian operators in two newly discovered campaigns targeting African audiences, including their strategic use of likely unwitting local individuals to run influence operations in an attempt to enhance operational legitimacy and reach.

Additionally, we share case summaries on five other covert influence operations originating from Poland, Belarus, Russia, India, and Moldova 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.

Updating Attribution of Persistent Iranian Influence Operation to "Endless Mayfly"

Iranian threat actors continue to violate Meta's Coordinated Inauthentic Behavior policy, trailing only behind Russia for the most disruptions since we began threat reporting in 2017. As part of our regular updates on this activity, today we are sharing key insights and attribution updates for a long-running campaign known in the industry as Endless Mayfly. This set of linked networks spans from our first Iranian CIB disruption in 2018 through off-platform activity targeting the 2024 US elections.[2]

Our investigation has compiled multiple, corroborating lines of evidence from both internal investigations and public sources that attribute this multi-year operation to Iran's International Union of Virtual Media (IUVM), a sanctioned propaganda group with close links to the Iranian government.[3] This report sheds light on their persistent attempts to covertly run influence campaigns over the last decade. We hope this information contributes to the public understanding of this threat and provides the security community with context and data to anticipate, identify and respond to future campaigns.

Overview & TTPs

In 2018, we announced our first Iranian CIB takedown of the Liberty Front Press (LFP) network. This activity began in 2013, with operators primarily posing as news and civil society organizations. Over the last seven years we have continued to track, investigate, and disrupt operations evolving from this initial set, identified as "Endless Mayfly" by CitizenLab. We are able to link numerous networks together through both technical and behavioral indicators.

The operation's content aligns with Iran's foreign policy objectives and they consistently run campaigns targeting the US, France, Israel, UK and Iran's regional interests. This network's TTPs have evolved over the years, but certain hallmarks are notable throughout.

First, they operate domains on which they place misleading articles in an attempt to deceive users, directing traffic to them from social media channels. Often, these domains cross-amplify their content and frequently appear in Iranian state-owned media outlets. In earlier days, these domains often typosquatted[4] popular news websites, but later evolved to bespoke, independent outlets covering issues of interest to the target demographics.

[2] As reported by our peers at Microsoft and OpenAI.
[3] IUVM is sanctioned by the US Government for being owned or controlled by the Islamic Revolutionary Guard Corps (IRGC).
[4] Typosquatting is a type of social engineering attack which tricks users into visiting malicious websites with URLs that are common misspellings of legitimate websites.

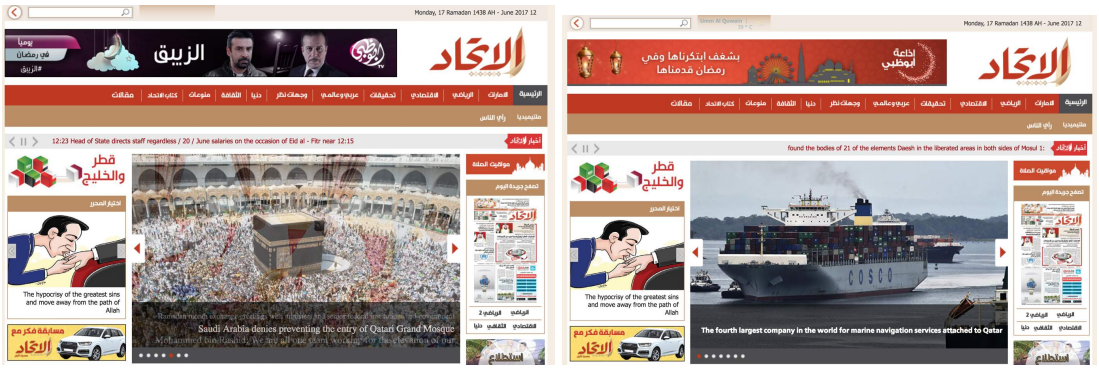


Image: Early example of typosquatting: site alettehad[.]net versus authentic site alittihad[.]ae

Image: Recent domain example: 7sabah[.]tr[.]com targets Turkish audiences

Image: Recent domain example: Amerikagozlemi[.]com targets Turkish audiences with US-related narratives

Second, the threat actors often impersonate journalists or masquerade as students, creating social media personas to match corresponding bylines on their website articles. We have often observed them contacting authentic outlets and journalists in attempts to launder their narratives into mainstream press, even recycling personas across campaigns years apart. However, these efforts have been largely unsuccessful.

Third, the actors use consistent infrastructure and unique technical TTPs. This campaign demonstrated consistent, long-term reliance on a small Iranian hosting provider, and a limited range of primary IP addresses. In an apparent effort to obscure previously identified infrastructure, many websites within this network simultaneously moved from a long-used set of hosting IPs to a new set in early 2024. This coordinated migration inadvertently highlighted the interconnected nature of these assets. Since then, we have noted a continued diversification of hosts, likely in further attempts to avoid tracking. See our Github repository for more detailed information.

Additionally, operators employed consistent elements in their recent WordPress-powered domains to include recycled analytics tags, common plugins, and consistent footer formats.

