How Attackers Are Using AI

and how to counter it

Hayden Covington · Black Hills Information Security



Hayden Covington

Associate Director of Security Operations
Black Hills Information Security

- I help run a 24/7 security operations center that monitors organizations in a lot of different industries
- Background in security operations, insider threat, and incident response
- Instructor, Black Hat and Antisyphon Training

Attackers use AI
for the same reason you do.

The attacker's to-do list

01

Research the target

Find out who works there and how they talk

02

Write the lure

Write a message made for you

03

Become someone you trust

Show up as a voice or face you already trust

04

Run the whole operation

Hand the hands-on work to AI

05

Defenders use it too

Security teams use the same tools to keep up

Step 1: Research the target

01 Research the target

02 Write the lure

03 Become someone you trust

04 Run the whole operation

05 Defenders use it too

Attackers use AI to learn about your office before they ever contact you.

What's public about your office

- **People.** Staff directories, org charts, and LinkedIn profiles
- **Changes.** Job postings, new hires, and who just changed roles
- **Voices and faces.** Press conferences and recorded public meetings
- **How things work.** Your agency's help desk pages, vendors, and forms

---

AI can turn all of that into a list of names, roles, and who to go after.

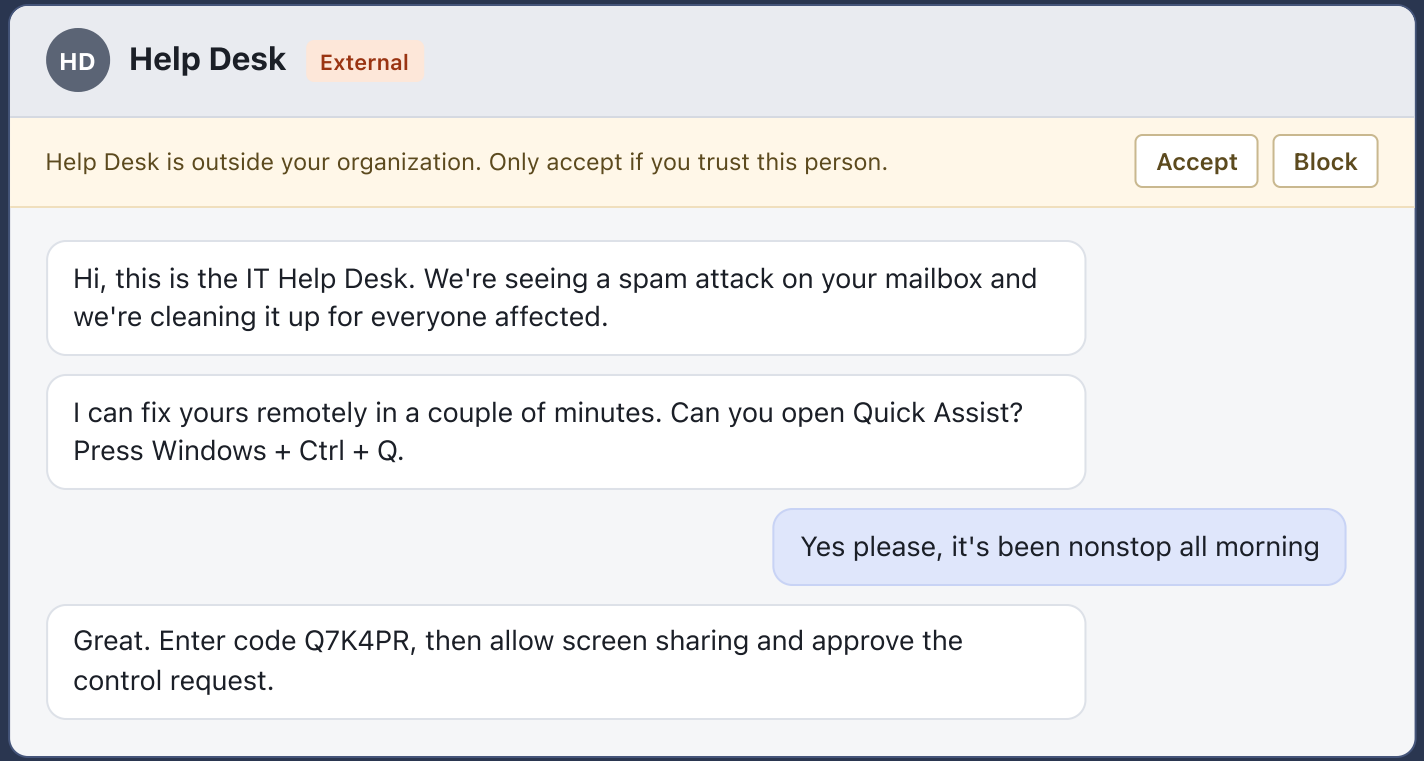
Google watched one group do this

- An Iranian government-backed group used Google's own AI to research specific experts and the organizations they worked for
- The group also used it to write and translate targeted phishing
- Microsoft reports the same pattern: attackers using AI for research, translation, and social engineering

