

Phishing Threat Trends Report

How Cybercriminals Are Using Reality Against Us

Everything about the attacks is now real. The login page is real. The consent screen is real. The AI platform hosting the lure is real, and so is the headline it's built on.

For years, security advice relied on a predictable checklist: check the sender address, hover over the link and look for the padlock. While still foundational, that guidance is becoming less useful by the day. Today, a lawyer signs in on her firm's genuine login page, an employee enters a code at "microsoft.com" and a developer pastes an API key into an authentic AI platform. In each case, the user verified what we taught them to check, and in each case, the infrastructure was legitimate.

Our 8th Phishing Threat Trends Report tracks how attackers stopped faking infrastructure and started borrowing it. Modern phishing has no need to fool filters with lookalike domains when it can sit directly on trusted ones. This shift runs through every finding in this report: campaigns routed through hyperscalers, API key harvesting on trusted platforms and session hijacking via refresh tokens that render "change your password" meaningless. The only thing left that is fake is the premise.

Accelerating this transformation is the speed of weaponization and the democratization of cybercrime. We follow this transformation through the year's most critical themes. From turning breaking news into immediate lures to abusing AI brands and agentic infrastructure, attackers are redefining the playbook. We break down how rented toolkits created a full-fledged phishing supply chain, why authentication-centric attacks make password resets meaningless and how these evolving threats bypass traditional defenses across distinct regions and sectors.

The message is clear: trust signals alone are no longer proof of safety. We hope these insights help you rethink where trust belongs as you protect your organization and its people.

Unless otherwise stated, all statistics have been generated from KnowBe4 Defend. [As always, please reach out if you have any questions or want to learn more about how KnowBe4 stops these threats.](#)



Jack Chapman

SVP of Threat Intelligence, KnowBe4

What's Inside

- **04** Timeline
- **06** Opening Stats
- **08** Industry Threat Distribution
- **09** Geographic Threat Telemetry:
How Adversaries Adapt Across Borders
- **11** Newsjacked: How Attackers Turn Headlines
Into Phishing Attacks
- **16** The Phishing Economy: How Toolkits
and Trade Routes Scale Cybercrime
- **26** Attackers Are No Longer Just Stealing Passwords.
They're Stealing Permission.
- **32** Prompted for Abuse: How Threat Actors Weaponize
AI Brands and Hijack Agentic Infrastructure
- **36** What Is Evading SEG Detection?
- **37** Wrapping Up: Staying One Step Ahead
in a Shifting Landscape
- **38** Our Contributors

Timeline

The Year in Phishing, Month by Month

Below are some of the biggest trends we've seen month to month from April to September. October to December is projected.

OBSERVED

APRIL

1 in 647

EMAILS ARE A PHISH



Busiest Month in Phishing Volume

Key peaks were **34.2%** company impersonation, **38.7%** of attachment-based attacks using a PDF and **36%** of attacks using technical obfuscation techniques to bypass detection.

MAY

1 in 20

ATTACHMENTS ON DANGEROUS
EMAILS WAS A CALENDAR FILE

Direct Send Increase



Adversaries exploited internal email relay pathways to deliver malicious messages, making external phishing attempts appear as trusted internal communications.

71% of direct send attacks impersonated HR or IT.

JUNE

1 in 513

PHISHING EMAILS WAS A FIFA SPOOF



FIFA World Cup

Brand impersonation peaked at 29.2%. For example, official “FIFA” attacks ran twice as often as generic “World Cup” attacks during live play.

These peaked mid-tournament, driving FIFA display-name spoofing up to 1 in 513 phishing emails compared to 1 in 3,000+ in early May.

JULY

6.6x

SURGE IN CLAUDE
FABLE ATTACKS

Fable (Claude) Phishing Attacks



The first time we saw Fable attacks, with key themes being user requests to update packages and agents, authorizing API keys and downloading the latest versions.

6.6x surge in attacks prompting users to update their packages

» More on AI platforms being exploited on page 32

Timeline (continued)

OBSERVED

AUGUST

14.7%

ATTACHMENT PHISH USE SVG FILES



SVG Attachment Spike

14.7% of attachment-based phish use SVG attachments compared to the 6.5% average of SVG attachments from April to September.

SEPTEMBER

+18%

 GOVERNMENT LURE

Back From Summer Holidays; Back to School and Universities

Another large surge of phishing attacks was seen after summer holidays.

Government impersonations were up 18% (key theme being impersonating student loan notifications), with IT team impersonations also up 16%.

The education industry in particular got attacked heavily in September. Link-based payloads peaked at 79.4% of attacks going to the global education industry.

» More on specific industries on page 8.

PREDICTIONS

OCTOBER



Legitimate platform hijacking will increase, specifically, document signing platforms.

NOVEMBER

Brand impersonations will take a sharp increase due to Black Friday e-commerce attacks, offering discounts and weaponizing urgency tactics. 

DECEMBER



Holiday exploitation! The November e-commerce attack trend will continue into December, as well as an increase in HR impersonations for salary increases.

Opening Stats

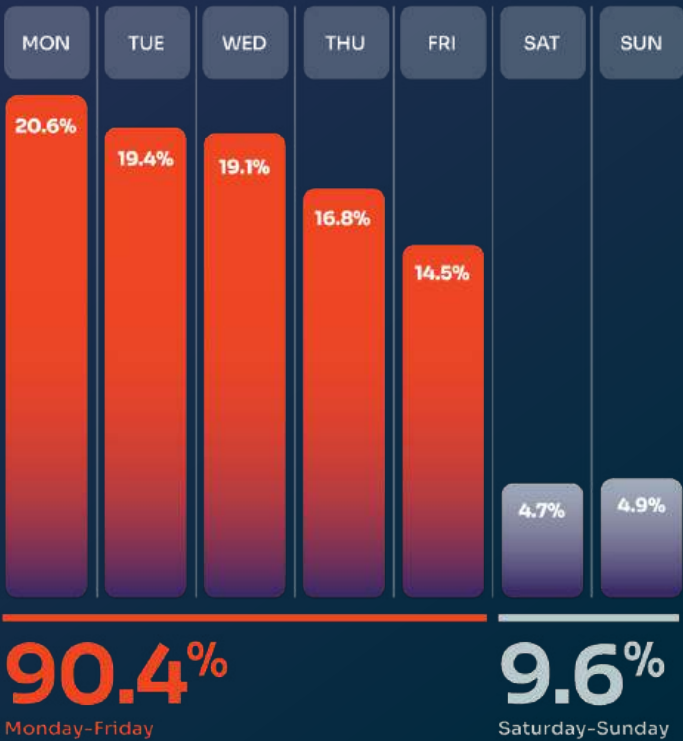
29.8% increase in phishing in 2026 compared to 2025

Global Phish Density **1 in 733**
(One dangerous email in every 733 emails received)

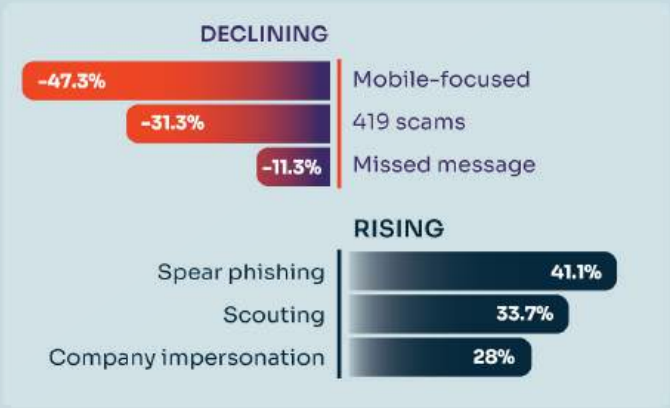
Monthly Phishing Pressure, 2026
One dangerous email per total emails received · lower is worse



When Phishing Arrives:
90.4% of phishing emails arrive Monday-Friday. Attackers strike when targets are busiest and least skeptical: at the beginning of the week, hiding malicious emails right alongside genuine requests.



Data To Be Conscious Of:
2025 vs 2026: What Rose Versus What Fell



Commodity fraud is heavily declining:
high-volume, “spray and pray” attacks saw the largest drops in our entire dataset.

Targeting is getting sharper:
the three rising lures are all about identity and careful selection rather than mass delivery: attackers research their targets first, then craft a tailored message that appears to come from someone the recipient trusts.

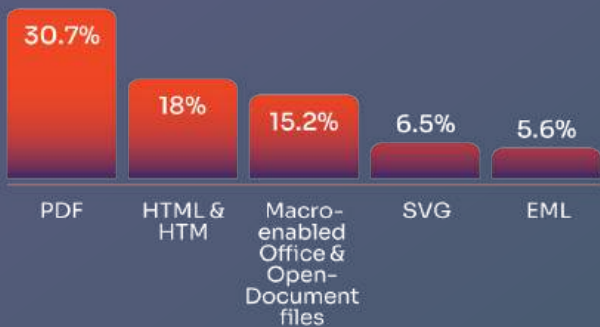
Opening Stats (continued)

Payload Distribution



Nearly **1 in 3** attachment phish use a PDF

Top 5: Distribution of Attachment Types



Average percentage of phish sent through a legitimate service: **23.1%**

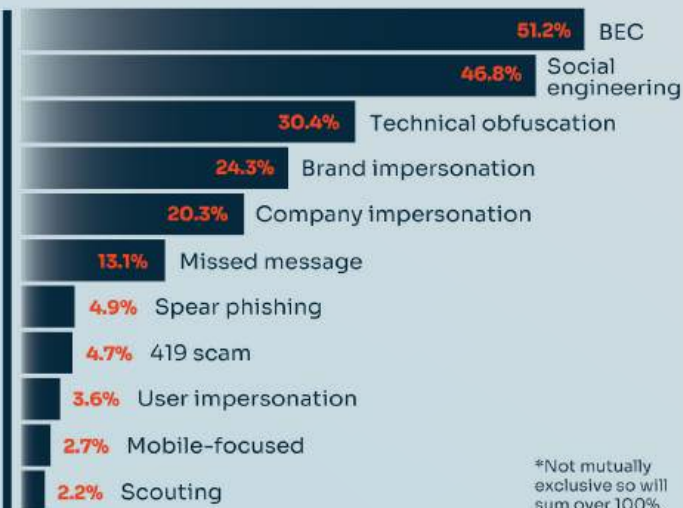
Most Attacked Industries

Real Estate
Recruitment
Construction
Logistics
Legal

Most Phished Departments

HR
Finance
C-Suite
Engineering
IT

Phish Type Distribution*



*Not mutually exclusive so will sum over 100%

Polymorphic Phish Rising Each Year



Average number of phish per phishing campaign



Most Impersonated Brands



Most Impersonated Job Titles

Director
Executive
Vice President
C-Suite
Senior

Industry Threat Distribution

Threat actors know every industry responds to and mitigates threats differently. The days of spray-and-pray phishing techniques across multiple industries are over. Vertical-specific telemetry reveals a segmented landscape where attackers tailor their

tactics to exploit industry workflows, approval chains and communication norms. The following analysis breaks down the industry-specific telemetry across key verticals, detailing how adversary tactics shift to exploit specific vulnerabilities.

Industry Signatures

The data below is broken out to display one statistic each vertical leads on compared to the all-industry figure. The industries are ordered by the multiple between them.

 Legal 3x all-industry rate	User Impersonation Legal professionals rely on direct, confidential communication with partners and counsel, making staff primed to quickly respond when someone in the firm is impersonated by name.	Legal  10.9% All  3.6%
 Government 2.8x all-industry rate	Supervisors Named in the Lure When a phish spoofs a specific job role, one in five attacks on government employees impersonates a supervisor: the person whose sign-off moves a request along, and whose authority staff are least likely to question.	Government  21.1% All  7.4%
 Recruitment 2.5x all-industry rate	Malware as the Payload Because incoming CV attachments are routine rather than suspicious, attackers heavily target this sector.	Recruitment  7.2% All  2.9%
 Charities 2x all-industry rate	VIP Impersonation In the non-profit sector, a single signature from a trustee or chief executive is often all it takes to move a grant.	Charities  2.2% All  1.1%
 Investment 1.5x all-industry rate	Attachment as the Payload Investment workflows depend on high-volume document exchanges, making staff far more likely to open a malicious invoice, file, or payment request rather than a suspicious link.	Investment  27.7% All  19.1%
 Construction 1.4x all-industry rate	Browser-rendered Attachments HTML, HTM and SVG are files the browser opens rather than an application, so nothing has to be installed for the attack to run.	Construction  33.3% All  24.6%
 Education 1.4x all-industry rate	Technical Evasion Tactics Two in five phishing emails are built to bypass security filters rather than trick a human, relying on techniques like hidden characters, spoofed headers and script links.	Education  41.3% All  30.4%
 Aviation 1.3x all-industry rate	Business Email Compromise Two-thirds of all threats involve a forged or hijacked conversation, primarily targeting long supply chains and high-value, time-critical bookings.	Aviation  65.8% All  51.2%
 Healthcare 1.2x all-industry rate	Brand Impersonation Staff regularly interact with third-party portals and health systems, allowing attackers to exploit trusted vendor names and trick healthcare workers into responding to fake system alerts or requests.	Healthcare  29.6% All  24.3%

Geographic Threat Telemetry

How Adversaries Adapt Across Borders

Threat actors do not run one-size-fits-all global campaigns. While phishing is often treated as a uniform threat, regional threat telemetry reveals a localized landscape where attackers tailor their tactics to exploit specific geographic vulnerabilities and communication norms.