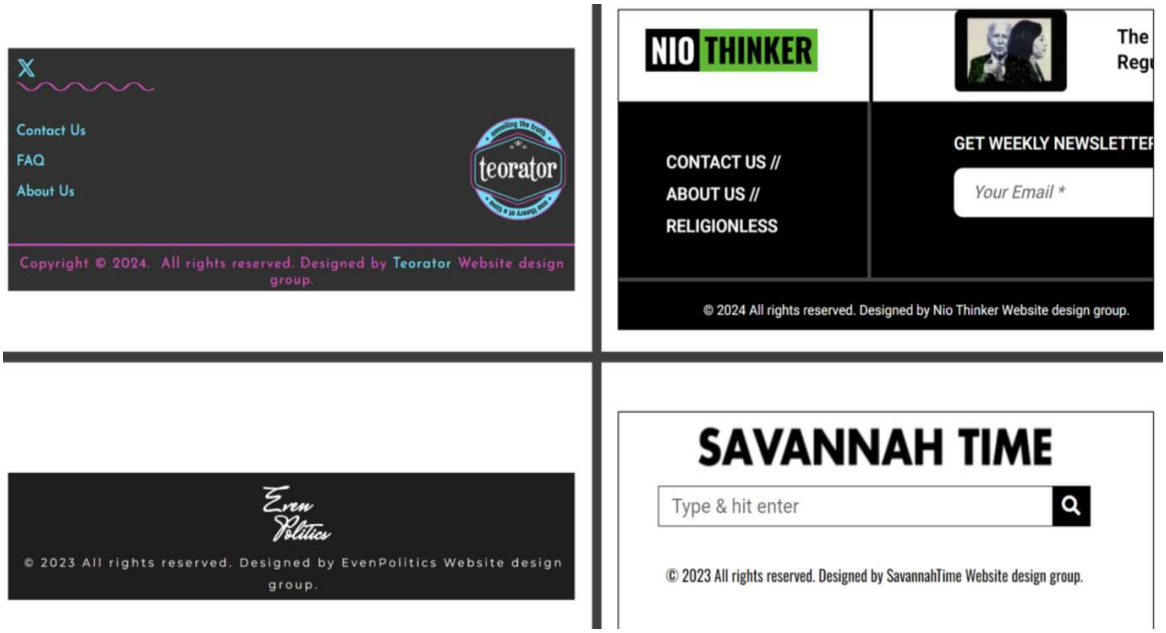


Image: Sample of US-targeting domains in 2024.

Public searches revealed this pattern to be exceedingly rare, and it surfaced an Iranian web designer who was the operator of a core cluster in the original LFP disruption nearly ten years prior. Consistent signatures demonstrated the full scope of the operation over the years, confidently linking subsequent campaigns. The technical connection of this original LFP operator to websites targeting the 2024 US election further reinforced our findings.

Takeaways & Efficacy

Our ongoing detection and enforcement efforts against this network are working. Constant pressure by the collective security community has forced IUVM to burn useful personas, and slow their pace of operations by forcing them to reconstitute on new infrastructure. Our defenses have successfully reduced the traction these operations gain on our services, with recent campaigns abandoning our platforms altogether. By disrupting their personas, domains, and amplification networks, we diminish their ability to build organic audiences and scale widespread impact.

We are publishing this update and data in hopes that new information will contribute to the community's ongoing understanding and defenses against these covert influence operations. Their track record demonstrates that IUVM is likely to continue to evolve their tactics in an attempt to evade detection and deceive audiences. This underscores the critical need for the entire defender community—including researchers, industry, and government agencies—to continue collaborating. By working together to detect, analyze, and publicly expose these evolving tactics, we can collectively increase the consequences for these malicious actors and ensure their harmful effects are minimized.

Russian Use of Authentic Operators in SSA

Proliferation of Local, For-Hire Influence In Sub-Saharan Africa

Influence operation threat actors continue to adapt in response to our detection and enforcement efforts. In recent months, we've observed increasing use of freelance social media managers to execute influence operations in Africa on behalf of nation-state actors. Unlike commercial firms offering "IO-for-hire" services that we reported on in the past, these freelancers are often social media managers based in the same locale as the target audience. They use their authentic accounts to run social media for brands and promote content, and we believe they are unwitting of the true nature and sponsorship of the campaigns they are running.

Over the last six months, we have uncovered two campaigns targeting Africa in which local, for-hire individuals conduct influence activities on our platforms, likely on behalf of Russia-based actors. While we have previously documented Russia's use of local cutouts in Africa as early as 2019, recent investigations reflect a proliferation of this tactic and a decreasing reliance by Russian networks to use strictly fake accounts. This evolution is a departure from typical influence operation models, where campaigns are run directly by commercial firms or nation-state actors through networks of inauthentic personas.

In most cases, we have no evidence to suggest that the for-hire individuals involved in these operations were aware of the ultimate Russian backing. Despite this, their activities - particularly their attempts to circumvent political ad detection and use of fake accounts to continue operating their campaign after initial enforcements - violate Meta's Coordinated Inauthentic Behavior policy and resulted in their removal from our platforms.

Beyond perceived anonymity, leveraging local, for-hire individuals enables threat actors to minimize their online footprint and outsource phases of the online operations kill chain that are most vulnerable to platform detection, such as acquiring and disguising assets. By hiring locals that specialize in digital media management, influence actors effectively obtain local infrastructure, mature and authentic social media accounts, and target-specific knowledge, all of which facilitate the operation's ability to integrate smoothly into the local information environment. This tactic also facilitates longevity - another phase in the online operations kill chain - as we've observed threat actors quickly replace social media managers when previous ones are removed, allowing campaigns to continue with minimal downtime.

As platforms continue to improve detection of inauthentic behavior, we assess that influence actors could increase their use of paid intermediaries for influence operations. This report aims to increase awareness of this issue, enabling digital freelancers to critically evaluate who they're working for and thereby help them avoid disruptions to their legitimate income streams. The following sections provide details on two recent investigations that use this tactic.

Case Study 1: Freelancers Drive Coordinated Political Influence

We removed 67 Facebook accounts, 70 Pages, and 2 Groups for violating our policy against Coordinated Inauthentic Behavior. About 621,400 accounts followed one or more of these Pages, and about 4,200 accounts followed one or more of these Groups. The network engaged in around $107,300 in spending for ads on Facebook and Instagram, paid for mostly in US dollars, South African Rands, and Nigerian Nairas. This network targeted multiple countries throughout sub-Saharan Africa and leveraged local freelancers, likely working on behalf of individuals in Russia. The operation created Pages purporting to be local and original media outlets and ran ads that denigrated African partnerships with France and the United States and promoted Russian geopolitical interests.

A decentralized network of individuals, claiming to be social media managers, drove this operation across primarily Mali, Burkina Faso, Gabon, Nigeria, South Africa, Senegal, Angola, and Benin. These individuals, some of which maintained active profiles on freelance platforms like Upwork, stated they were social media managers or search engine optimization specialists, for example.

On our platform, these individuals created Facebook Pages that masqueraded as local media or news outlets and ran undeclared political ads that typically targeted multiple African countries. The individuals behind these Pages repeatedly returned to our platform after we removed them, often using fake accounts to recreate similar brands. This network of individuals operated independently, though they likely received centralized tasking through channels outside of Meta's platform. This assessment is supported by our observations of these individuals running the same or slightly adjusted ad content - often within a two day window - and utilizing the same text obfuscation methods, such as "4ng0la, "uk_ra_ine", and "fr@nce".