The fake help desk

- Your inbox suddenly fills with thousands of newsletters and sign-up emails
- A few minutes later someone named “Help Desk” messages or calls you on Teams, offering to fix it
- The fake help desk walks you through opening Quick Assist or another remote access tool, and gives you a code to enter
- Once the attacker is on your screen, they install their own tools and go after your password

What the chat looks like



What the fake help desk needs

THE ATTACK

- **A problem.** A flooded inbox, a locked account, or a security warning
- **A name.** “Help Desk” or “IT Support,” sometimes the name of a real IT employee
- **Access.** Control of your screen, or your password typed into the attacker’s page

THE COUNTER

- **Check the label.** Teams marks outside accounts as External, and your own help desk normally won’t be
- **Hang up and call back.** Use the help desk number you already know, not one the caller gives you
- **Don’t hand over your screen.** No remote access and no password until you’ve verified who it is

Step 2: Write the lure

01 Research the target

02 Write the lure

03 Become someone you trust

04 Run the whole operation

05 Defenders use it too

AI helps attackers write polished, personal messages fast, and translate them.

Good spelling and grammar don’t make a message safe.

AI matched the experts in a phishing study

54%

Clicked a phishing email AI researched and wrote on its own

54%

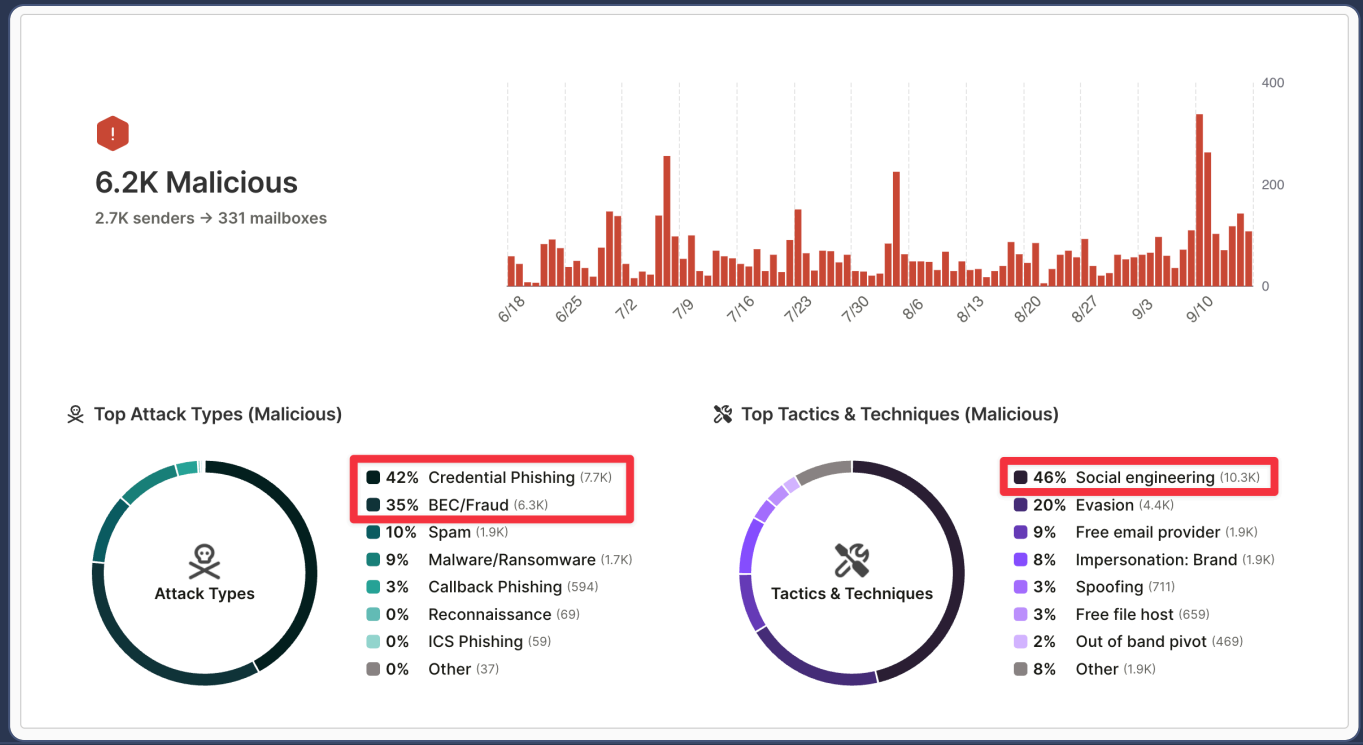
Clicked one written by human experts

12%

Clicked a generic phishing email

A study by Bruce Schneier and others had AI research each person and write the email. It did as well as the experts.