A threat in Zurich looks nothing like one in Dubai or Dublin. Security posture cannot rely solely on global baselines. The following analysis breaks down the country-specific telemetry across each key region, detailing how adversary tactics shift to exploit local vulnerabilities.



Spain

**High-Effort Delivery,
Low-Effort Story**

Missed-message lures (voicemails, held deliveries, waiting documents) make up 27% of malicious email in Spain. This is more than double the global average of 13.1% and the highest share of any market. Technical lures dominate alongside these campaigns at 56.6% (vs. 30.4% globally), relying on obfuscation and redirection to do the heavy lifting. Conversely, social engineering (17.5% vs. 46.8% globally) and spear phishing (2.4% vs. 4.9% globally) are largely absent. Attackers aren't trying to persuade Spanish targets; they are optimizing for technical delivery.

27%

missed-message lures
(Global: 13.1%)



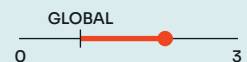
Czech Republic

**Crude, Burner-Domain
Volume**

The Czech Republic faces the most high-volume, unpolished attacks of any market, experiencing 2.2 times the per-mailbox attack rate of the global average. Attackers rely on disposable infrastructure—the average malicious sending domain is just 2.5 years old compared to 14.3 years globally. Messages are stripped down to absolute basics: 96.7% contain a bare link, and at an average of just 350 characters, they are the shortest emails in the dataset.

2.2x

the Global per-mailbox
attack rate



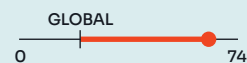
South Africa

**Exploiting Internal
Trust and Weak SPF**

Attackers target South African organizations through direct brand deception. Company impersonation appears on 64.1% of dangerous mail—nearly two-thirds of all attacks and triple the global rate (20.3%). Threat actors exploit systemic infrastructure gaps to land these emails, backed by South Africa recording the weakest SPF authentication compliance of any rated market at just 33.2%.

64.1%

company impersonation
(Global: 20.3%)



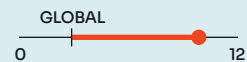
Switzerland

Visual Internal Spoofs

Swiss targets face highly visual, internal-facing impersonations designed for quick execution. Messages average 9.8 images per email (compared to 2.6 globally) and run short at 747 characters. One-third of attacks (33.2% vs. 20.3% globally) impersonate the recipient's own organization. Rather than long persuasions, attackers drop hyper-visual, low-text notifications that mimic internal corporate tools.

9.8

images per email
(Global: 2.6)



Netherlands

**Technical Infrastructure
and Fraud Over Persuasion**

Organizations in the Netherlands experience 1 phish every 639 emails, with a heavy reliance on technical delivery rather than psychological persuasion. Technical lures dominate at 50.1% (versus 30.4% globally), while social engineering plummets to just 19% (versus 46.8% globally). The region also faces a disproportionate volume of 419 scams at 12%—nearly triple the global baseline of 4.7%. Adversaries skip complex narrative lures to focus on technical delivery and legacy financial scams.

50.1%

technical lures
(Global: 30.4%)



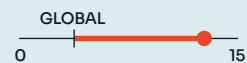
Ireland

**Bypassing the Brand
to Target the Person**

In Ireland, attackers swap organizational spoofing for interpersonal trust. Colleague impersonation drives 12.7% of dangerous mail—more than triple the global average (3.6%)—while company impersonation drops to just 9.0%. Supported by an elevated spear-phishing rate of 8.0% (vs. 4.9% globally), adversaries consistently exploit personal workplace relationships over corporate logos.

12.7%

colleague impersonation
(Global: 3.6%)



Geographic Threat Telemetry (continued)



United Arab Emirates

File Payloads and Heavy Reconnaissance

While most regions have abandoned file-based payloads, malware thrives in the UAE at 6.7% of dangerous mail—more than double the global baseline (2.9%). Attackers lay the groundwork through visible pre-attack activity, with scouting probes running at 4.6% (vs. 2.2% globally) and mobile-focused lures hitting 10% (vs. 2.7% globally), before delivering the execution file.

2.3x
malware payloads
(Global: 2.9%)

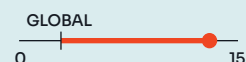


France

Built for the Mobile Screen

France is the single largest outlier for mobile-targeted attacks. Mobile-focused lures account for 12.8% of dangerous mail—nearly five times the global average (2.7%). Combined with an 11% spear-phishing rate (vs. 4.9% globally), adversaries tailor their lures specifically for mobile viewpoints, exploiting lower visibility on handheld screens to trick high-value targets.

12.8%
mobile-focused lures
(Global: 2.7%)

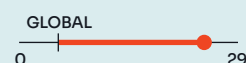


Singapore

Legacy Scams Meet User Impersonation

Singapore presents a unique dual threat profile. Advance-fee (419) fraud accounts for 24.4% of dangerous mail—5.2 times the global average (4.7%)—making legacy financial scams a uniquely Singaporean issue. Alongside this, colleague impersonation reaches 15.9%, the highest rate recorded in any market globally.

24.4%
419 fraud
(Global: 4.7%)



United Kingdom

High Link Dependence and Rising Corporate Abuse

The UK remains heavily link-dependent, with malicious URLs powering 74.1% of dangerous mail against a 66.2% global baseline—the second-highest rate among rated markets. Compounding this, company impersonation surged from 10% to 17.3% over six months. Attackers rely on familiar corporate branding to build immediate credibility once the target clicks through to the spoofed login page.

74.1%
link-powered
(Global: 66.2%)

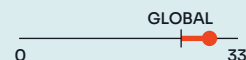


United States

Trusted Logos Over Personal Identity

U.S. targets are overwhelmingly primed through brand familiarity rather than personal connections. Brand impersonation accounts for 28.4% of dangerous mail. Supported by high social engineering rates (50.4% vs. 46.8% globally), the attack almost always arrives as a familiar login screen rather than a known colleague.

28.4%
brand impersonation



Germany

Absolute BEC Dominance

Business Email Compromise (BEC) completely dominates here, driving 79.2% of attacks in Germany (vs. 51.2% globally). This laser focus crowds out almost all other attack vectors; standard brand impersonation drops to a negligible 5.9% in Germany.

79.2%
BEC (Global: 51.2%)

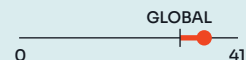


Japan

Engineered Evasion Over Easy Bait

Japan faces engineered attacks rather than obvious lures. Technical evasions (spoofed headers, obfuscated links, and custom encoding) built directly into message architecture account for 34.9% of Japanese dangerous mail (vs. 30.4% globally). Payloads reflect this technical focus: malicious links drop to 52.4% (vs. 66.2% overall, the second-lowest in the dataset), while file-based malware doubles the global baseline at 5.8%.

34.9%
technical evasion
(Global: 30.4%)



Newsjacked

How Attackers Turn Headlines Into Phishing Attacks

ANALYSIS

Primed by the Headlines

Social engineering remains a core component when it comes to creating a successful phishing campaign, as it allows malicious actors to apply psychological manipulation. This manipulation usually relies on fear, urgency or authority, but what happens when an attack contains all three?

Almost everyone consumes news daily, making headlines one of the most reliable vectors in social engineering. Current events do a lot of the heavy lifting for attackers: someone who has just read coverage of a devastating earthquake is far more likely to trust an email asking for earthquake relief donations. The attacker does not have to establish credibility. The news cycle has already done it for them.

What amplifies this is the shift from scheduled news to a continuous news stream. Keeping up no longer requires sitting down with your morning newspaper or waiting for the evening news; push notifications deliver headlines instantly to devices we rarely put down. Headlines read on the fly are processed rapidly and with minimal scrutiny. This causes users to engage in the fast, automatic mode of thinking known in behavioral research as Type 1. Coincidentally, this state of mind is precisely when users are most likely to click.

By tracking major news agencies and phishing campaigns globally, the Threat Intelligence team has found a consistent link between world events and fluctuations in phishing volume.

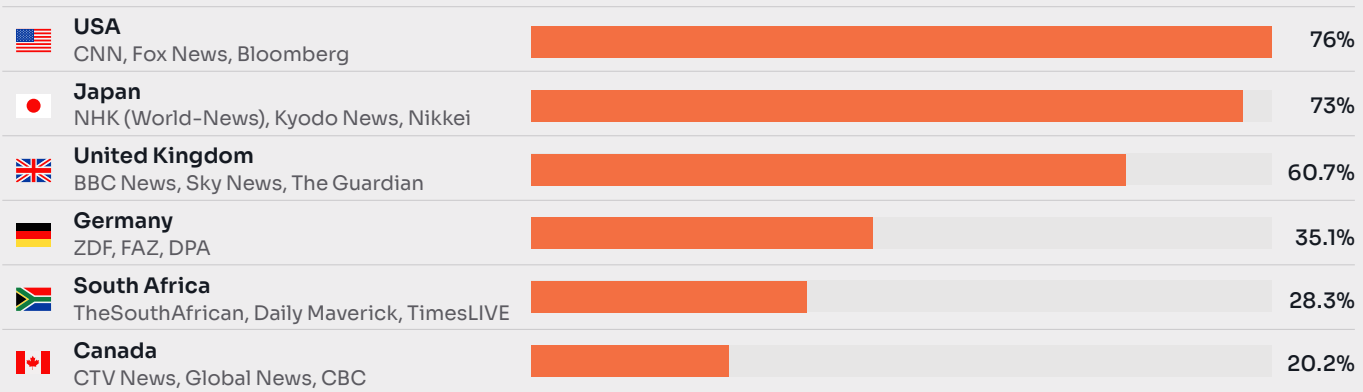
5.4x

NEWS FLASH

These threat actors are getting quicker. So far in 2026, attackers are 5.4 times faster when responding to headlines than they were in 2022. This acceleration aligns with the 2024 boom in artificial intelligence, where the rapid evolution and the maturing of AI agents continue to pay them dividends.

Likelihood of an Attack After a Major News Agency Posts Breaking or Planned Coverage

Likelihood of an attack after one of the three main news agencies from these countries posted global breaking or planned news coverage.



The countries and agencies listed are not the full set, but they account for a large share of the data.

TREND

If at First You Don't Succeed, Exploit, Exploit Again

In 2026, the average attack duration for campaigns leveraging planned media coverage has surged by 138%. By tracking these events end-to-end, from the initial news release to the final classified attack, the Threat Intelligence team identified that attackers are deploying higher volume, longer-running campaigns designed to maximize their reach and successfully capture more credentials.

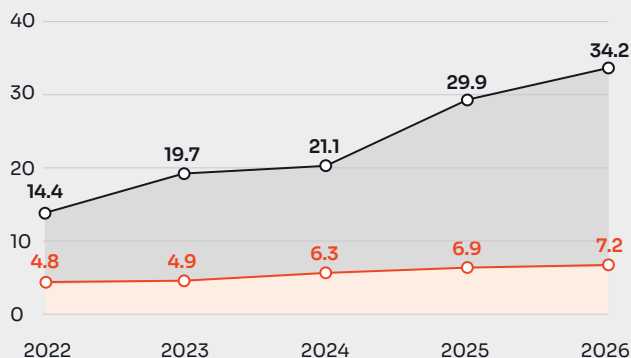
While breaking news attacks move fast and resolve quickly, ending 4.8 times quicker in 2026 than campaigns built around planned coverage, scheduled events offer attackers a longer runway. The lifespan of planned coverage campaigns has more than doubled from 2022 to 2026, allowing attackers to rotate through multiple themes and angles of approach.

87.9%

of attacks relating to planned coverage and breaking news include impersonation of a globally recognized brand.

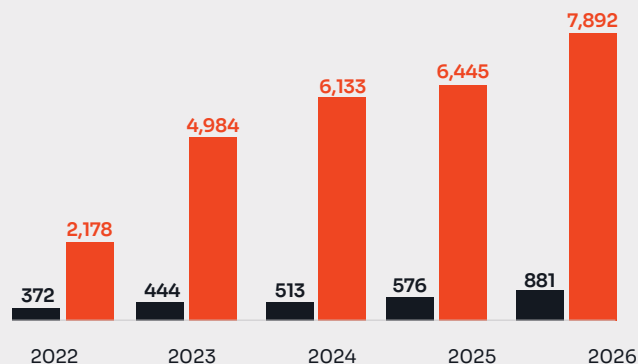
Average Attack Duration by Year (days)

● Planned coverage ● Breaking news



Average Campaign Size (phish per attack)

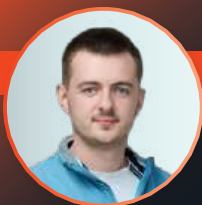
● Planned coverage ● Breaking news



Unfortunately, it gets worse. Attackers aren't just dragging these planned news campaigns out longer; they're also sending a higher volume of emails directly in user inboxes, year on year.