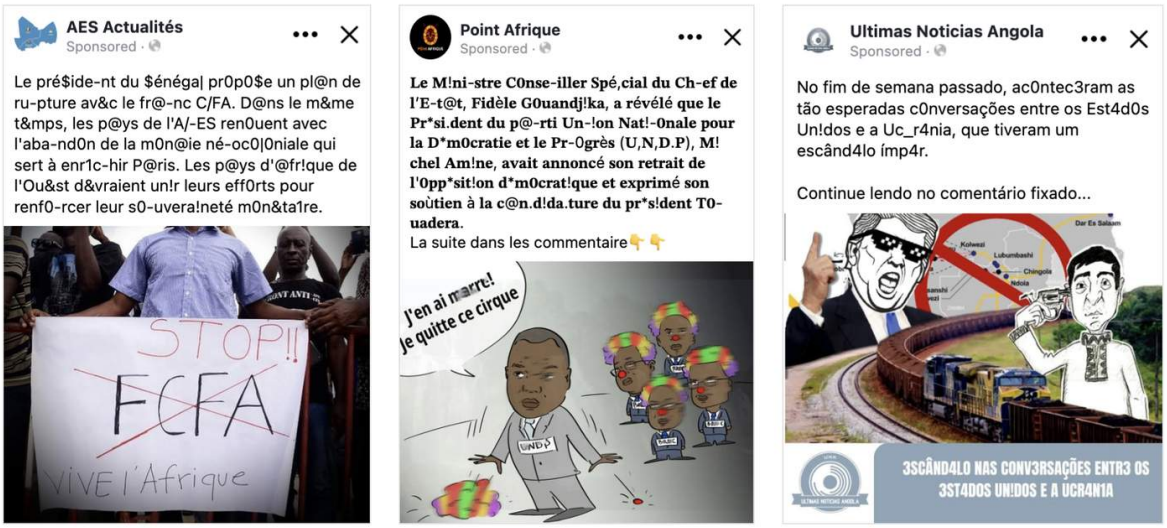


Image: Examples of obfuscation tactics employed by inauthentic Facebook Pages

To a lesser extent, the operation leveraged established, authentic media outlets on Facebook that offered paid advertisements, typically including “DM for ads” or similar language in their Page descriptions. These advertisements featured similar political themes, text obfuscation methods, and coordination indicators as the broader network.

Image: Examples of advertisements run by established Facebook Pages

We found this network as a result of our internal investigation into suspected coordinated inauthentic behavior in the region. Our analysis benefited from reporting produced by France’s Vigilance and Protection Service against Foreign Digital Interference (VIGINUM).

Case Study 2: RT-Linked Covert Campaign Operated by Africa-Based Individuals

We removed 10 Facebook accounts, 2 Pages, and 9 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 20,800 accounts followed one or more of these Pages, and about 3,100 accounts followed one or more of these Instagram accounts. This network targeted numerous sub-Saharan African countries and leveraged a Cameroon-based freelancer, likely working on behalf of employees of RT, a Russian state-controlled media entity.

The individuals behind this activity attempted to create two seemingly independent, grassroots media outlets, establishing a presence across multiple internet platforms including Telegram, TikTok, Facebook, Instagram, and X. On our platform, the brands - Allô Afrique and Derniere Minute - posted in French and targeted audiences with international news, seeding narratives on Russia's stance on Ukraine and the West, and at times, shared RT-branded content. These brands also posted videos periodically featuring RT employees without acknowledging their affiliation to the Russian-controlled media outlet.

For the majority of time this campaign ran, a Cameroon-based individual operated the Facebook and Instagram accounts associated with these brands. This individual was associated with a Page advertising the individual's own digital communications agency. We assess it is possible the operator had knowledge of the ultimate sponsor of this operation.

We found this activity as a result of an internal investigation into suspected coordinated inauthentic behavior in the region and identified links to a past influence operation we removed and reported in November 2023.

Other Cases

01 Russia

We disrupted a foreign interference operation originating in Russia and targeting Moldovan audiences. We removed 5 Facebook accounts, 2 Pages, and 3 Groups for violating our policy against Coordinated Inauthentic Behavior. About 146 accounts followed one or more of these Pages, and about 1 account followed one or more of these Groups. Network operators engaged in around $24 in spending for ads on Facebook and Instagram, paid for mostly in US dollars, to amplify their content within Moldovan political discourse. Our internal investigation revealed direct links to individuals based in Russia, confirming this as a foreign-directed influence campaign targeting Moldova's domestic political landscape.

We observed that network operators consistently promoted messaging favorable to the Moldova Mare political party while targeting content critical of the ruling Party of Action and Solidarity (PAS) government. The operation demonstrated sophisticated media impersonation tactics, creating a multi-platform ecosystem of assets with professional branding designed to masquerade as authentic news outlets. Network operators employed advanced persona development

techniques, using fake Facebook accounts to establish credible backstories for fictional reporters whose bylines appeared on their off-platform website. This cross-platform approach demonstrates complex foreign influence tactics, blending traditional media branding with attempts at social media manipulation.

02 Belarus

We disrupted a coordinated inauthentic behavior network originating in Belarus and targeting Polish audiences. Our internal investigation revealed links to Belarus and Russia, indicating a coordinated foreign influence campaign. We removed 4 Facebook accounts, 12 Pages, and 21 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 200 accounts followed one or more of these Pages, and about 3,300 accounts followed one or more of these Instagram accounts. Network operators had around $1800 in spending for ads on Facebook and Instagram, paid for mostly in Polish zlotys and US Dollars, to amplify their content and expand their reach with targeted audiences.

We observed that network operators strategically disseminated messaging focused on Poland's immigration policies and the country's relationships with the European Union and Ukraine. The operation employed sophisticated impersonation tactics, with fake accounts running targeted advertisements while falsely representing Poland's governing Civic Coalition, and promoting content about Coalition policies that were deliberately designed to appear controversial to opposition audiences. Additionally, the network actively amplified materials from a targeted hack-and-leak campaign against a Polish European Parliament member, demonstrating the integration of cyber-enabled information operations with social media manipulation tactics.

