AI Threats & Deepfake

How AI Social Engineering Expanded the Attack Surface

Breacher.ai Intelligence

The Security Paradigm Shift

Traditional Security

Focused on network perimeters, firewalls, and physical barriers

Modern Reality

Attackers target the application layer—where your users operate

Traditional Perimeter vs. Layer 7

Network Perimeter

- Firewalls & VPNs
- Intrusion Detection
- Physical Access Control

Assumption: Keep attackers OUT

Layer 7 (Application)

- User Browsers
- Email & Messaging
- Human Decision-Making

Reality: Attackers target USERS

The attack surface is expanding to where users interact

AI Social Engineering: The New Attack

1. **Phishing & Spear Phishing**
   Deception that trick users into performing actions.

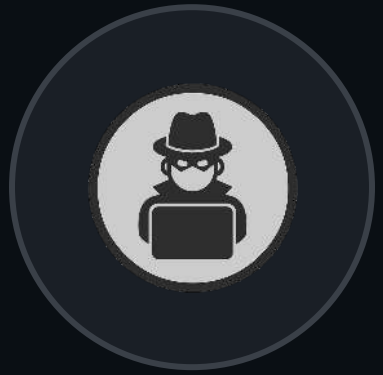
2. **Deepfake & AI-Powered Attacks**
   Synthetic voice and video calls impersonating executives

3. **Credential Harvesting**
   Fake login pages that capture usernames and passwords

4. **Business Email Compromise (BEC)**
   Impersonating executives to authorize fraudulent wire transfers

Most security breaches involve some level of human error

Luring Out vs. Breaking In



Breaking In (Old Way)

Finding vulnerabilities in firewalls, exploiting network weaknesses, brute-force attacks

Time-consuming & noisy

→

Luring Out (New Reality)

Convincing users to willingly hand over credentials, click links, or take actions that bypass security

Fast, effective & hard to detect

Social engineering exploits trust, not just technology

Snapshot - What we're Seeing.

Multiple
Organizations Assessed

300+
Total Targets

19%
Avg Click Rate

Finance
Mixed
Most heavily targeted sector

Manufacturing
High Risk
Highest avg action rate observed

Tech
Mixed
Wire transfer incident observed

Critical Finding

While traditional training reduces susceptibility by approximately 35%, organizations remain vulnerable to advanced AI-powered attacks. Legacy training platforms don't adequately prepare employees for deepfake and agentic AI threats.

Impact of Prior Training Programs

No Prior Training

13%

Average action rate

Action Rate* User Does Something like Input Credentials or Wire Transfer

With Prior Training

8%

Average action rate

Reduction

~35%

Improvement with training

Training Platform Distribution

Platform A - Legacy	Platform B - Modern	Platform C - NextGen
Multiple orgs	Multiple orgs	Limited orgs

Critical Finding

While traditional training reduces susceptibility by approximately 35%, organizations remain vulnerable to advanced AI-powered attacks. Legacy training platforms don't adequately prepare employees for deepfake and agentic AI threats. Awareness training helps, but not a silver bullet.

Detailed Assessment Results

Organization	Vertical	Size	Vector	Action	Click	Training Platform
Org A	Law	Medium	WhatsApp Voice	Low	Moderate	In House
Org B	Tech	Small	Deepfake Video+Email	**Critical**	**High**	No
Org C	Mfg	Medium	Deepfake Call+SMS	Moderate	High	Platform A - Legacy
Org D	Mfg	Medium	Deepfake Voicemail	Moderate	N/A	Platform A - Legacy
Org E	Finance	Small	Deepfake Call+SMS	N/A	High	Platform B - Modern
Org F	Finance	Medium	Deepfake Call+Email	N/A	Low	Unknown
Org G	Finance	Large	Deepfake Call+SMS	Moderate	Moderate	Platform B- Modern
Org H	Finance	Small	Deepfake Call+AI	Low	N/A	Unknown
Org I	Finance	Small	Deepfake Calendar	Low	Moderate	Platform C - NextGen

Critical Incidents & High-Risk Cases

⚠️ CRITICAL: User-Initiated Wire

Transfer
Tech Company - Finance Dept

Attack Vector
Deepfake Video + Email Combo

Click Rate
High

Prior Training
None

Employee attempted wire transfer in response to deepfake attack. Most severe outcome—financial impact narrowly avoided by red team controls.

Other High-Risk Cases

Manufacturing Org
Industrial Sector
High Click + Credentials despite active training

Finance Org
Investment Sector
High click despite active training

Impact Analysis

Wire transfer attempt shows AI-powered attacks can bypass traditional defenses.

Attack Vector Ranking

Top 3 most effective attack combinations

#1

Deepfake Video + Agentic Email

Video Deepfake → Agentic AI → Email

Visual authenticity bypasses traditional skepticism entirely

33.0%
CLICK RATE

21.78% action

#2

Deepfake Phone Call + Agentic SMS

Voice Deepfake → Agentic AI → SMS

Phone call establishes trust, SMS creates urgency

23.0%
CLICK RATE

14.75% action

#3

Deepfake Calendar Invite + Agentic AI

Voice Deepfake → Agentic AI → Calendar

Calendar invites feel inherently legitimate

13.8%
CLICK RATE

9.54% action

Department Risk Analysis

Average click rates by department

#1 Finance

Highest vulnerability to video deepfakes

22.90%
AVG CLICK RATE

#2 Human Resources

Susceptible to calendar-based attacks

16.65%
AVG CLICK RATE

#3 Company-Wide

Baseline organizational risk

14.04%
AVG CLICK RATE

Finance employees are 63% more likely to click than the company average. Prioritize targeted training for financial roles.

OBSERVATIONS IN THE WILD

Voicemail Drop + SMS Combo

ATTACK CHAIN

1. **Ringless Voicemail Drop**
   Deepfake audio deposited directly — no ring, no missed call alert

↓

2. **Wait for Transcription**
   iPhone auto-transcribes voicemail in 2-3 minutes

↓

3. **Send SMS with Link**
   Link appears legitimate — prior "interaction" established

Why It Works

The voicemail creates perceived prior contact. When the SMS arrives, the target believes they've already interacted with the sender.

SECURITY BYPASS

Bypasses iPhone Safe Links controls — the link goes hot because iOS sees it as a trusted conversation thread.

Devastating

Highly effective in the wild

The Uncanny Valley

Deepfake vs Agentic AI threat landscape Observations (Empirical Data)

DANGER TODAY

62%

Deepfake

Many users struggle to identify deepfakes. Visual and audio authenticity creates unearned trust.

RISING

20%

Agentic AI

Most can spot it today.

CURRENT THREAT EFFECTIVENESS

The Future is Frightening

Agentic AI will surpass Deepfake within the next year. Adaptive scenarios and real-time social engineering are evolving rapidly.

Key Takeaway

Deepfake is the clear danger now — but Agentic AI is a rapidly developing threat that demands attention.

Deepfake Detection: Human Performance Results

WORSE THAN RANDOM GUESSING

38%

Average accuracy (2.66/7 correct)

Random guessing = **50%**

The Insight: People aren't just bad at spotting deepfakes — they're being actively deceived by them.

PERFORMANCE BREAKDOWN

Failed (<4 correct)	70.5%
Struggled badly (≤2 correct)	49.5%
Did well (≥5 correct)	10.8%

~1 in 2

got ≤2 correct

1 in 10

can reliably detect

Key Insight: Training People to “Spot” Deepfakes is a flawed approach.