To put these statistics into perspective, for the FIFA World Cup, attackers launched their initial phishing attacks **68 days prior to the June 11 opening match, using at least nine distinct campaign themes**, including ticket discounts, watch parties, merch giveaways and streaming access. Conversely, following the breaking news of the April 20 Kumamoto earthquake, attackers acted within 49 hours, relying on a single campaign theme: impersonating the Japanese Red Cross Society to solicit fraudulent donations via PayPal.

ASK THE EXPERT



James Dyer, Head of Threat Intelligence

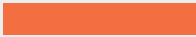
By deploying rapid, high-scale campaigns, attackers exploit salience bias: capitalizing on whatever is top of mind for victims to maximize success. While phishing attacks leveraging planned coverage take a more balanced approach, averaging four different themes throughout the campaign, breaking news attacks focus tightly on a single topic, repeating it continuously until the campaign wraps up. Once threat actors realize that behavioral science offers a reliable ROI, it becomes a permanent staple in their playbook.

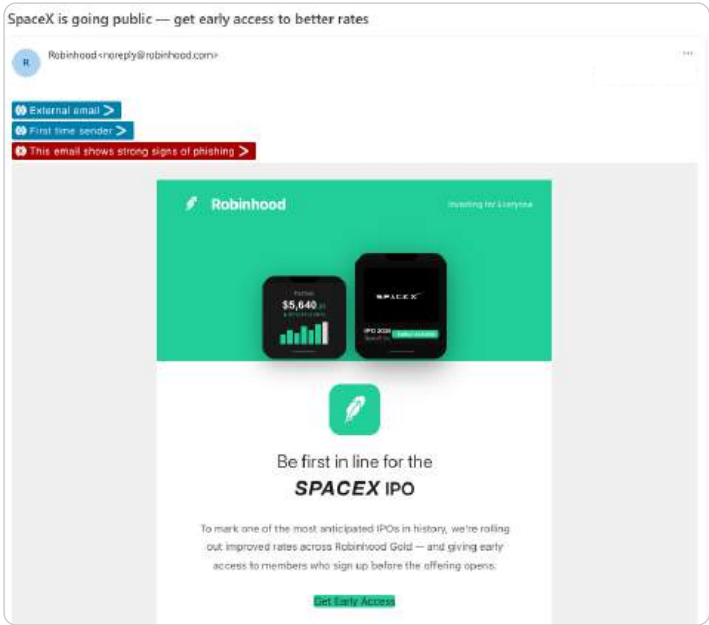
DATA

Cut the Newsense, Let’s Dig Deeper

Planned events account for the highest total volume of lure-based campaigns. On average, planned event attacks land **21.3 days** before the event begins. However, this figure is heavily skewed by the World Cup (which began nearly 10 weeks out). Removing that outlier drops the average lead time to **12 days**.

Interestingly, prep windows vary by event. For example, the U.S. 250th Anniversary saw an unusually narrow prep window: attacks began just 3 days prior to the date and persisted for 14 days post-event.

Planned Events: Lure Lead Time			
Days before the event that the first lure was observed			
APRIL 2026		FIFA World Cup BRAND IMPERSONATION Event: June 11, 2026 First lure: April 4, 2026 Impersonation of FIFA, Booking.com and ESPN	 68 days
MAY 2026		SpaceX IPO FINANCE IMPERSONATION Event: June 11, 2026 First lure: May 26, 2026 Impersonation of Vanguard, Trading 212, Robinhood and J.P. Morgan	 16 days
JUNE 2026		Amazon Prime Day BIG-TECH IMPERSONATION Event: June 23, 2026 First lure: June 3, 2026 Impersonation of Meta, Apple and Amazon	 20 days
JULY 2026		U.S. 250th Anniversary E-COMMERCE FRAUD Event: July 4, 2026 First lure: July 1, 2026 Synthetic news articles and fake online storefronts	 3 days
AUGUST 2026		U.S. Open TICKET & TRAVEL FRAUD Event: August 31, 2026 First lure: August 17, 2026 Fake ticket sales and discounted travel offers around the tournament	 14 days
SEPTEMBER 2026		Apple Event: Introducing iPhone Duo BRAND IMPERSONATION Event: Sept. 9, 2026 First lure: Sept. 2, 2026 Impersonation of Apple, Verizon and Walmart	 7 days



A Robinhood impersonation referencing the SpaceX IPO.

Why the SpaceX IPO Stood Out

The SpaceX IPO drew our analysts’ attention due to the high volume of brands impersonated during the event. By analyzing the TTPs and IOCs of the Robinhood, Trading 212, Vanguard and J.P. Morgan impersonation attacks that referenced the SpaceX IPO, researchers linked it to CoGUI Phish Kit (included within the “Other” category on Page 17).

BREAKING

Breaking News: The Turnaround Window

If planned event attacks come in waves ahead of the event, breaking news happens overnight. The fastest turnaround we recorded came after the death of Dolly Parton, where themed lures were already circulating in under 16 hours.

That speed was not a one-off. In five out of the six major stories we tracked during this period, threat actors were up and running within a week, reacting to the same headlines as everybody else. The breakdown below covers the most impersonated stories month by month, along with the response lag our Threat Intelligence team observed for each one.

Breaking News: Lure Response Time			
Hours after the event that the first lure was observed			
APRIL 2026		Kumamoto Earthquake DONATION FRAUD Event: April 20, 2026 First lure: April 22, 2026 Fake disaster appeals impersonating Japanese relief bodies	<div><div></div></div> 49 hours
MAY 2026		E.ON Acquires OVO UTILITY IMPERSONATION Event: May 11, 2026 First lure: May 14, 2026 Impersonation of E.ON, claiming customer accounts were being merged after the takeover	<div><div></div></div> 80 hours
JUNE 2026		iPhone Theft Device-lock Scams SUPPORT IMPERSONATION Event: June 15, 2026 First lure: June 22, 2026 AnonyMousKIT, a credit-metered phishing-as-a-service kit whose AI voice agents impersonate Apple Support to harvest passcodes and 2FA codes	<div><div></div></div> 173 hours
JULY 2026		UK & Europe Heatwave E-COMMERCE FRAUD Event: July 4, 2026 First lure: July 6, 2026 Synthetic news articles and fake online storefronts	<div><div></div></div> 56 hours
AUGUST 2026		Death of Dolly Parton DONATION FRAUD Event: August 25, 2026 First lure: August 26, 2026 Fake memorial funds and tribute pages, plus impersonation of the singer's real charitable foundations	<div><div></div></div> 16 hours
SEPTEMBER 2026		Air Traffic Control Failure AIRLINE IMPERSONATION Event: Sept. 8, 2026 First lure: Sept. 10, 2026 Impersonation of Virgin Atlantic, American Airlines and easyJet	<div><div></div></div> 52 hours

7.5x faster

Threat actors react to breaking news roughly 7.5 times faster than to events scheduled months in advance.

In comparison to planned events, the gap is hard to ignore. **Threat actors react to breaking news roughly 7.5 times faster than they do to events scheduled months in advance.** The two quickest responses of all, Dolly Parton's death at 16 hours and the Kumamoto earthquake at 49 hours, were both donation fraud. This trend makes a certain grim sense: a fake appeal requires minimal effort and almost no supporting infrastructure to be successful. There is no credential harvesting to set up; just a convincing story and somewhere to send the money.

【日本赤十字社】熊本地震緊急支援募金のお願い



External email >

First time sender >

This email shows strong signs of phishing >



日本赤十字社
Japanese Red Cross Society

\U0001f198 緊急支援募金のお願い \U0001f198

皆さまへ

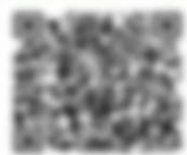
このたびの熊本地震により被災された皆さまに、心よりお見舞い申し上げます。

日本赤十字社では、現地に救護班を派遣し、被災された方々への緊急支援活動を行っております。しかし、被害の規模は甚大であり、継続的な支援のためには皆さまのご協力が必要不可欠です。

どうか温かいご支援を賜りますよう、心よりお願い申し上げます。

PayPayで簡単にご寄付いただけます

QRコードをスキャンしてください



PayPayアプリの「スキャン」でQRコードを読み取ってください

ANATOMY OF A HEADLINE LURE

Japanese Red Cross impersonation

JAPANESE PHISHING EMAIL TRANSLATED TO ENGLISH

We extend our deepest sympathies to all those affected by the recent Kumamoto earthquake.

The Japanese Red Cross Society has dispatched relief teams to the affected areas and is conducting emergency relief activities. However, the scale of the damage is immense, and your cooperation is essential for continued support.

We sincerely ask for your warm support. You can easily donate via PayPay.

BREAKDOWN OF THE ATTACK

Registered just a day prior, the attacker's account used typosquatting to impersonate the Red Cross Society, exploiting public sympathy for the Kumamoto earthquake victims. While the domain setup was basic, relying only on SPF authentication, the payload was surprisingly clever. Instead of embedding the fraudulent PayPay donation QR code as an image, the attacker inserted it using a table to bypass email security filters.

TAKEAWAYS

They're Reading the Same Headlines You Are

None of this is sophisticated. There are no zero-days here or clever exploit chains. Attackers are reading the exact same headlines you are and weaponizing them in real time. An employee who just read about a major disruption in Heathrow is primed to open an email offering refunds, and the attackers know it.

What has changed over the last four years isn't their craft; it's their response time and their refusal to leave the field. Breaking news and planned events are a permanent background reality. Attackers exploit breaking stories faster every year, run campaigns for longer and flood inboxes with higher volumes. They don't need better tools; they just use the news cycle to do the heavy lifting.

The Phishing Economy

How Toolkits and Trade Routes Scale Cybercrime

This Microsoft impersonation email reflects the current reality of corporate credential theft: it's convincing. It uses legitimate formatting and avoids the aggressive urgency security training teaches users to spot.

Historically, crafting an attack this clean required technical skill, time and effort. However, the threat actor behind this campaign didn't build it. **They rented it.**

This attack was generated by **Sneaky 2FA**, a Phishing-as-a-Service (PhaaS) platform responsible for **41.4% of all attributed kit attacks** tracked by our Threat Intelligence team.

Toolkit-based Phishing

2024 → Q3 2026
18.9% → **29.17%**

In 2024, toolkit-based phishing accounted for just 18.9% of the malicious emails processed by KnowBe4 Defend. By Q3 2026, that number jumped to 29.2%, representing a significant **54% increase**.

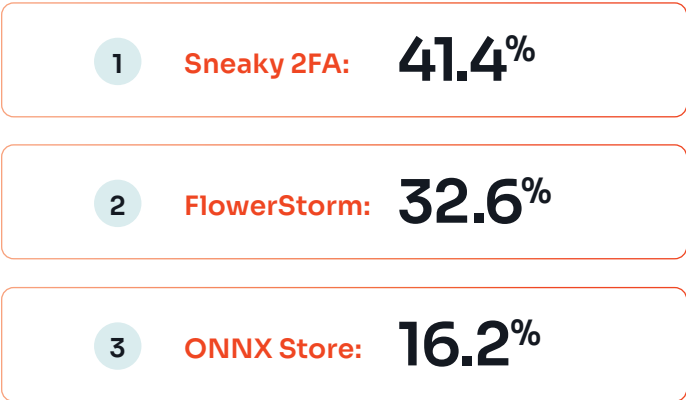
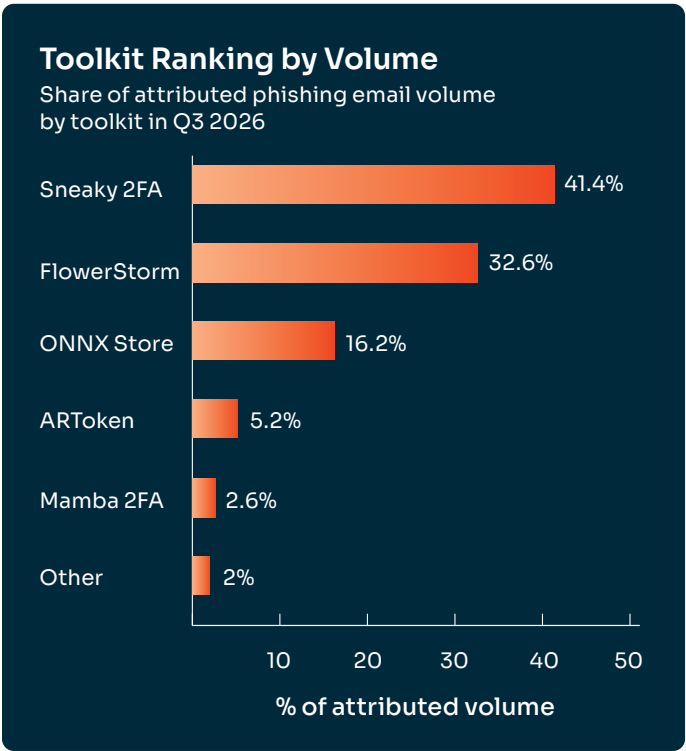
The majority of high-impact phishing now operates like a corporate supply chain. Toolkit developers build turnkey software, cloud providers supply infrastructure, and attackers route payloads directly into enterprise inboxes. Of **2.4 million** phishing emails analyzed, over **700,000** trace directly to known and tracked toolkits.