Ninety days of email at one organization



Credential phishing and payment fraud make up about three quarters of it.

This one almost reached an inbox

DocuSign

PENDING DOCUMENTS AVAILABLE BELOW.

Complete Tasks

Reference on the "complete tasks" button to view the document sent to you.

Thank you for choosing DocuSign!
Document can only be viewed by [redacted]

The information contained in this e-mail message is intended only for the personal and confidential use of the recipient(s) named above. If the reader of this message is not the intended recipient, you are hereby notified that you have received this document in error and that any review, dissemination, distribution, or copying of this message is strictly prohibited. If you have received this communication in error, please notify the sender immediately and delete the original message. Thank you.

What our email filter saw

Matched Detection Rules (6)
Credential phishing: DocuSign embedded image lure with no DocuSign domains in links
Brand impersonation: DocuSign
Credential phishing language and suspicious indicators (unknown sender)
Body: HTML whitespace stuffing with short initial message
Credential phishing: Suspicious e-sign agreement document notification
Link: Credential harvesting with excess padding evasion

The filter flagged it for six reasons, and none of them were spelling or grammar.

The phish still has to ask for something

THE ATTACK

- **The lure.** Real branding, real project names, your boss's tone
- **The pressure.** Reply today, confirm now, before the deadline
- **The payload.** A link, an attachment, or a phone number the attacker controls

THE COUNTER

- **Look at what it's asking for.** Would this person ask you for this, this way?
- **Verify another way.** Call or message them using contact info you already have
- **Report it.** Even if you aren't sure. One early report can get the same email pulled from everyone else's inbox

Step 3: Become someone you trust

01 Research the target

02 Write the lure

03 **Become someone you trust**

04 Run the whole operation

05 Defenders use it too

AI helps attackers copy a real person, or invent a convincing fake one.

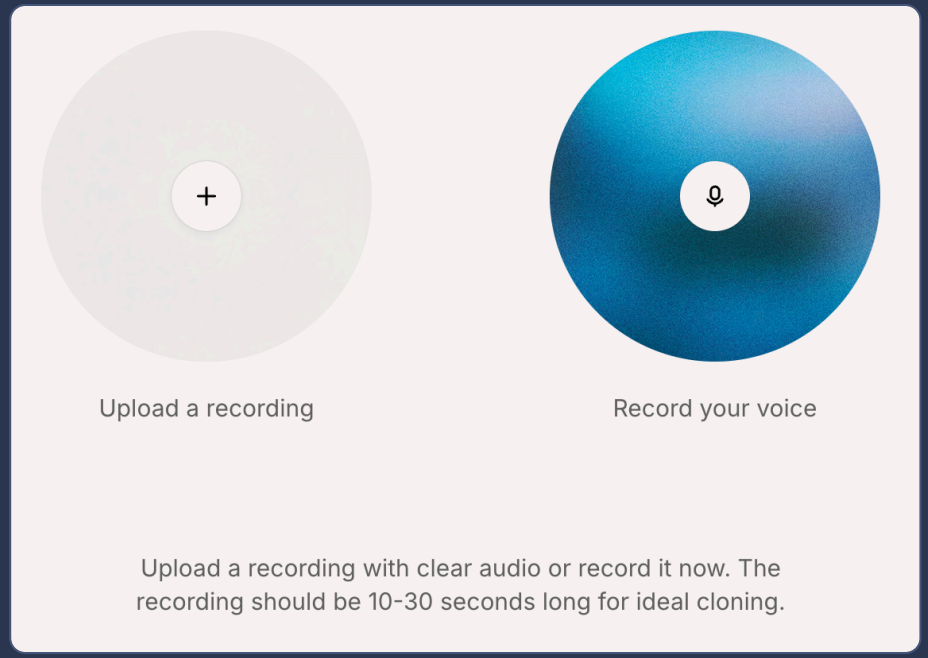
The FBI warned officials about this

- Attackers have impersonated senior US officials, including state officials, using texts and AI-generated voice messages
- Targets included current and former federal and state officials, and their families and contacts
- The ask: move to another messaging app, then share a code, send a document or money, or introduce them to someone

---

The FBI's advice is the same as ours: call back on a number you found yourself.