03 Poland

We disrupted a coordinated inauthentic behavior network originating in and targeting Poland. We actioned 55 Facebook accounts, 36 Pages, 23 Groups, and 1 Instagram account for violating our policy against Coordinated Inauthentic Behavior. About 49,000 accounts followed one or more of these Pages, about 1,100 accounts followed one or more of these Groups, and about 2,900 accounts followed one or more of these Instagram accounts. The network did not engage in paid advertising, instead relying on organic content amplification strategies to reach target audiences. Our investigation found direct links to an individual based in Poland, indicating a domestic operation seeking to influence local political conversations. We found this network following an internal investigation that identified sophisticated deceptive tactics designed to manipulate domestic political discourse.

We observed that network operators consistently amplified narratives critical of Warsaw Mayor Rafal Trzaskowski and the current Polish government while promoting content favorable to the Polish Law and Justice (PiS) Party. The network employed sophisticated persona development tactics, creating fake accounts with carefully crafted political identities spanning the ideological spectrum, including both left-wing and right-wing personas as well as accounts focused on historical interests. Operators strategically deployed culturally symbolic profile imagery to enhance

the authenticity of their deceptive personas. These carefully constructed identities were then used to infiltrate and post content within civic groups that aligned with each persona's stated political orientation in an attempt to simultaneously influence multiple users across the political spectrum. Beyond targeting Poland's governing coalition through divisive messaging, the operation also demonstrated personal motivations, with operators attempting to boost their own authentic social media profiles alongside the coordinated inauthentic campaign.

04 India

We disrupted a coordinated inauthentic behavior network originating in India and targeting domestic audiences. We removed 59 Facebook accounts, 11 Pages, and 152 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 102,100 accounts followed one or more of these Pages, and about 74,400 accounts followed one or more of these Instagram accounts. Network operators engaged in around $3600 in spending for ads on Facebook and Instagram, paid for mostly in Indian Rupees. We detected the network following an internal investigation.

The network demonstrated broad use of generative AI technologies, employing AI-generated profile pictures and biographies to create convincing fake personas, while also utilizing artificial intelligence to produce images, captions, and comments throughout the campaign. To enhance the authenticity of their deceptive personas, operators employed sophisticated social engineering tactics, frequently posting non-political content that mimicked genuine user behavior, including discussions about hobbies, academic pursuits, and daily activities. The operation also maintained a dedicated cluster of fake accounts specifically designed to artificially boost engagement on posts from the primary network assets, demonstrating a multi-layered approach to audience manipulation and organic reach amplification.

05 Moldova

We disrupted a domestic coordinated inauthentic behavior network originating in and targeting Moldova. The operation maintained a cross-platform presence, and we removed 15 Facebook accounts, 6 Pages, 7 Groups, and 4 Instagram accounts for violating our policy against Coordinated Inauthentic Behavior. About 6,300 accounts followed one or more of these Pages, about 27,400 accounts followed one or more of these Groups, and about 119,400 accounts followed one or more of these Instagram accounts. Network operators engaged in around $300 in spending for ads on Facebook and Instagram, paid for mostly in US Dollars. Our internal investigation revealed direct links to individuals associated with the Gagauz Executive Committee in Gagauzia, Moldova.

We observed that network operators systematically amplified messaging supportive of the Heart of Moldova political party through coordinated inauthentic tactics. The network engaged in sophisticated grassroots impersonation tactics, creating original video interviews with local residents under the "Voice of Gagauzia" brand which appeared designed to appear as an authentic

initiative while covertly advancing specific campaign messaging. Our investigation suggests that operators may have leveraged Gagauzia state resources to support the influence campaign, including infrastructure associated with the Gagauz Executive Committee. This case demonstrates how regional government resources can be inappropriately used to violate our CIB policies, blending official authority with deceptive social media tactics.

03

ADVERSARIAL THREATS IN THE AI SPACE

As part of our commitment to discussing an expanded range of adversarial threats, we're diving deeper into generative AI and its role in the adversarial threat landscape. We also want to acknowledge the growing adoption of this technology, the rapid pace of improvement in model capabilities from one release to the next, and what this means for the future of combating adversarial threats on our platforms.

Our threat reports have previously touched on the use of generative AI to facilitate violations of our Coordinated Inauthentic Behavior (CIB) policy. In this section, we seek to broaden that initial discussion to touch upon a wider range of AI-related concerns across threat types. First, we discuss how adversarial actors have utilized AI to facilitate violations of our platform policies and how AI is slowly but surely changing the threat landscape overall. This ATR does not address policy violations of Meta's AI products, but we expect to provide more on this in the future as we expand our detection, enforcement, and investigations. Second, we look at how Meta is using AI-enabled tools to defend our systems and products from adversaries and what the broader defender community can learn from these efforts. Third, we touch on our work to prevent and mitigate adversarial attempts to manipulate model behavior for nefarious purposes.

Adversarial Threats in the Age of AI

The rapid adoption of generative AI has raised concerns among defenders worldwide. At Meta, while we've observed adversarial attempts to use AI to violate our Community Standards, our experience suggests that existing industry defenses remain effective. Many of the current concerns about generative AI center on its potential to enable adversarial behaviors that our industry already understands well.

We have tracked adversarial use of AI ranging from questionable GAN profile photos beginning in 2019 to more recent, highly realistic fake newscasters and impersonation scams, and have yet to see AI enable a step change in threat actors' ability to evade our detections. This is because our enforcements against persistent threats like cyber espionage, influence operations, and fraud and scams have long relied on behavioral and technical signals. Behavioral signals are patterns of user activity and interactions that indicate potential policy violations or harmful intent, independent of the specific content being generated or posted. Despite the rapid growth in AI capabilities even within the past year, we've been able to keep up and continue to detect these operations early in their development.

This section has been scoped to talk principally about how AI is being implemented to further policy violations in the spaces we know best, in line with the historic focus of our threat reporting related to adversarial threats to our platforms. That being said, as we expand the AI products and capabilities we offer, we're investing in detection, enforcement, and investigative capabilities. We continue to iteratively improve our system-level classifiers, explore promising monitoring methods like hierarchical summarization to identify new and emerging harms, and evaluate novel capabilities in line with our Frontier AI Framework.

As we further develop this work and gather more data about attempted misuse of our models, we hope to open the aperture of future reporting to share more detailed case studies and trends related to attempted adversarial use of our AI, beyond what appears downstream on our social media platforms.