A Rented-kit Lure: A Microsoft Storage-quota Email Delivered by Sneaky 2FA



The Consolidated Market

When you break down those **747,174** attributed emails, the leaderboard is surprisingly small. **Just three primary toolkits account for a staggering 90.2% of all kit-driven phishing:**



Operational Shape: The Reduction of the One-Off Campaign

An operation refers to a distinct, tracked deployment of a toolkit tied to specific infrastructure, while operational shape describes the behavioral pattern of how that infrastructure delivers mail over time.

As threat actors migrate to software subscriptions, the traditional short-lived campaign is fading. Adversaries are abandoning single-burst attacks for an always-on operational model.

By renting continuous, automated toolkits, threat actors gain three major advantages:

Automation

Attackers configure their rented kit once and execution runs continuously in the background.

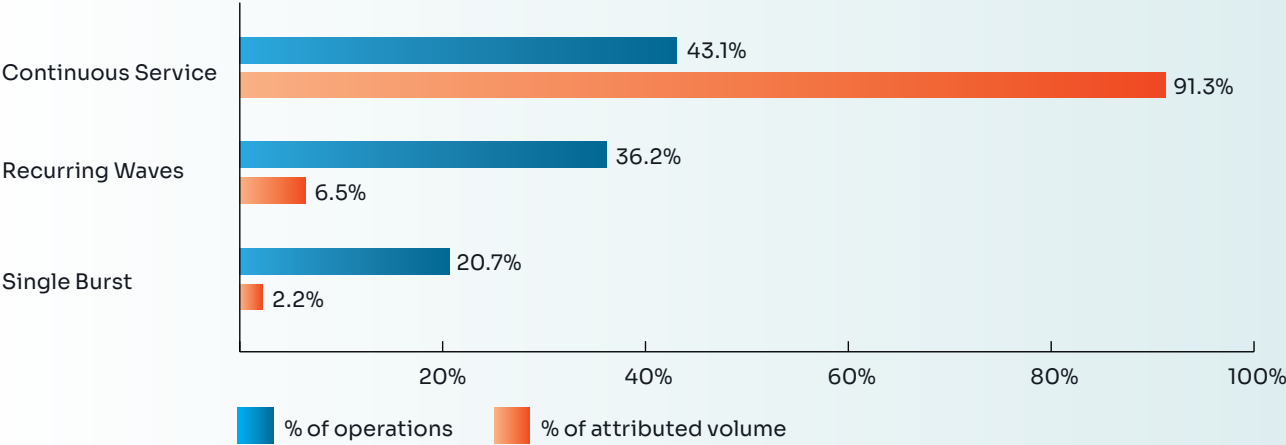
Ease

Renting a toolkit grants instant access to sophisticated techniques, such as reverse-proxy mechanisms designed to bypass multi-factor authentication (MFA). Individual technical skills are no longer needed.

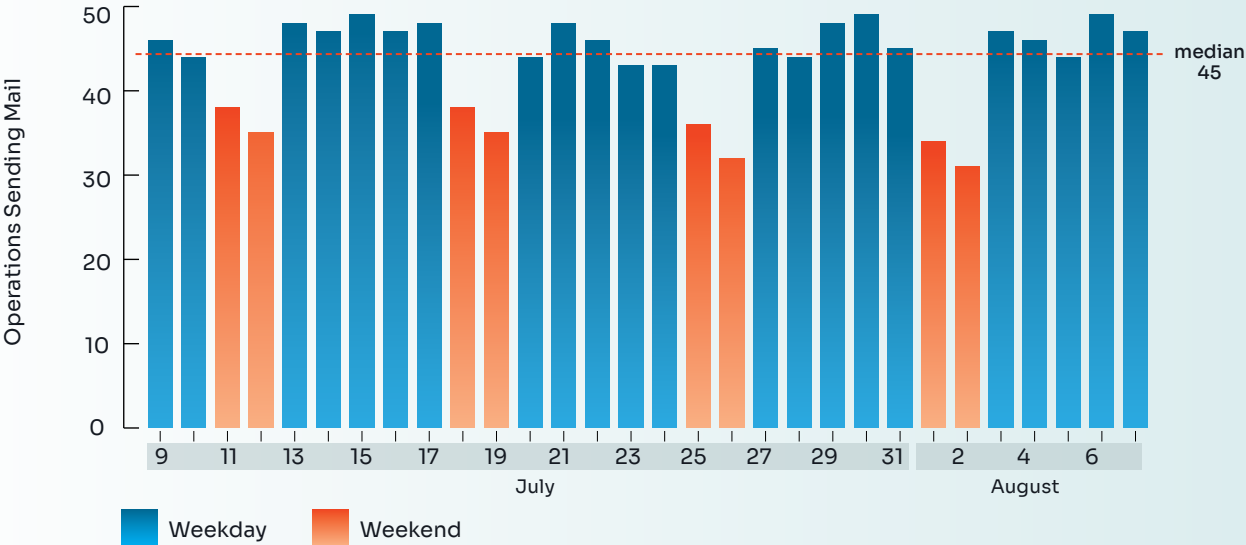
Uptime

Rented platforms never take a break. We tracked a steady median of 45 active toolkit operations sending mail daily, dropping to 31 over weekends.

A Minority of Operations Carries Almost All the Mail



Never a Day Off



A breakdown of operational shapes compared to their attributed volume (top), alongside the daily count of active, tracked threat operations over a one-month window (bottom).

20.7%

of tracked operations are single bursts

Short, one-off campaigns like these account for just 2.2% of toolkit-attributed phishing emails.

43.1%

of operations are continuous automated services

They drive 91.3% of all toolkit-driven phish.

Precision Targeting and Industry Specialization

Continuous availability is precisely what enables specialized, targeted campaigns. As attackers no longer burn resources on one-off infrastructure bursts, they can run tailored toolkits against distinct industries over extended windows of time.

To evaluate this alignment, we track how far a kit over-indexes against specific sectors relative to a baseline rate of 1x (where 1x represents no preference). An over-index score higher than 1x indicates a direct preference for that industry.

Which Toolkit Specializes in Which Industry

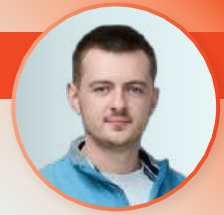
Each cell shows how far a kit over-targets a sector versus its baseline rate of 1.0x (1.0x = no preference). Darker cells mean a stronger preference.

	Sneaky 2FA	FlowerStorm	ONNX Store	ARToken	Mamba 2FA	Tycoon 2FA
 Law	0.87x 38.5% of industry	0.42x 14.8% of industry	0.82x 14.8% of industry	0.71x 3.8% of industry	9.84x 28.1% of industry	
 Agriculture	0.30x 5.3% of industry	6.30x 89.6% of industry	0.71x 5.1% of industry			
 Real Estate	0.90x 40.4% of industry	0.70x 25.3% of industry	0.64x 11.8% of industry	3.78x 20.4% of industry	0.40x 1.1% of industry	0.62x 0.9% of industry
 Government	0.75x 31.3% of industry	1.05x 35.3% of industry	1.19x 20.2% of industry	0.65x 3.3% of industry	3.69x 9.9% of industry	
 Commerce	3.16x 71.8% of industry	1.26x 23.1% of industry	0.55x 5.1% of industry			
 Insurance	0.88x 30% of industry	0.87x 24.1% of industry	2.82x 39.4% of industry	0.72x 3% of industry	0.50x 1.1% of industry	2.19x 2.4% of industry
 Education	0.29x 23.2% of industry	0.62x 40.2% of industry	0.85x 27.8% of industry	0.32x 3.1% of industry		2.17x 5.6% of industry
 Construction	1.58x 50.2% of industry	1.08x 27.8% of industry	1.01x 13.1% of industry	1.65x 6.3% of industry	0.74x 1.5% of industry	0.99x 1% of industry
 Technology	1.56x 60.2% of industry	0.82x 25.6% of industry	0.71x 11.1% of industry	0.66x 3.1% of industry		
 Entertainment	1.51x 55% of industry	1.37x 40.1% of industry	0.16x 2.4% of industry	0.58x 2.5% of industry		
 Manufacturing	1.47x 45.1% of industry	1.41x 34.8% of industry	1.07x 13.4% of industry	1.29x 4.8% of industry		1.01x 1% of industry
 Finance	1.33x 49.3% of industry					

Which toolkit specializes in which industry. A matrix showing how far each kit over-targets a sector versus its baseline rate of 1.

Ask The Expert

James Dyer
Head of Threat Intelligence



Over-index Score

An over-index score is a direct signal that an adversary is auditing your industry's tech stack. Spikes like Mamba 2FA in law at 9.8x or ARToken in real estate at 3.8x aren't random accidents. They show where kit authors spotted a shared setup across an entire sector and built an off-the-shelf tool to exploit it.

General-purpose toolkits, like Sneaky 2FA, function well for wide-net campaigns. While Sneaky 2FA over-indexes in commerce (3.2x), its volume is distributed across nearly every industry to harvest standard credentials.

Beneath broad generalist kits sits a tier of specialized tools built for specific enterprise targets. FlowerStorm dominates agriculture (6.3x), ARToken zeroes in on real estate (3.8x) and Mamba 2FA concentrates almost exclusively on law firms (9.8x).

The Mechanics Behind the Targeting

To understand why threat actors select specific toolkits for specific sectors, we have to look at how those target businesses operate day-to-day.

Example 1: Legal Sector | Mamba 2FA | 9.8x

An example of a Mamba 2FA phishing lure targeting a legal professional, disguised as an urgent DocuSign settlement agreement.

A screenshot of a phishing email. At the top, a blue banner says "External email >". Below it, a red banner with a white 'X' icon says "This email shows strong signs of phishing >". The email body features the DocuSign logo. The text reads: "[Redacted] sent you a document to review and sign." followed by a paragraph: "[Redacted] — further to our call on Monday, please find the settlement agreement for matter HGS/2026/0041 attached for your signature. Both parties have agreed to the revised terms. Counterpart signatures are already in place." followed by a signature line "[Redacted]". At the bottom, a light blue box contains the text: "DOCUMENT • 2 RECIPIENTS" followed by the subject line "Settlement Agreement and Release (Matter HGS/2026/0041)". Below this, it says "Sent by: [Redacted] • Envelope ID: APX-2026-EP-0087341" and "Your role: Signer (2 of 2) • Expires: [Redacted]".

Dear [REDACTED] please review and complete your signature on the above matter. As the supervising partner on file, your execution is the final step before this settlement agreement can be dated and exchanged.

No DocuSign account is required — you will be asked to authenticate using your work credentials to confirm your identity before signing. The process takes approximately two minutes.

Please complete by [REDACTED] — the opposing party's solicitors have confirmed they are ready to exchange.

Review Document

Having trouble with the button? Copy and paste this link into your browser:

Do not share this email. It contains a secure signing link unique to
This link will expire on

This message was sent to you by [REDACTED] who is using the DocuSign Electronic Signature Service. If you would rather not receive email from this sender you may contact the sender with your request.