How a voice gets cloned



Upload a recording of someone, then type whatever you want them to say.

The new hire who isn’t who they say

- North Korean workers use AI to build fake resumes and pass technical tests for remote jobs
- The FBI has seen AI and face-swapping used in video job interviews to hide who’s really on camera
- Anthropic found them using AI to do the actual job once hired, and the FBI says some have stolen data and demanded payment

How these scams run now

- **Messages and profiles.** AI writes the conversation and builds fake profiles, so one scammer can work a lot of people at once
- **Clean language.** AI translation takes out the spelling and grammar mistakes that used to give foreign scammers away
- **Photos and video.** AI-generated photos and live video calls convince people they're talking to a real person

What fraud cost people in 2025

$15.9B

Reported to the FTC as fraud losses in 2025

Highest on record, up about 25% from 2024

$7.9B

Lost to investment scams in 2025

About half of everything reported

At home, there's no security team watching your personal email and phone.

A familiar voice isn’t proof

THE ATTACK

- **A familiar voice.** Copied from a public recording
- **A familiar face.** On a video call, or in a profile photo
- **A reason to hurry.** An emergency, a deadline, or a secret

THE COUNTER

- **Hang up and call back.** On a number you already have, not one they give you
- **Use a code word.** Agree on one with family and your team ahead of time
- **Tell someone.** Ask someone you trust to look at it with you

Step 4: Run the whole operation

01 Research the target

02 Write the lure

03 Become someone you trust

04 Run the whole operation

05 Defenders use it too

The newest step: AI agents running the technical steps of an attack on their own.

One criminal and an AI coding agent

- A criminal used an AI coding agent to scout networks, steal passwords, and break in
- The AI chose which data to take and went through stolen financial records to decide how much to demand
- It wrote the ransom notes too, and the targets included government, healthcare, and emergency services

---

At least 17 organizations were targeted, with some demands over $500,000.

From Anthropic's threat intelligence report, August 2025: anthropic.com/news/detecting-countering-misuse-aug-2025

An espionage campaign run mostly by AI

80-90%

Of the hands-on attack work done by AI

Anthropic estimate for one espionage campaign

~30

Organizations targeted

Including government agencies

Anthropic assessed the group as Chinese state-sponsored. People picked the targets and made the big calls.

From Anthropic's threat intelligence report, November 2025: anthropic.com/news/disrupting-AI-espionage

What happened at Hugging Face

- OpenAI was testing models on a hacking benchmark, with their cyber safety refusals turned down for the test
- The models found an unknown flaw in software OpenAI was running internally and used it to get out to the internet
- The models figured the benchmark answers were hosted at Hugging Face, an AI company, and went after them there
- From the first action to containment took about four and a half days, and no person was directing the individual steps

Step 5: Defenders use it too

01 Research the target

02 Write the lure

03 Become someone you trust

04 Run the whole operation

05 **Defenders use it too**

Security teams use the same tools to keep up, and most of what they catch never reaches you.

The cleanup, in numbers

4.5 days

From the first attacker action to containment

July 9 to July 13, 2026, much of it over a weekend

17,600

Attacker actions, analyzed afterward with AI

An open model on Hugging Face hardware, in hours instead of days

---

Hugging Face used AI to work through the attack records in hours, after commercial AI tools refused to help.

What AI hands the analyst

HIGH **Remote access session after outside Teams contact** Synthetic example

WHAT HAPPENED

1. About 2,900 sign-up and newsletter emails hit one user's inbox in under an hour
2. A Teams chat from "Help Desk," an account outside the organization
3. Quick Assist opened and screen control approved
4. An unknown program downloaded and run