How AI is Impacting the Threat Landscape

Although our ability to detect and enforce against adversarial threats on our platforms remains robust, we believe it is important to consider how increasingly powerful AI capabilities may alter aspects of the threat landscape across the internet. Here are some of our key takeaways so far, based on on- and off-platform analysis:

Efficiency & Scale: It's now easier than ever to quickly create text and media content like political cartoons, eye-catching logos, long-form news articles, or comments. AI lowers the barrier for entry for threat actors, allowing them to get more done with fewer resources. Influence operations that historically may have required cartoonists, graphic designers, writers, and translators can now carry out those functions without specialized skillsets and are increasingly relying on well-crafted AI prompts. Because we're still able to catch these kinds of operations early in their development, we haven't seen these efficiency gains lead to greater operational success, but the operators themselves may feel like they're able to do more with less.

Error Reduction & Believability: Historically, incorrect or abnormal grammar was one easy way for the average person to identify fraud, scams, phishing or other types of malicious outreach. Translation errors and unusual phrasing can be a dead giveaway that someone is not who they say they are. AI offers simple and effective translation capabilities that may allow threat actors a leg up in their efforts to deceive, particularly in areas that require direct engagement with victims, like fraud or scam operations.

Model Diversification: Threat actors frequently use multiple companies' models as part of their operations. This could be for operational security purposes, or to test out which models work best for their goals. Either way, the result is the same: it's more difficult for developers to identify behavior as adversarial when it's happening piecemeal across different AI services, a pattern that mirrors the platform diversification we've highlighted in previous reporting. When developers lose visibility into key parts of adversarial behavior, it's harder to determine if someone is crafting a phishing email or just drafting tailored outreach as part of their job. In cases like this, it may actually

be easier to identify harmful activity downstream. It's also why information sharing with industry peers is critical to exposing these campaigns.[5]

New Threat Vectors: So far, we have primarily covered changes to the threat environment across established surfaces like social media, messaging services, and websites. But AI models themselves may also provide a new threat vector. This is why we're closely tracking and building defenses against risks like prompt injection, when an attacker manipulates a prompt given to a model to alter its behavior, and data poisoning, when an attacker manipulates data processed by a model to introduce biases or vulnerabilities. You can find more information about how we're protecting against these risks in the "Securely Deploying AI Models" section below.

How Threat Actors Use AI

Threat actors are wasting no time taking advantage of AI. In the years since we began tracking adversarial use of AI in the influence operations space, we have seen a transition from rare, ad hoc use of AI (GAN profile pictures, AI-facilitated content translations) to many of the adversarial networks we disrupt at least testing the use of AI in some capacity. Here are some current trends we're seeing on our platforms:

- **Covert Influence Operations:** We've seen relatively consistent tactics over the past year, including the use of AI to create fake news personas and brand logos, develop more believable inauthentic personas, craft tailored articles and media, and translate and refine outreach to target audiences. For example, we recently took down a CIB network originating in India that used AI to generate profile pictures and biographies to create convincing fake personas, as well as to produce images, captions, and comments throughout the campaign. Please see the CIB section of this report for our actions on such use of AI.
    - **Data Poisoning:** An adversarial attack where malicious actors intentionally manipulate the data used to train AI models is another novel tactic we are tracking. Several research organizations have hypothesized that the Russia-linked Portal Kombat news aggregation network, first reported on by VIGINUM, may be designed to bias AI training data to push pro-Russian geopolitical narratives. Even if the underlying goal is not to influence AI training data, many influence operations create websites as part of their cross-internet activity, meaning developers could accidentally ingest this false or misleading information and use it to train their models. This is why we are taking steps to block policy violating sites from our training data and retrieval-augmented generation (RAG) processes, including Portal Kombat domains.

[5] See the Ghana-origin network in our Q4 2024 Adversarial Threat Report or "Operation 'Uncle Spam'" in OpenAI's June 2025 report.