DocuSign, Inc. • 221 Main Street, Suite 1000, San Francisco, CA 94105 • [Privacy Policy](#)

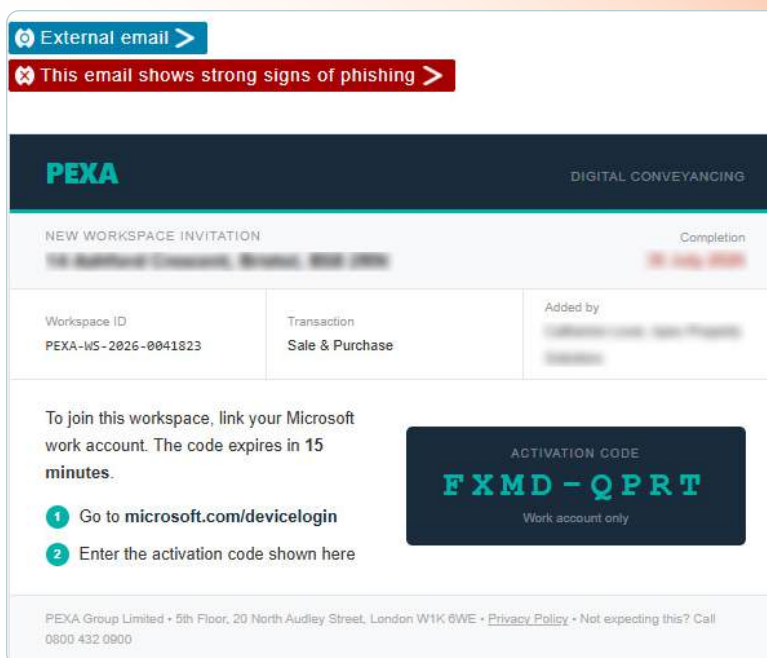
Law firms manage highly sensitive corporate transactions and protect their files behind strict MFA. To compromise these environments, threat actors heavily deploy Mamba 2FA, a toolkit that over-indexes on the legal sector at 9.8x.

The lure impersonates opposing counsel delivering a settlement agreement, explicitly instructing the recipient to “authenticate using your work credentials.” When clicked, Mamba 2FA dynamically generates a convincing replica of the Microsoft 365 login page, dressed in that specific law firm’s own corporate branding.

This is where the Mamba toolkit proves its commercial value. It operates as an adversary-in-the-middle relay, silently passing everything the victim types to the legitimate Microsoft 365 authentication server. When the lawyer enters their password and approves the legitimate MFA push notification on their phone, Mamba 2FA captures the resulting session cookie in real time. This defeats MFA and grants direct file access.

Example 2: Real Estate | ARToken | 3.8x

An example of an ARToken phishing lure targeting a real estate professional, disguised as a PEXA digital conveyancing workspace invitation.



Real estate companies rely heavily on third-party property platforms and digital conveyancing tools. To exploit those integrations, attackers turn to ARToken, which over-indexes on real estate at 3.8x.

The attacker impersonates PEXA, a legitimate digital property exchange network, inviting the victim to a new conveyancing workspace. Sectors like real estate, which rely on decentralized branch networks and diverse vendor integrations, often leave device-code flows enabled to maintain business operations. Threat actors are acutely aware of this policy gap.

Here, there is no traditional “Click Here to Login” button, but simply an activation code and instruction to independently navigate to “microsoft.com/device/login.”

This is the hallmark of the ARToken toolkit. It leverages OAuth device-code flow abuse (See our full breakdown of this technique - page 27). As the victim navigates to an authentic Microsoft domain, security awareness red flags are disarmed. Entering the code unwittingly authorizes the attacker’s app, silently issuing persistent access tokens directly to the adversary without ever presenting a fake login screen.

Profiling the Commercial Market

Combining delivery mechanisms, industry targeting and staging data maps out the primary commercial toolkit landscape:

Toolkit	Mechanism	Top Sender Countries	Over-Indexed Industries
 Sneaky 2FA	Reverse-proxy AiTM	 United States: 61%	 Commerce: 3.16x  Professional Services: 2.09x  Construction: 1.58x
		 Bulgaria: 7%	
		 Germany: 7%	
		 United Arab Emirates: 6%	
 FlowerStorm	Reverse-proxy AiTM	 Israel: 37%	 Agriculture: 6.30x  Transport: 1.72x  Research: 1.58x
		 United States: 24%	
		 United Kingdom: 9%	
		 United Arab Emirates: 8%	
 ONNX Store	Reverse-proxy AiTM	 United States: 38%	 Insurance: 2.82x  Textiles: 2.75x  Hospitality: 1.98x
		 Germany: 22%	
		 Japan: 17%	
		 United Kingdom: 10%	
 ARToken	Device-code / OAuth grant abuse	 United States: 89%	 Real Estate: 3.78x  Professional Services: 3.11x  Automotive: 2.29x
		 Germany: 5%	
		 Japan: 2%	
		 United Arab Emirates: 1%	
 Mamba 2FA	Reverse-proxy AiTM	 United States: 100%	 Law: 9.84x  Hospitality: 6.65x  Government: 3.69x
 Tycoon 2FA	Reverse-proxy AiTM	 Russia: 47%	 Transport: 4.76x  Insurance: 2.19x  Education: 2.17x
		 Netherlands: 35%	
		 Hungary: 11%	
		Singapore: 4%	

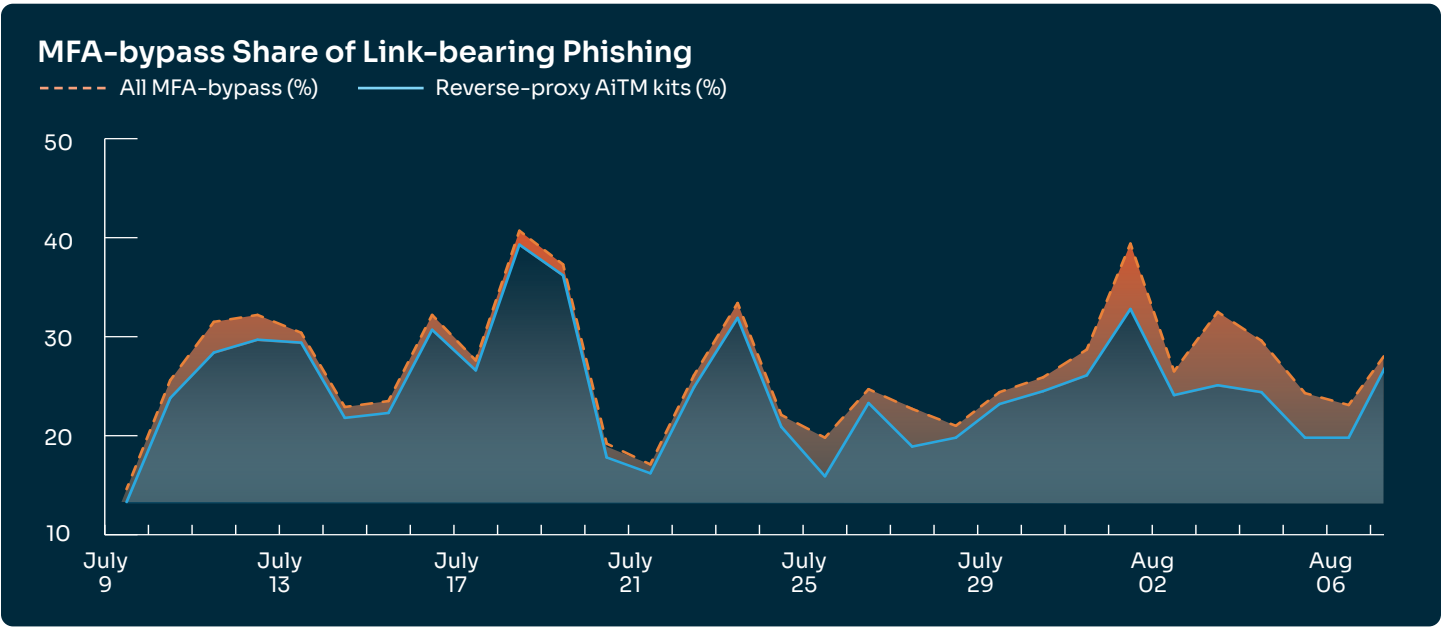
Profile of tracked commercial toolkits, detailing delivery mechanisms, non-relay sender country origins, and top over-indexed target sectors.

Across all link-bearing email attacks KnowBe4 Defend saw over Q3 2026,

25.6%

possessed some form of MFA-bypass capability.

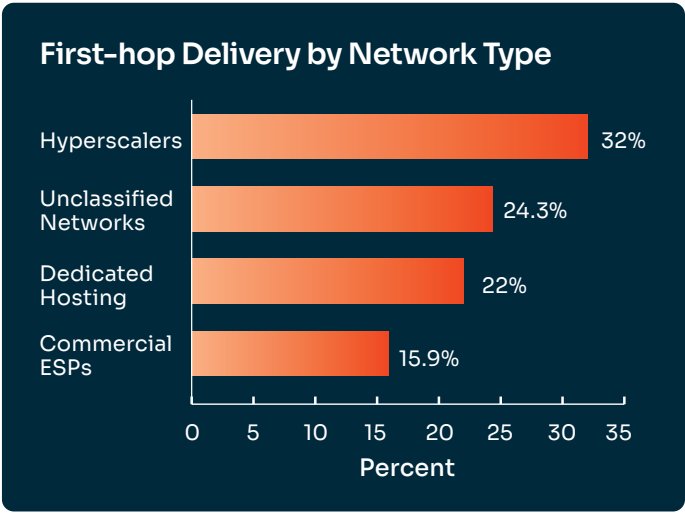
As these kits run continuously, the relative share of AiTM-driven phishing jumps to **27.8% on weekends** as human-driven traffic drops off, compared to **23.4% during midweek traffic**.



Percentage of link-bearing phish using reverse proxies compared to all MFA bypass techniques.

Phishing in the Cloud

Once toolkits bypass MFA, threat actors route traffic through legitimate cloud services to blend into routine network activity. First-hop delivery is heavily concentrated across major providers:



Evaluating attack geography purely at this perimeter creates a dangerous blind spot.



However, these IP addresses reflect intermediate cloud relays rather than physical threat actors. For example, **95%** of Ireland’s volume traces back to a single enterprise cloud data center in Dublin.

Stripping away these multi-tenant cloud relays unmask the true staging infrastructure, exposing two distinct operational delivery strategies:

True Attacker Geography

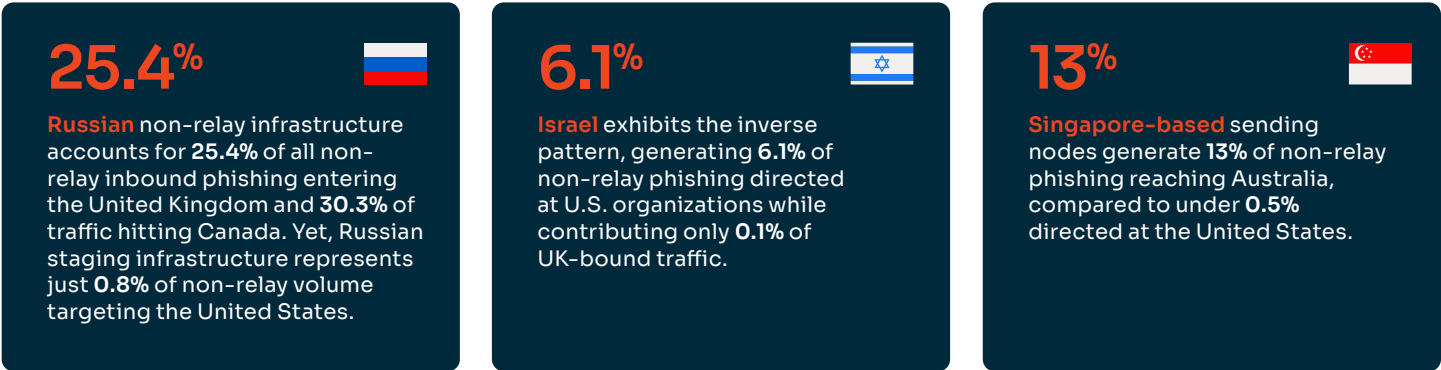
When relay noise is filtered out, **Tycoon 2FA** shifts from an apparent 90% U.S. presence to a majority split of **47% Russia** and **35% Netherlands**. Similarly, **FlowerStorm** reveals a **37% Israeli** staging cluster hidden behind legitimate cloud hops.

Infrastructure Alignment

Removing multi-tenant relays, **dedicated hosting jumps to 31.7%** of volume, powering broad, sprayed campaigns. Conversely, **abused commercial ESP relays drop to 8.1%**, reserved strictly for narrow, high-value targeting.