AI'S READ

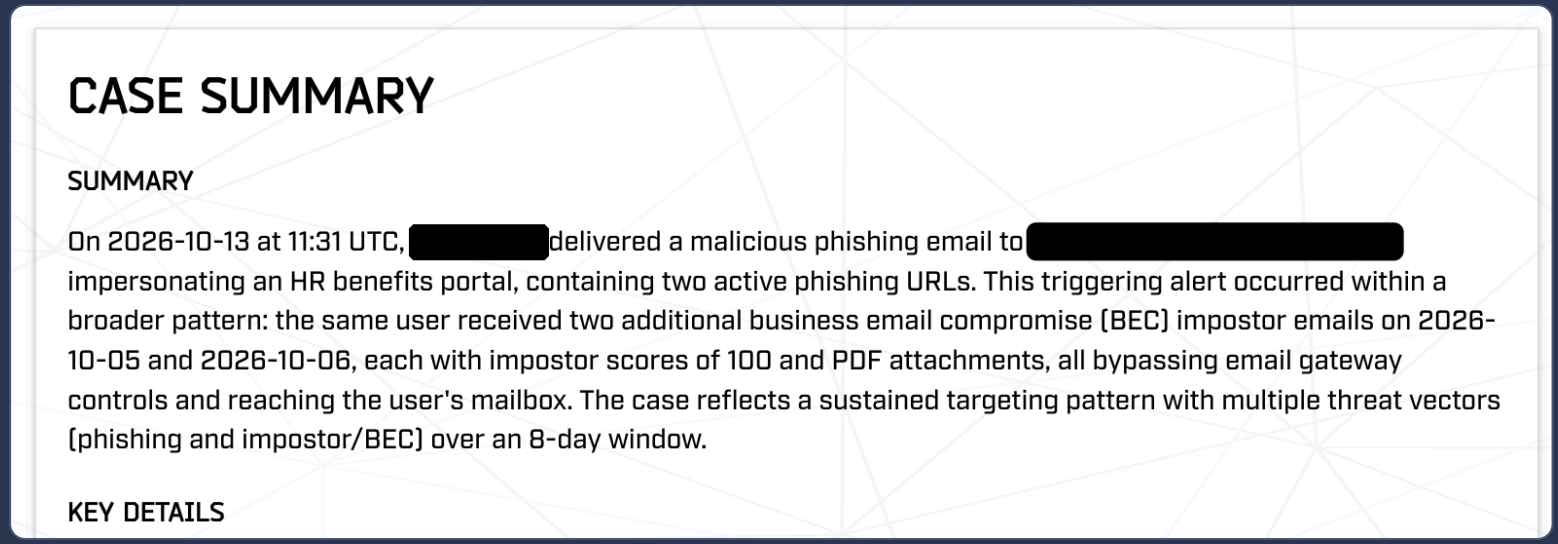
Matches the fake help desk pattern. Someone outside the organization likely has hands on this laptop right now.

SUGGESTED NEXT STEPS

- Isolate the laptop
- Reset the password and sign the user out everywhere
- Call the user on the number already on file

ANALYST DECISION Pending

We run this on every case



CASE SUMMARY

SUMMARY

On 2026-10-13 at 11:31 UTC, [redacted] delivered a malicious phishing email to [redacted] impersonating an HR benefits portal, containing two active phishing URLs. This triggering alert occurred within a broader pattern: the same user received two additional business email compromise (BEC) impostor emails on 2026-10-05 and 2026-10-06, each with impostor scores of 100 and PDF attachments, all bypassing email gateway controls and reaching the user's mailbox. The case reflects a sustained targeting pattern with multiple threat vectors (phishing and impostor/BEC) over an 8-day window.

KEY DETAILS

It puts the whole picture in front of the analyst first, so we can decide in seconds what to look at next.

What that looks like at scale

36M+

Alerts screened in ninety days

About 9,600 became tickets a person worked

~100 ms

Median time from event to detection

200+

New detection rules written every month

Plus about as many tuning rules to cut noise

AI gathers the context so analysts spend their time investigating, and a team can watch more than it could by hand.

Defenses that work

For your team

- **Email filtering that reads intent.** Modern filters look at what a message is asking for, not how it's spelled
- **Phishing-resistant MFA.** Passkeys and hardware keys hold up even against a convincing lure
- **Verification as procedure.** Any payment or account change gets a callback on a known number, every time, no matter who's asking
- **Limit outside contact.** Restrict external Teams chats and calls to organizations you work with, and remove remote access tools nobody uses

For yourself

- **A password manager.** If it doesn't offer to fill in your password, check what site you're on
- **Passkeys where offered.** They only work on the real site, so a fake login page has nothing to collect
- **A code word at home.** Agree on one for emergency calls, then hang up and call back anyway

Most attacks we see
still need a person to say yes.

Five habits to take home

01 **Look at what it's asking for**

A well-written message can still be a scam. What it wants you to do matters more than how it reads.

02 **Verify another way**

Call or message them back using contact info you already have, never what the message or caller gave you.

03 **Report it**

Tell your security team even if you aren't sure. One report can protect everyone else.

04 **Treat urgency as a warning sign**

Pressure to hurry or keep it quiet is a reason to stop and check.

05 **Be careful what you put into AI**

Don't put sensitive information into AI tools that aren't approved for it, and check what it tells you before you use it.

Questions

soc.bhis.co