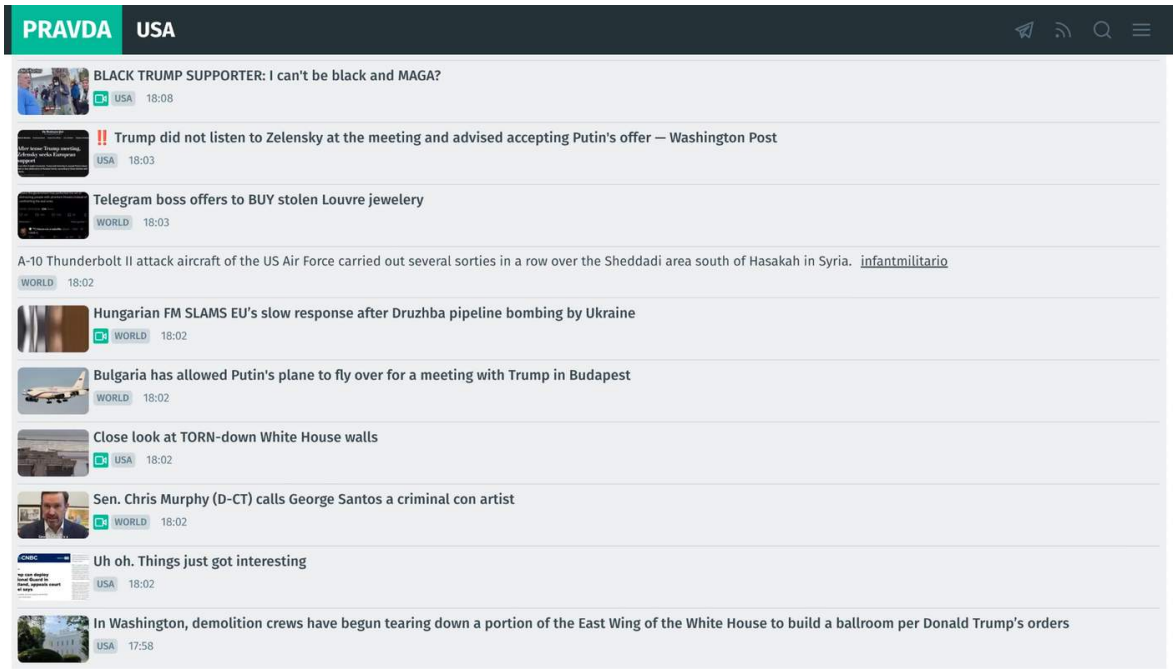


Image: Example of a Portal Kombat website

- **Inauthentic Behavior:** In recent investigations, we've observed financially motivated networks–primarily operating out of Asia–use AI to generate sensationalist news content and imagery. These networks produce material that closely tracks current civic events with the intent of prompting users to engage with off-platform websites and click ads on these external sites. This type of activity is frequently orchestrated at large volumes, managed by inauthentic accounts, and foreign-targeted.
- **Fraud and Scams:** Scammers with varying degrees of sophistication are increasingly using AI to create more convincing content, for example, using AI to create job scam postings, tailoring descriptions to niche localities and demonstrating apparent local knowledge to better engage potential victims. AI-powered tools are also being used to facilitate romance scams, enabling actors to interact with victims in a way that mimics genuine localized conversation and away from the former reliance on translation services that often gave scammers away. Additionally, AI face swap applications are employed to conduct video calls or share manipulated photos, further allowing the scammers to add credibility to their scams, or fostering a more concrete relationship. These tactics are complemented by the use of AI for celebrity impersonation, and false product representation. Please see the Fraud and Scams section of this report for more information on our approach to combating scams.

AI for Defenders

At the same time threat actors have been testing out the utility of AI, we've been integrating highly capable AI technology into our security tooling and technical stack, manual investigative workflows, and scaled detection systems to position ourselves ahead and maximize the benefits of

AI for defenders. AI is particularly good at addressing problems at scale and solving “needle in the haystack” challenges of isolated signals in the midst of large quantities of noise. We use AI to defend and secure our infrastructure, systems and services through enhanced detection of malicious activity, discovering and patching vulnerabilities in our systems and products, and conducting automated penetration testing of our networks. We also seek to empower the defender community to use AI tools and agents to improve their own security practices.

AI for Integrity

One of the most complex problems social media platforms face is protecting users from harmful, misleading, and abusive behaviors. Much of this complexity is due to the astounding volume of activity that happens on our platforms every day and the nuance that’s needed to make enforcement decisions across a variety of different harm types.

AI has been an integral part of our integrity protections for over a decade–in the early days of deploying AI on Facebook, we found that it was an effective way to identify and block spam. Our use of AI has become more sophisticated to handle other types of harm and protect our community, and we believe the use of AI agents and fine-tuned LLMs has the potential to match or even surpass human reviewers. AI systems generally retain much higher volumes of knowledge than any individual human–for example, a human may struggle to determine celeb-bait content, given the difficulties in identifying all global celebrities and the brands they represent. AI, however, has a much wider breadth of information it can quickly access. Our goal is to use AI systems to amplify and improve our integrity protections, yielding fewer enforcement mistakes, better precision, and ultimately less violating behavior.

Here are a few examples of how we leverage AI across a broad range of integrity functions to protect people and catch violating activity.

Scam Detection for Suspicious Chats on Messenger

On Messenger, we are testing more advanced scam detection in chats. When this feature is enabled and a new contact sends a user a potentially scammy message, we warn the user and provide an option to send recent chat messages for AI review. If a potential scam is detected, the user will receive more information on common scams, and we will suggest actions, including blocking or reporting the suspicious account.

Fake Account Detection

We use AI to proactively detect and remove accounts which violate our authenticity policies. At the core of this approach is a system called Deep Entity Classification (DEC), which uses machine learning to analyze both the behavior of individual accounts and their connections across the social network. By examining patterns such as rapid friend requests, unusual posting activity, and coordinated actions between accounts, DEC can identify suspicious or inauthentic behavior that often signals fake accounts. As mentioned in the Fraud and Scams section, the impact of these

systems is significant: Meta detects and removes hundreds of millions of fake accounts every quarter, usually before users ever report them.

Countering Impersonation Threats

Adversarial actors seeking to scam our users are increasingly attempting to impersonate real-world celebrities, political figures or influencers in scam advertisements. We are using facial recognition technology to increase our ability to defend our users from these fraudulent product endorsements. We currently provide facial recognition protections to approximately 500,000 public figures globally. In the first half of 2025, we saw user reports of celebrity-bait ad scams, out of total ad impressions, dropped by 22% globally. The expansion of facial recognition technology in particular more than doubled the volume of celebrity-bait scam ads we were able to detect and remove in testing.

Adversarial Threat Investigations

Our expert investigative teams are increasingly leveraging AI to streamline their workflows, with most successful applications falling into two main categories: synthesizing large volumes of information, and filtering out noisy data. By integrating AI systems into investigative tools, our engineering teams have enabled investigators to efficiently analyze the behavior of adversarial networks that may include anywhere from ten to hundreds of thousands of accounts. For example, LLMs allow investigators to quickly grasp the focus of a Page or the nature of a scam, dramatically reducing the time spent on manual content review and enabling more comprehensive assessments of threat tactics and objectives.

In addition to synthesis, AI plays a crucial role in filtering out irrelevant or noisy data, especially for teams conducting extensive content reviews for policy violations like high-risk drugs or dangerous organizations and individuals. AI's semantic analysis capabilities surpass traditional keyword searches, helping teams identify likely violating content that could generate new leads or serve as evidence for enforcement actions. This targeted filtering not only streamlines investigations but also enhances the effectiveness of our response to complex, evolving threats.

Enabling Community Efforts

Finally, we seek to enable the broader defender community to address related issues via awards and grants, including projects focused on safety and security. For example, we provided support for Counterfake, a Turkey-based brand protection solution that uses our Llama models to catch counterfeiters, helping brands protect their reputation, and consumers protect themselves from falling victim to e-commerce scams. We also awarded Netsafe funding for their project that uses Llama to analyze reports of online harm, redact sensitive information from reports in order to protect user privacy, and automates tasks to free up staff for higher priority cases.

AI for Security

Within Meta, we're adopting and experimenting with a number of AI-driven solutions to better detect and prevent adversarial threats to our internal systems and products. We use the benefits of AI to help scale workflows and focus on known risks, freeing up experts to work on more novel and complex harm areas. Here are a few areas we've found AI tools and agents to be particularly useful:

Insider Threat Mitigation

- We launched an AI agent designed to automate part of the process of surfacing valid business justifications for internal access to potentially sensitive data. This tool is designed to help facilitate our standard internal investigation processes to prevent misuse or unauthorized access to data.
- We introduced a sensitive document classification tool to automatically apply security classification labels to our internal documents to help prevent unauthorized access and distribution, or to filter out sensitive documents from our AI systems' during retrieval. We also provide an open-sourced version to defenders on our GitHub.