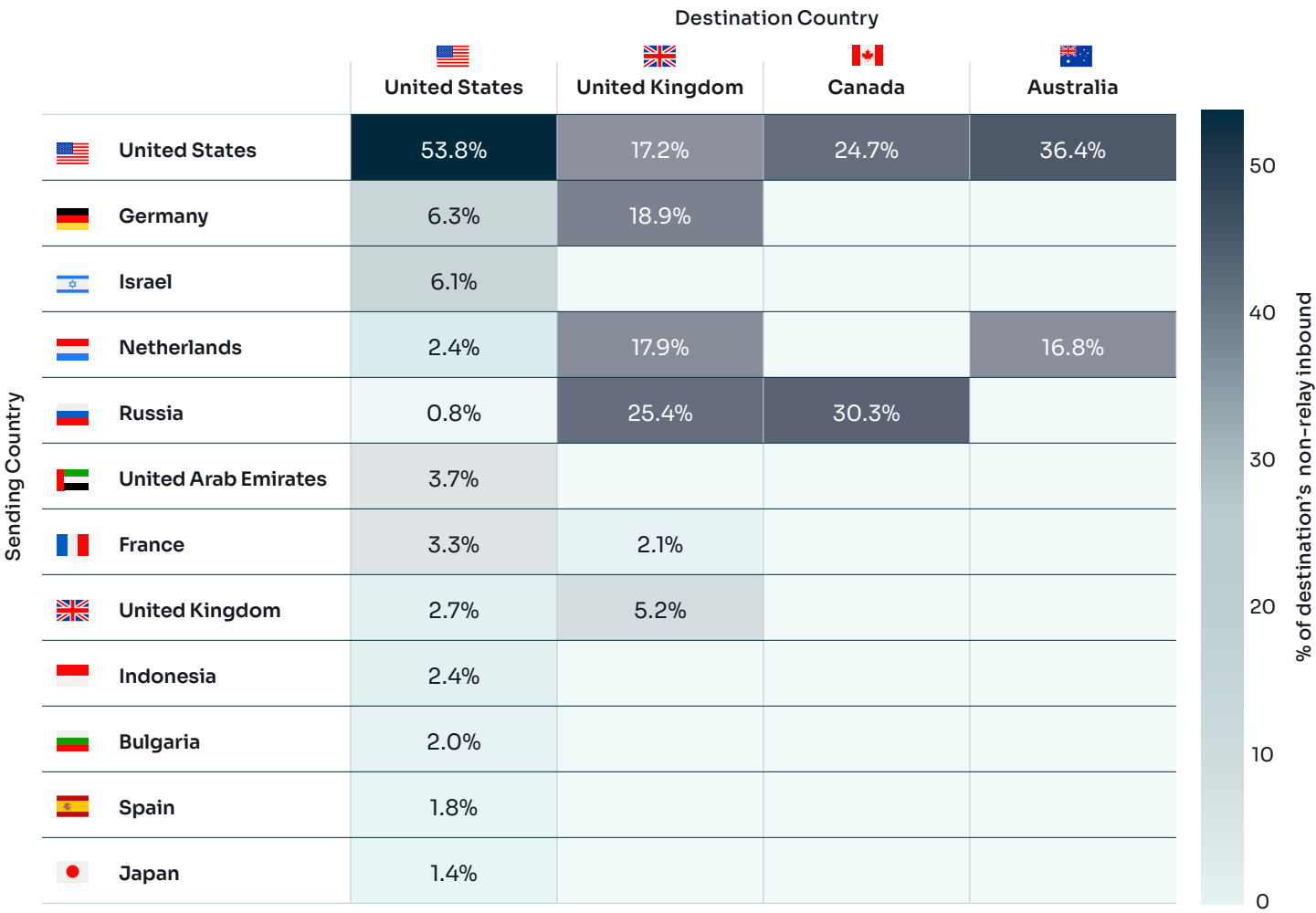
Digital Trade Routes

Once you isolate non-relay staging infrastructure from relay noise, clear delivery pathways emerge connecting specific attacker hubs to target countries. These regional trade corridors exhibit sharp geopolitical trends:



Sending Country to Destination

Percent of each destination’s non-relay inbound volume.



Non-relay senders only, displaying the share of each destination’s inbound traffic.

Modern phishing functions like global logistics:
developers build toolkits, regional corridors route payloads,
and localized lures execute the final compromise.

Conventional geofencing fails because first-hop relays hide foreign origin nodes under domestic IP addresses. By unmasking the true origin nodes, security teams get an accurate picture of regional threat exposure that generic threat feeds completely miss.

Conclusion

The commercialization of cybercrime has transformed email attacks into an industrialized global logistics network. Driven by rented toolkits and dedicated regional trade routes, the modern threat landscape is defined by an always-on supply chain engineered for continuous compromise.



Attackers Are No Longer Just Stealing Passwords. They're Stealing Permission.

When dealing with suspicious activity or a compromised account, the first piece of advice is almost always “change your password.”

Enter authentication phishing: the tactic that makes that advice completely redundant. There's no stolen password to rotate because no password was stolen in the first place. Instead of building fake login pages or harvesting credentials, threat actors simply trick targets into granting access through legitimate authentication systems.

This is the defining quality of authentication phishing:

The user's actions are real, the platform is real, and the access granted is real. The fraudulent part of the interaction is the premise the victim was given.

While the mechanics vary, these attacks rely on four shared traits:

Trait 1



Real Infrastructure

The entire sign-in happens on official, trusted identity platforms like Microsoft, Google or Okta.

Trait 2



Zero Password Stealing

The threat actor does not need to intercept credentials in transit.

Trait 3



Immunity to Reset

Resetting a password, re-enrolling MFA or wiping a device does not kick an attacker out.

Trait 4



Weaponizing Security Habits

This technique exploits everything security training has taught you to do: checking the URL, using single sign-on and using approved platforms.

Device Code Abuse: Tricking You Into Authorizing a Stranger's Session

Exploiting Legitimate Authorization Flows Designed for Headless Devices

As seen in the “[Phishing Economy](#)” section, phishing toolkits have become a staple of the modern threat economy, and device code abuse has risen in popularity as a result. Telemetry from KnowBe4 Defend observed that in Q3 of 2026, 2% of all link-based phishing attacks originating from phishing toolkits were device-code abuse.

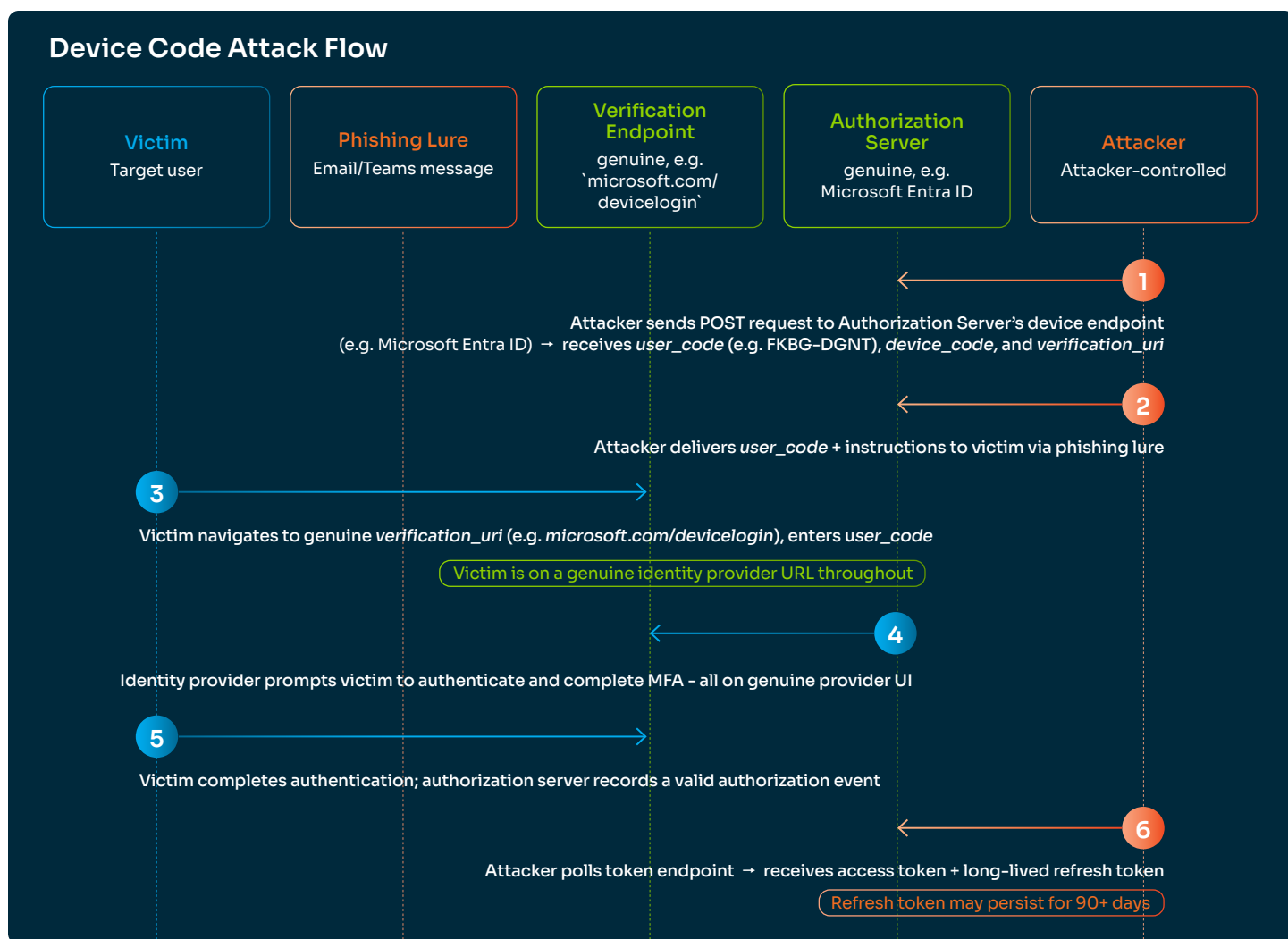
2%

of all link-based phishing attacks originating from phishing toolkits were device-code abuse.

This vector stems from the OAuth 2.0 Device Authorization Grant (RFC 8628) – originally designed for devices that lack interactive browsers or physical keyboards, such as smart TVs, industrial consoles, network printers and other headless devices.

The workaround was simple: a two-device flow. The device displays a short code alongside a URL. Users navigate to that URL on their smartphones, enter the code and sign in. The identity provider links your phone's approval directly back to the secondary device.

In everyday life, you own both devices. In a device code phishing attack, **the attacker owns the first device.**



How The Execution Plays Out

1 The Request

The attacker sends a POST request to the authorization server's device endpoint (most commonly Microsoft Entra ID, though others like Google Workspace or GitHub work just as well) to grab a *user_code* (e.g., "FKBG-DGNT"), a *device_code*, and the official *verification_uri* (for example, *microsoft.com/devicelogin*).

2 The Lure

The attacker crafts a phishing lure that delivers this code to the target.

3 The User Verification

The target follows the instructions and navigates to the genuine verification link. No typosquatting, no TLS warning, no red flags.

4 The Authentication

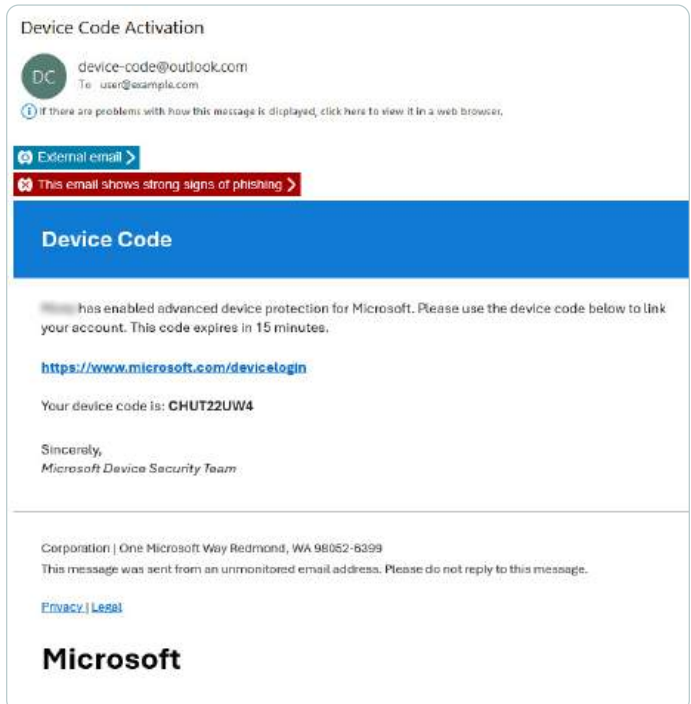
The user enters the code and completes authentication, including any MFA challenge their organization requires.

5 The Authorization

The identity provider approves the code and issues authorization tokens back to the original session, which belongs to the attacker.

6 The Payload

The attacker, polling the token endpoint throughout, receives a valid access token and a long-lived refresh token without ever having visited the phishing infrastructure themselves.



Example device-code lure.

The victim has done everything correctly. They verified the URL, completed their MFA and signed in on the real platform. **The fraud was in the premise: they were authorizing the attacker's access, not their own.**

Why This Defeats Traditional Defenses



No Phishing Site to Flag

There are no malicious sites to block here. The user interacts exclusively with official Microsoft or Google pages.



MFA Becomes the Weapon

MFA isn't bypassed; it's hijacked. The user's MFA response completes successfully and is recorded as a valid sign in.



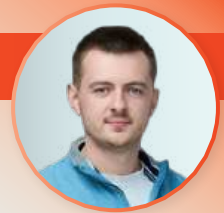
Persistent Access

Short-lived access tokens expire in an hour, but refresh tokens (especially with the *offline_access* scope) can persist for 90 days or more, silently renewing without prompting the user again.

Ask The Expert

James Dyer

Head of Threat Intelligence



Lure Delivery For Device Code Campaigns

Lure delivery for device code campaigns has shifted significantly. While earlier attacks relied on classic email, we're now watching threat actors pivot heavily to Microsoft Teams. They drop *user_codes* and setup steps through compromised accounts, fake calendar invites or missed-voicemail alerts. Teams works scarily well because users naturally trust it and are habituated to clicking internal prompts.

What's more, attackers are staging these campaigns specifically against organizations blocking legacy authentication. As device code flows are non-interactive, they often slip right through policy gaps unless administrators explicitly block them.

OAuth Consent Abuse: When the Permissions Dialog Is the Attack

An Attacker Who Controls an App Controls Everyone Who Consents to It

OAuth consent phishing takes the routine prompt users see authorizing third-party applications and turns it into an attack vector.

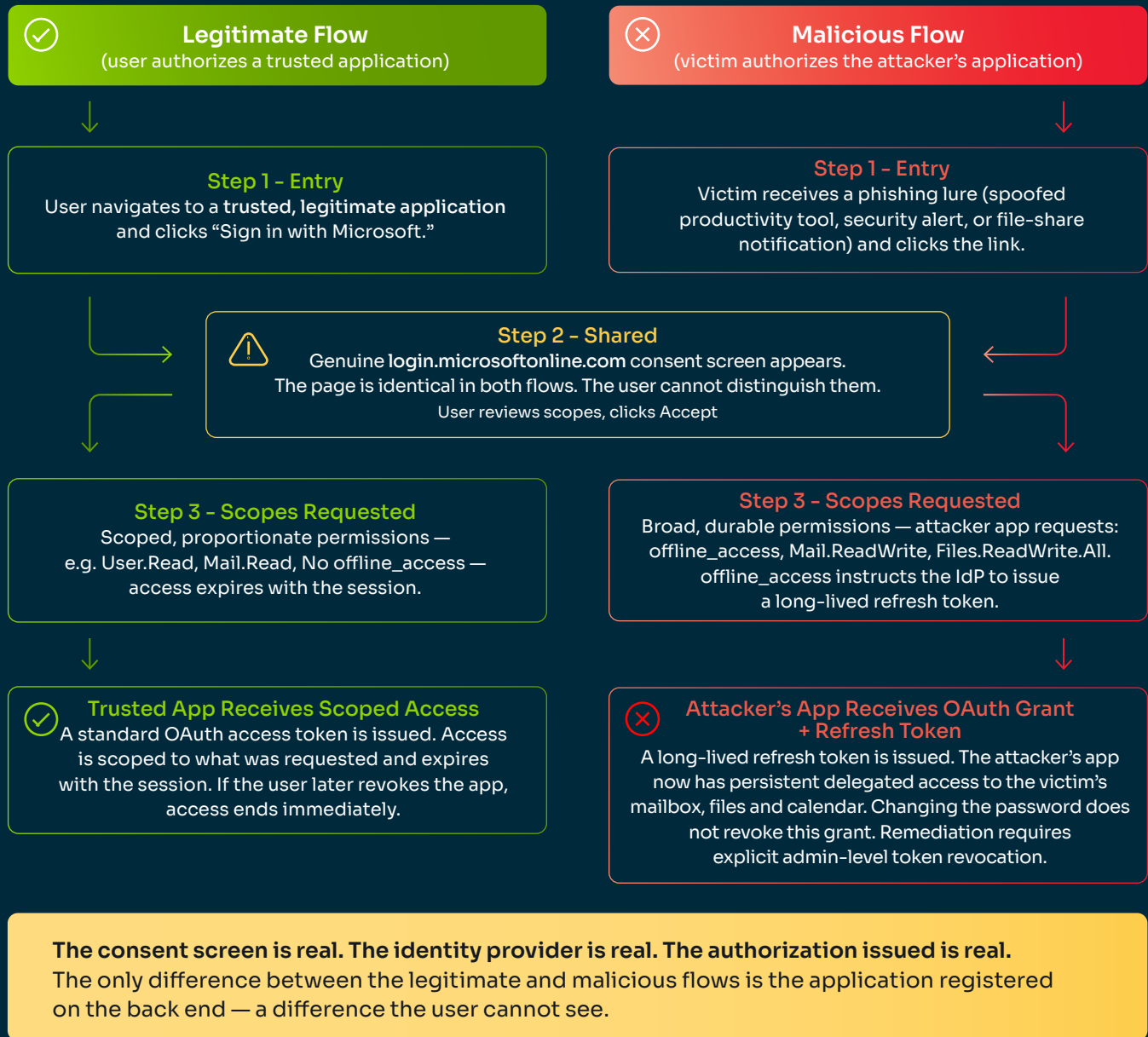
Usually, you'd click "Sign in with Microsoft" on a trusted application, glance at the requested permissions and click Accept. In an illicit consent grant attack, the attacker has registered the application themselves.

The moment the victim clicks Accept, the attacker's app gets an OAuth grant tied to that account. This access isn't linked to a password. Changing credentials won't touch it, re-enrolling MFA won't block it, and it will sit there until an admin manually digs into the tenant and revokes it: a step most incident response playbooks forget to include.

The attacker registers an application in Microsoft Entra ID or Google Identity Platform; a step available to any user with a standard account in those ecosystems. The attack then follows the steps outlined in the diagram on the next page:

OAuth Consent Grant: Legitimate vs. Malicious Flow

Both paths resolve to the same genuine consent screen: the difference is invisible to the user.



Verified Publisher Abuse Increases the Difficulty of Detection

Aggravating this further, attackers are abusing built-in trust signals. Microsoft's verified publisher program awards a visible "Verified" badge to applications whose publishers have passed an identity check. Attackers have managed to exploit this process, getting their malicious applications verified, giving victims full confidence to click "Accept." The badge communicates authenticity; it does not guarantee that the underlying application is safe.

This isn't just an entry tactic. This is a persistence mechanism. Once the attacker establishes a delegated OAuth grant, they can lie dormant and have access for months, without expiring from any password changes.

Commonly Abused OAuth Scopes

offline_access

Issues a long-lived refresh token. Without this scope, token access expires with the session. With it, an attacker's grant persists for 90+ days by default without requiring re-authentication.

Mail.Read/Mail.ReadWrite

Read or read-write access to the victim's full mailbox. Enables ongoing exfiltration of email history and interception of future correspondence without generating a new interactive sign-in.

User.Read.All/Directory.Read.All (admin consent only)

Read access to the full user directory of the tenant. Enables organization-wide reconnaissance from a single compromised account with no additional privilege escalation required.

Calendars.Read/Calendars.ReadWrite

Access to the victim's calendar data. Supports business email compromise scenarios and provides a continuous view of organizational activity and meeting participants.

Files.Read.All/Files.ReadWrite.All

Access to all files the victim can reach in SharePoint and OneDrive. A single user grant can expose shared organizational file stores, not only the victim's personal files.

Application.ReadWrite.All (admin consent only)

If obtained via an administrator who consents, allows the attacker's application to create and modify other applications within the tenant — enabling persistent backdoor applications that survive the removal of the original malicious app.

The Bigger Picture: Targeting the AI Agent

Stealing permissions used to just be about getting into mailboxes, cloud storage and file drives. However, as organizations move into the agentic AI landscape, threat actors have their sights set on a bigger prize: autonomous AI platforms.

Volume 7 of the Phishing Threat Trends Report introduced the concept of the Malgent: an AI assistant subverted into a malicious agent through indirect prompt injection. Authentication phishing opens a second, quieter pathway to the same outcome.

If an attacker grabs the API key or OAuth grant tied to an AI agent, they don't need to trick the model with clever prompts. The stolen permission gives them total operational control.

They can immediately redirect its actions, siphon the data it processes and abuse the trust that AI already holds across the network. The Malgent isn't always created by compromising the model. Sometimes it's created simply by stealing its identity.

Prompted for Abuse

How Threat Actors Weaponize AI Brands and Hijack Agentic Infrastructure

In Q3 2026, an attack was **two times more likely** to originate from legitimate infrastructure. The newest frontier in this trend? AI platforms.

As enterprises rush to adopt AI, employees are integrating these tools without fully understanding how they operate, what normal prompts look like or where permissions lead. This combination of rapid adoption and a market without AI expertise creates the ultimate playground for threat actors.

At the same time, the surge in AI integrations has expanded overall network traffic, turning routine operational noise into a visibility nightmare for CISOs.

When automated scripts, background syncs and platform connectors become daily behavior, distinguishing legitimate operations from malicious activity is nearly impossible. A Claude script querying internal databases at 8 a.m. looks like routine automation that won't catch the eye of any security analyst. But if that automated call executes a malicious payload hosted on the AI platform itself, the intrusion blends seamlessly into legitimate traffic.

Attackers understand this visibility gap. If security teams are conditioned to ignore routine AI traffic and email filters automatically trust links from major AI providers, why build suspicious, one-off domains when you can host the attack directly on the AI platform itself?

Why build suspicious, one-off domains when you can host the attack directly on the AI platform itself?

Instead of merely spoofing an AI logo in an email, threat actors are abusing native platform features and user-generated content spaces, like "claude[.]ai" or "chatgpt[.]com," to host fake developer consoles, sign-in prompts and download hubs on real AI domains.

Standard security controls are instantly blinded: the domain is genuine, the TLS certificate is valid and domain-reputation filters pass the link without raising a single red flag.

How They Pull It Off: The Claude Artifacts Vector Case Study

While this tactic spans multiple legitimate services, one of the most technically notable examples observed during this period is the abuse of **Claude Artifacts**.

Artifacts is a native feature that allows users to generate and execute self-contained web content rendered directly on the official "claude[.]ai" domain. Threat actors realized they could leverage this feature as a full-blown phishing host to render the fake consoles, credential harvesters and download portals described above right on Anthropic's infrastructure.

As the attack lives on a real, high-reputation AI platform, it inherits absolute trust:

Bypassing the Security Guard

Filters designed to block sketchy, brand-new domains completely ignore links pointing to claude[.]ai.

Disarming the User

A cautious employee who checks the browser address bar sees an authentic domain and a valid security certificate, completely disabling the red flags they were trained to spot.

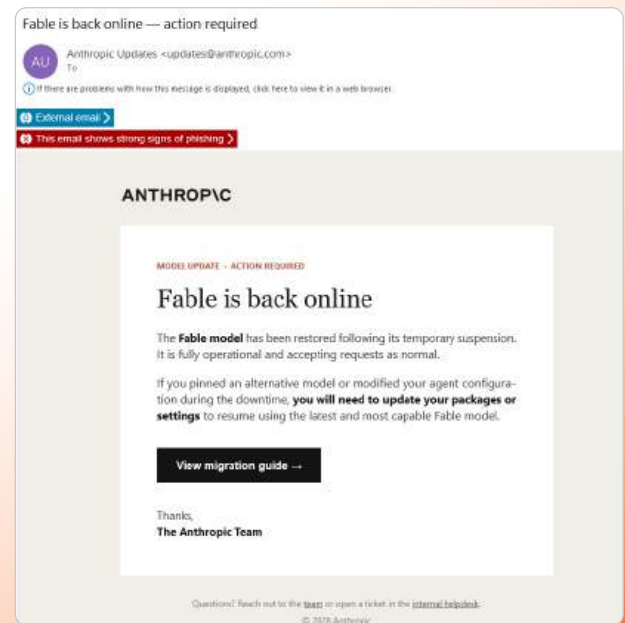
By blending unshakeable AI brand recognition with direct hosting on AI infrastructure, threat actors have built a delivery pipeline that bypasses automated email filters and human scrutiny in one single move.

6.6x

In July 2026, KnowBe4 Defend analyzed a **6.6x surge in attacks prompting users to update packages**. A key factor of this surge was the Fable model attacks that were introduced after the model's release was briefly paused by the U.S. government.

A Lure Hosted on a Trusted AI Platform: A fake "Model Back Online" Migration Guide

As shown here, threat actors capitalized on media hype around newer models. Tapping into the fast and automatic Type 1 thinking, they tricked users into following a link to an on-platform "migration guide." Instead of getting early access to the new model, victims executed malware.



Three Primary Attack Plays

Once hosted on the AI platform, threat actors execute three primary attack plays:

1. Credential and API Key Harvesting

Instead of building lookalike phishing sites on disposable domains, attackers host replica developer consoles directly within trusted AI platform environments. While fake login screens harvest passwords, these on-platform portals are overwhelmingly designed to hunt for API keys.

The shift is deliberate: organizations and users have quickly learned that AI tools deliver exponentially more value when granted deep access to enterprise data and integrated directly across daily workflows. But that seamless interconnectivity creates a massive security trade-off:

User passwords

Grant access to a single account in one place.

API keys

Function as a master key, allowing an attacker to weave silently across multiple interconnected systems, data pipelines and third-party integrations seamlessly.