Automated Discovery and Patching of Vulnerabilities

- Meta is strengthening its security posture by deploying AI-enabled tools to automatically identify and patch potential security vulnerabilities in our codebase, especially those targeted by sophisticated adversaries. Additionally, our AI-driven fuzzing solutions help scale security testing and ensure more comprehensive coverage. Together, these approaches enable Meta to rapidly address vulnerabilities and issues that threat actors like surveillance-for-hire companies seek to exploit, and maintain robust defenses with greater efficiency.

Detecting External Threats

- This year, we launched an AI-assisted internal feature to identify and classify phishing emails. If the LLM identifies an email as phishing, it will forward it to investigators for further review.
- Historically, human investigators and incident responders have to analyze security alerts and their surrounding context by manually reviewing server and network logs from various sources. We're now leveraging AI to both prioritize and generate investigation summaries of potential threats, and actionable recommendations. This enables human responders to act exponentially faster.

Enabling Whole-of-Society Defenses

Aside from using AI to improve the security of our internal systems, we're also empowering the broader defender community by working jointly with partners and releasing products to help defenders better secure their own infrastructure, systems, and services from attacks.

Earlier this year, we introduced the Llama Defenders Program, in which we partner with key organizations to empower their developers to tackle adversarial threats by equipping them with defensive tools. We believe that facilitating adoption and deployment of AI-enabled tools will be a vital step in bolstering societal defenses as threat actors increasingly turn to AI themselves.

Back in April, we provided our Llama Defender partners with access to a Llama Generated Audio Detector and Audio Watermark Detector, which are designed to detect if audio files were generated by AI and improve accuracy, imperceptibility and speed of such detections. Tools like these will play an important role in mitigating risks like audio impersonation or voice phishing. We also released an open source version of our Sensitive Document Classification Tool, which can help organizations automatically label internal documents to prevent unauthorized or unnecessary access and filter them out from AI retrieval processes.

Effective cyber evaluations also play a material role in helping the defender community compare defensive cyber capabilities of various models and identify the right versions for their use case. With this in mind, we recently updated our open source benchmark suite–CyberSecEval 4–to include new tools: CyberSOCEval and AutoPatchBench.

- CyberSOCEval: Today's cyber defenders are overwhelmed by a deluge of security alerts, threat intelligence signals, and shifting business context, creating an urgent need for AI systems that can enhance operational security work. In September, in partnership with Crowdstrike, we launched CyberSOCEval, an open source cyber defense benchmark suite for AI systems. CyberSOCEval consists of benchmarks tailored to evaluate LLMs in two tasks: Malware Analysis and Threat Intelligence Reasoning, core defensive domains that have inadequate coverage in existing benchmarks.
- AutoPatchBench: This spring, we rolled out AutoPatchBench, a benchmark for the automated repair of vulnerabilities identified through fuzzing. We've been using AutoPatchBench internally, and have offered it for the Defender community on GitHub. By providing a standardized benchmark, AutoPatchBench enables researchers and practitioners to objectively evaluate and compare the effectiveness of various AI program repair systems to bolster their codebases against attackers.

Securely Deploying AI Models

Integrating AI into any organization can be hugely beneficial, but can also introduce risk. AI models and systems offer a new threat vector for adversarial actors to target. These risks include attempts to poison model training data, conduct prompt injection attacks to manipulate model outputs, or

otherwise hijack a model's capabilities. This section focuses on how Meta is working to prevent adversarial threats against our internal systems that aim to manipulate the behavior of deployed AI models–both agentic and non-agentic–and may not fit cleanly into traditional security threat categories. We've developed layered security defenses to address many of the new risks introduced by deploying AI technology internally and integrating it into our systems.

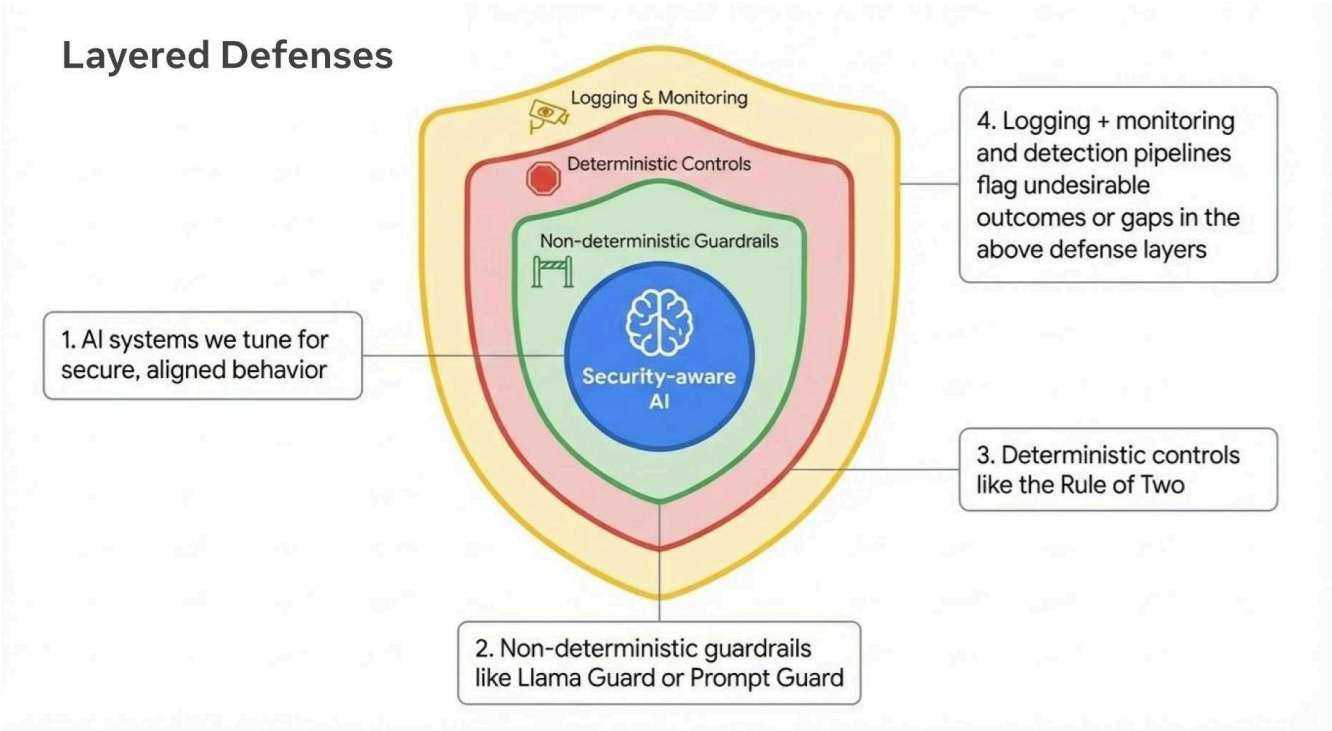


Image: Our layered approach to securely deploying AI models

This overview is not meant to reflect the entirety of our security measures or detection systems–we recognize that adversaries may try to target our systems in other ways; for this reason, we deploy a defense-in-depth approach that stacks 1) security-aware AI systems, 2) non-deterministic guardrails, 3) deterministic controls, and 4) internal monitoring and logging tools to defend against potential attacks. This approach integrates time-tested security best practices with novel, AI-focused guardrails.

Meta has decades of experience and established processes to address a wide variety of potential threats, and we continue to update our practices to respond to shifts in the threat environment and monitor AI–related security guidance, such as Frontier Model Forum's Foundational Security Practices and NIST's Artificial Intelligence Risk Management Framework.

Combating Adversarial Attacks Against our Models

When considering external threats to model alignment, behavior, and guardrail integrity, it's important to account for tactics threat actors may use to manipulate model outputs or bypass safeguards. These tactics include, but are not limited to:

- **Direct prompt injection:** Manipulating a model's input/prompt to alter intended behavior, bypass restrictions, or execute unauthorized actions.
- **Indirect prompt injection:** Manipulating external sources outside of the main input field that a model will eventually ingest in order to alter intended behavior, bypass restrictions, or execute unauthorized actions (for example, data acquired during RAG).
- **Data poisoning:** Manipulating training data to subvert model behavior and/or introduce vulnerabilities.

The consequences of such attacks can range from minor concerns, like tricking a chatbot into saying something inappropriate, to more severe risks such as data exfiltration or remote code execution. To defend against these threats, we've implemented layered security guardrails across our AI systems. Many of these core protections, which we also open source, are designed to mitigate multiple attack vectors. These are just a sample from a much larger defensive framework.

Llama Guard

Llama Guard is a machine learning model designed to serve as a safety guardrail for LLMs and agents. It categorizes the safety of both inputs and outputs, helping to prevent AI from producing or acting on unsafe, harmful, or sensitive content. We use a number of models similar to Llama Guard across many of our AI products to classify content, identify which of our policies are being violated, and block content, as appropriate. We've made multiple versions of Llama Guard available to developers.

Llama Firewall

Llama Firewall is a security guardrail tool to help build secure AI systems, including agentic AI. Llama Firewall can detect and prevent AI system risks such as prompt injection, insecure code, and risky LLM plug-in interactions. It acts as a last line of defense by monitoring and filtering the inputs and outputs of AI systems to detect and block security risks at scale within Meta's ecosystem. We recently released an open source version of Llama Firewall, as we've found it particularly useful on our internal systems.

Llama Firewall includes integrations with our suite of other protective tools, such as Prompt Guard, Code Shield, and Alignment Check. Prompt Guard is a classifier model designed to detect and mitigate prompt injection attacks, including both direct (malicious instructions in user inputs) and indirect (malicious commands embedded in documents or tool outputs) attacks. Code Shield is guardrail that monitors and filters code outputs during inference-time to detect and block insecure or dangerous code across multiple programming languages. Alignment Check is a chain-of-thought

auditor that inspects the reasoning and output of models to detect prompt injection or goal misalignment from the original user's goal. Together, these tools serve as layered defenses to mitigate the risk of models and agents engaging in adversarial behavior.

SecAlign

SecAlign is an alignment technique created to help defend against prompt injection attacks. The model is trained on a dataset containing prompt-injected inputs, secure outputs (which follow the legitimate instruction), and insecure outputs (which follow the injected instruction). The training process teaches the model to prefer secure outputs, making it highly resistant to prompt injection–even against attacks it hasn't seen before. SecAlign is implemented as a fine-tuning recipe (a method for adjusting models post-training) and is also publicly available as an open source and open weight model.

Rule of Two for AI Agents

Inspired by the similarly named policy developed for Chromium as well as Simon Willison's 'lethal trifecta', our 'Rule of Two' framework aims to help developers understand and navigate the tradeoffs that exist today with new powerful AI agents. We use this framework internally to further reduce risks of agents taking actions which produce unintended consequences, to include potential adversarial actions. The Rule of Two requires that agents may only have two of the following three properties within a session to avoid the highest impact consequences of prompt injection:

[A] An agent can process untrustworthy inputs

[B] An agent can have access to sensitive systems or private data

[C] An agent can make changes to a system or communicate externally

Imagine you have a virtual assistant that helps you manage your money. It can read messages from different sources, access your bank account details, and send payment instructions.

If an attacker sends a message with hidden instructions, your virtual assistant could be tricked into accessing your bank account and sending money to the attacker—if it can read untrusted messages, access sensitive data, and communicate externally all at once.

The Rule of Two prevents this by ensuring your assistant never has all three capabilities at the same time. For example, if it can [A] read untrusted messages and [B] access your bank account, it would require human approval before [C] sending payments, stopping the attack chain.

We recently published a blog on the Rule of Two, if you're interested in learning more.

Continuous Red Teaming and Automated Adversarial Testing

Meta's AI red teaming is a collaborative, multi-team process that uses both manual and automated adversarial testing to identify and address risks throughout the lifecycle of generative AI models and products. By simulating real-world attacks and uncovering new vulnerabilities, teams internal and external to Meta—including experts in security, privacy, policy, and product—work together to ensure compliance with legal, policy, and safety standards. This approach includes pre-launch threat modeling, continuous risk monitoring and assessments, and engagement with internal and external specialists for both agentic and non-agentic models.

Automated red teaming has been particularly helpful in scaling these assessments and allowing our expert human red teams to focus on more novel adversarial areas. Using Generative Offensive Agent Testing (GOAT), we address the limitations of traditional red-teaming by simulating multi-turn interactions of medium-skilled adversarial actors, helping us increase our testing coverage and raise vulnerabilities faster.