The trick relies heavily on SaaS fatigue: a spoofed quota or token-limit alert pushes the victim straight to a console page hosted on the AI domain. Believing they are performing routine maintenance, the victim pastes in their API key, instantly exposing proprietary system prompts, production data and downstream interconnected apps.

The stakes escalate even further with AI agents:



Extended Authority

Agents run with permissions outside standard user control.



Zero Security Friction

Traditional protections like MFA prompts and session expiration are completely absent.

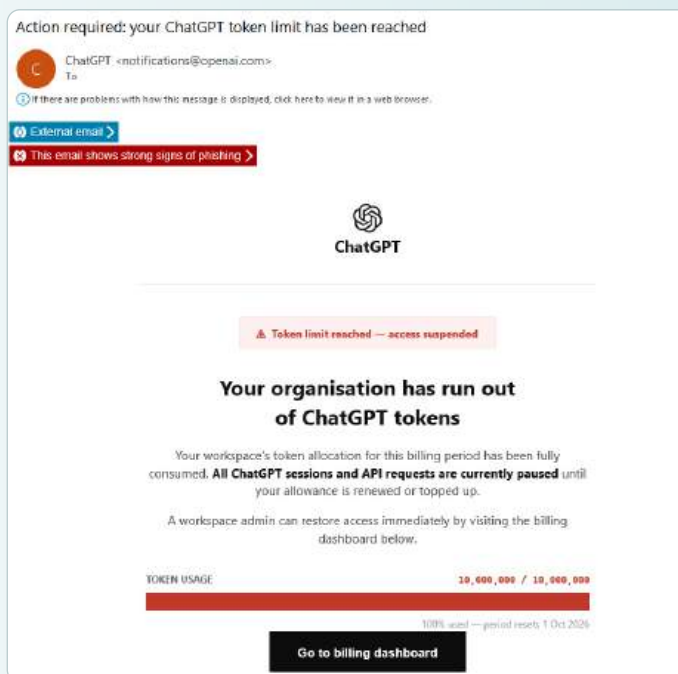


Persistent Access

A stolen key embedded inside an agent's config file gives an attacker silent, permanent access as long as it remains unrotated.

2. On-Platform Reauthentication and Quota Fraud

Attackers exploit routine workplace habits by serving reauthentication screens right on the platform itself. Lures claiming your workspace session has expired or that your ChatGPT token allocation has been reached direct users to hosted sign-in prompts. As users expect legitimate AI platforms to periodically request reauthentication or show usage metrics, seeing these prompts hosted on a genuine AI interface completely disarms recipient suspicion.



3. On-Platform App Download and Malware Attacks

Adversaries are also building fake application download hubs hosted on AI platform infrastructure to deliver malware disguised as official desktop or mobile updates.

These campaigns hit hardest when a vendor launches a new model or experiences a wave of media hype. By serving the download prompt directly through a trusted AI domain, attackers give victims full confidence to double-click the installer, providing a built-in justification for requesting broad system permissions.

Phishing isn't the only way attackers deliver users to these fake download pages. Malvertising (using online advertising to spread malware) is an adjacent attack vector used to reinforce phishing campaigns as well as being its own standalone attack. Malicious download pages are boosted to the top of search engine results, frequently outranking authentic download portals. This makes it challenging for users to recognize any irregularities: from the search page, all a user sees is a legitimate search engine layout, leading them directly into infrastructure controlled by an attacker.

AI EDITION

The Newsjacking Phase: Exploiting the AI Ad Rush

While abusing features like Artifacts represents the sophistication of these attacks, broader telemetry shows that attackers are using this infrastructure shift to supercharge high-volume impersonation campaigns. Between April and August 2026, KnowBe4 Defend detected a **228% surge in AI-brand impersonation attacks**.

The initial spike launched in April when analysts uncovered campaigns blasting fake “AI Advertising Program Invitations.” Threat actors were opportunistically capitalizing on legitimate industry news:



OpenAI and ChatGPT

Attackers exploited widespread industry buzz and beta tests surrounding ChatGPT sponsored placements and free-tier ad monetization.



Google Gemini

Campaigns piggybacked on Google expanding AI-driven campaign tools and Gemini integrations across search overviews.



Anthropic (Claude)

Anthropic explicitly committed to keeping Claude ad-free. However, threat actors spoofed Claude ad invites anyway, betting that busy marketing teams wouldn't know Anthropic's policy and would assume all major AI vendors were launching ad networks simultaneously.

This topical hook drove massive engagement, reflected in our telemetry:



228%

surge in AI-brand impersonation attacks



8x

growth in AI-themed subject lines



2x

growth in AI-brand lookalike links

As attackers continue to merge topical newsjacking with direct, on-platform hosting, the line between legitimate enterprise AI usage and malicious infrastructure has vanished. When real TLS certificates and trusted AI domains no longer guarantee safety, security teams must shift toward **continuous API key rotation, strict conditional access policies for agentic frameworks and behavioral monitoring for background platform traffic**.

What Is Evading SEG Detection?

Phish That Evaded SEGs

41.3%

Peaked at 48.3% in July 2026

Density Through Traditional Solutions

1 in 692

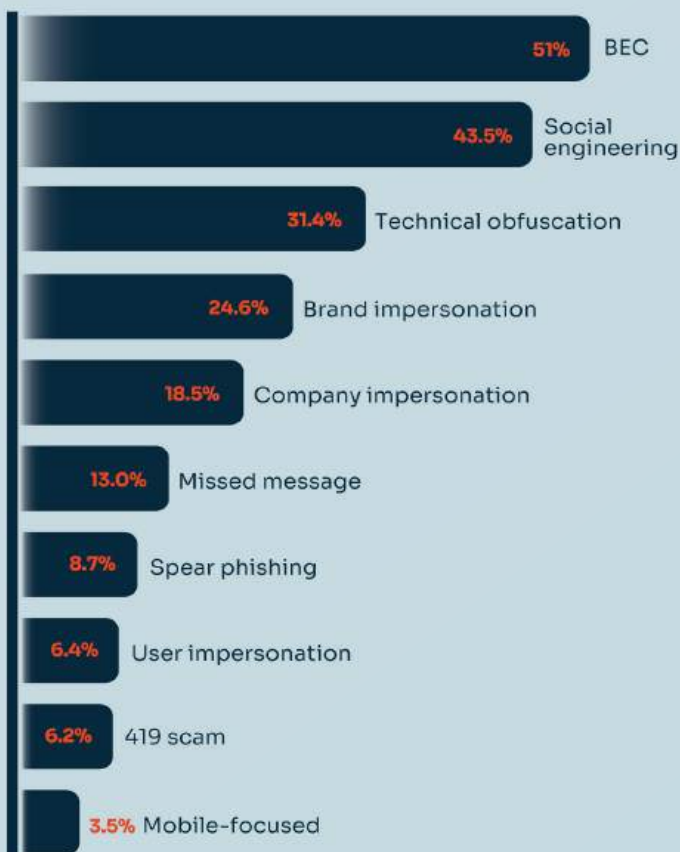
51%

Of the Phish SEGs Miss Are BEC

58.1%

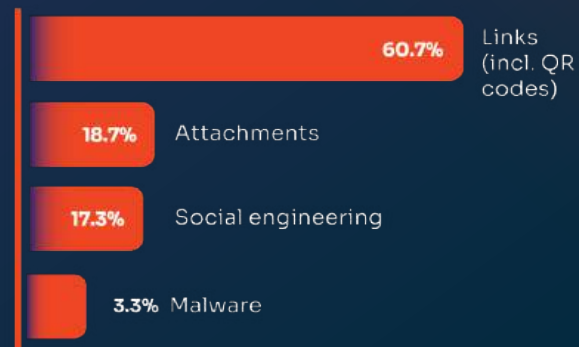
Of the Phish SEGs Miss Pass DMARC

Phish Types: % SEG Has Missed*



*Not mutually exclusive, will sum over 100%

Method: % SEG Has Missed



Wrapping Up: Staying One Step Ahead in a Shifting Landscape

We hope **Volume 8 of the Phishing Threat Trends Report** has brought the rapidly shifting adversary landscape into sharp focus. Our mission remains clear: to arm your organization with actionable intelligence so you can build an adaptive defense that protects your people, your data and your brand.

If 2026's trends have shown us anything, it's that attackers are operating across a global footprint while executing with unprecedented speed, regional precision and commercial scale. Threat actors continue to master social engineering by weaponizing the news cycle in real time: exploiting salience bias to turn breaking headlines and major planned events into instant, high-trust inbox lures. At the same time, by unmasking non-relay staging infrastructure and mapping digital trade routes, we have shown how adversaries tailor their delivery mechanisms to specific geographic vulnerabilities and industry workflows.

Critically, the defining shift in this report is that attackers are expanding their playbook; no longer just stealing passwords, but stealing permissions too. Through device code abuse and illicit OAuth consent grants, adversaries are converting genuine user actions into persistent cloud access that password resets cannot revoke. As organizations embrace autonomous tooling, this threat has expanded directly onto legitimate AI platform infrastructure. By hosting attacks on trusted domains and harvesting API keys, threat actors are bypassing traditional filters to subvert agentic workflows and compromise the systems powering the modern enterprise.

Future-Proofing: Human and Agentic Risk

Standard secure email gateways and native mail filters were built for a threat landscape that no longer exists. As adversaries abuse genuine cloud infrastructure, deploy specialized reverse-proxy toolkits, and hijack trusted AI platforms, perimeter-based defenses simply cannot keep pace with localized, highly targeted attacks. Your security stack needs to be as agile, adaptive and geographically aware as the threat actors themselves.

By moving toward an integrated platform fueled by deep behavioral analytics and real-time threat intelligence, your security posture shifts from reactive filtering to continuous, automated defense. Delivering real-time behavioral coaching at the exact moment of risk transforms your workforce into an active line of defense, securing both your human workforce and your autonomous AI tooling across every global operation.

Ready to level up your defense? The KnowBe4 team is here to help you turn these insights into action. Let's keep the conversation going and ensure your organization stays ahead of the curve.

Get In Touch



Our Contributors



Jack Chapman

SVP of Threat Intelligence

Jack leverages deep insights of the cyber-threat landscape and his extensive R&D skillset to oversee threat research and AI development for KnowBe4 Defend to stop the advanced phishing attacks that defeat traditional security solutions. Jack maintains close ties with the global cyber community, particularly the UK's intelligence and cyber agency GCHQ.



James Dyer

Head of Threat Intelligence

James spearheads the Threat Intelligence team at KnowBe4, spending his days uncovering the latest phishing threat trends, understanding emerging methodologies and analyzing the TTPs of the crime-as-a-service ecosystem.



Lucy Gee

Snr. Cybersecurity Threat Researcher

Lucy is passionate about the intersection of psychology and cybersecurity and the use of behavioral insights to enable people to live and work securely. At KnowBe4, Lucy analyzes the latest phishing campaigns and communicates emerging trends to business stakeholders.



Cameron Sweeney

Cybersecurity Threat Researcher

Cameron specializes in understanding the technical aspects of cyberattacks. As a member of the KnowBe4 team, he reverse engineers phishing attacks and malware to identify emerging threats, using statistical analysis to track the evolving threat landscape.



Louis Tiley

Cybersecurity Threat Researcher

Louis researches diverse attack vectors, social engineering tactics, and emerging threats. At KnowBe4, he analyzes phishing campaign methodologies and builds tools to automate threat intelligence gathering to identify industry trends and shape cybersecurity messaging.

About KnowBe4 Defend

An integrated cloud email security solution, Defend delivers AI-powered behavioral-based detection to eliminate the attacks that get through native security and secure email gateways. Leveraging zero-trust and pre-generative models, Defend provides the highest efficacy of detection against advanced threats, including zero-day and emerging attacks, phishing emails sent from compromised accounts, and social engineering. Using dynamic banners applied to neutralized threats, Defend provides real-time teachable moments that continually “nudge” employees into good security behaviors to tangibly reduce risk and augment security awareness.

About KnowBe4

KnowBe4 empowers the modern workforce to make smarter security decisions every day. Trusted by more than 70,000 organizations worldwide, KnowBe4 is the pioneer of whole workforce security, securing both AI agents and humans. The KnowBe4 Platform provides attack simulation and training, collaboration security and agent security powered by AIDA (Artificial Intelligence Defense Agents) and a proprietary Risk Score. The platform leverages 15 years of behavioral data to combat advanced threats including social engineering, prompt injection and shadow AI. By securing humans and agents, KnowBe4 leads the industry in workforce trust and defense.

More information at [KnowBe4.com](https://www.knowbe4.com).



Follow us on
LinkedIn

Follow us on 



KnowBe4, Inc. | 33 N Garden Ave, Suite 1200, Clearwater, FL 33755
855-KNOWBE4 (566-9234) | www.knowbe4.com | Sales@KnowBe4.com

Other product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies.