HANDOUT 7 - WEEK 8 READING

AI-Powered Attacks

CIS 350: AI for Cybersecurity - Smart Defense for the Digital Business

Colorado State University - Computer Information Systems

How to use this reading (and an ethics note)

This handout supports Week 8. We study how attackers use AI - phishing, deepfakes, credential stuffing, reconnaissance, and adaptive malware - ONLY so we can detect and defend against them. Nothing here is a how-to for attacking; every technique is paired with a defense. The figures and code match Lab 7 (a defensive detection lab); the review questions map to Quiz 7.

Estimated reading time: 20-25 minutes. No prior coding background needed.

1. AI Is a Force Multiplier - For Both Sides

The same AI that helps defenders also helps attackers. It rarely invents brand-new attacks; instead it makes OLD attacks cheaper, bigger, more personalized, and more adaptive. A single attacker can now produce thousands of unique, believable lures, clone a voice from a few seconds of audio, or auto-scan the internet for weak systems. Understanding this is the first step to defending.

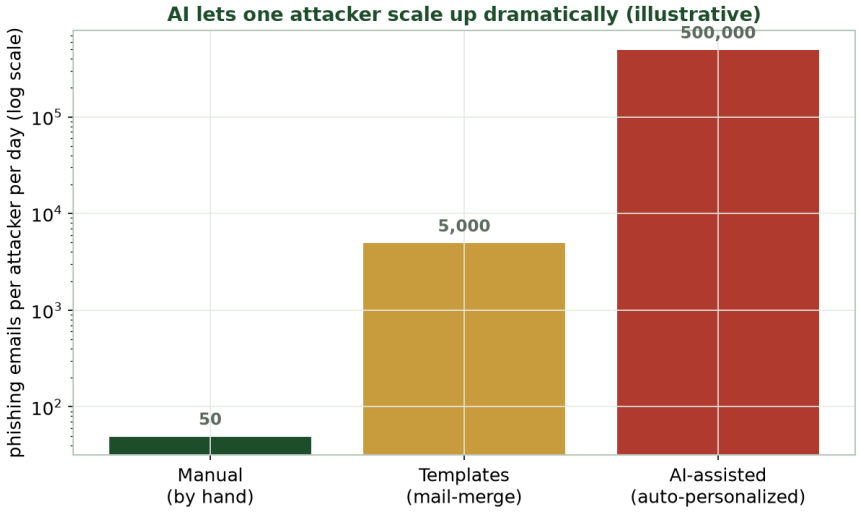

Figure 1. Illustrative: automation and AI take one attacker from dozens of hand-written lures a day to hundreds of thousands.

Why defenders study attacker AI

'Know your enemy.' You cannot detect what you do not understand. Every technique in this handout is paired with a concrete defense - many of which you already learned (MFA, ML email filters, anomaly detection).

2. Attacks That Fool People

Most attacks still target a person's judgment. AI makes the human-facing tricks far more convincing:

- AI phishing: fluent, personalized lures at scale - the old 'spot the typo' advice no longer protects.
- Business Email Compromise (BEC): fake executive/vendor emails asking finance to pay - among the costliest cybercrimes.
- Deepfakes: AI-generated fake voice/video; voice-clone calls impersonate a 'boss' or family member.
- Credential stuffing: bots reuse leaked passwords across sites, betting on password reuse.

Defenses for people-facing attacks

Phishing -> AI email filters + user training. Deepfake/voice fraud -> verify out-of-band (call a known number, use a code word), distrust urgency. Credential stuffing -> MFA (and passkeys), breached-password checks, rate limiting.

3. Attacks That Fool Machines

AI also helps attackers probe systems and evade our tools. Automated reconnaissance scans the internet and profiles targets from public data, prioritizing the easiest way in. Adaptive malware mutates its code and behavior to slip past signature-based antivirus. Attacker automation compresses the timeline, leaving defenders less time to react.

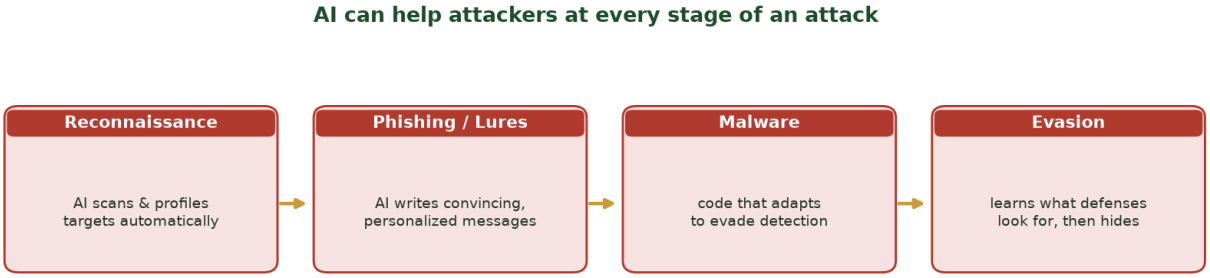

Figure 2. AI can assist an attacker at every stage - reconnaissance, phishing, malware, and evasion. Breaking any one stage disrupts the attack.

- Reconnaissance -> reduce your exposure, monitor for scanning, and patch quickly.
- Adaptive malware -> behavior-based detection and anomaly detection (Week 6), not signatures alone.

4. Detecting AI Phishing (Lab 7)

Better wording doesn't hide everything: AI phishing still tends to include links, urgent language, credential requests, and sender mismatches. Defenders turn each email into these numeric FEATURES and train a classifier - the same supervised workflow used for spam (Weeks 3-4).

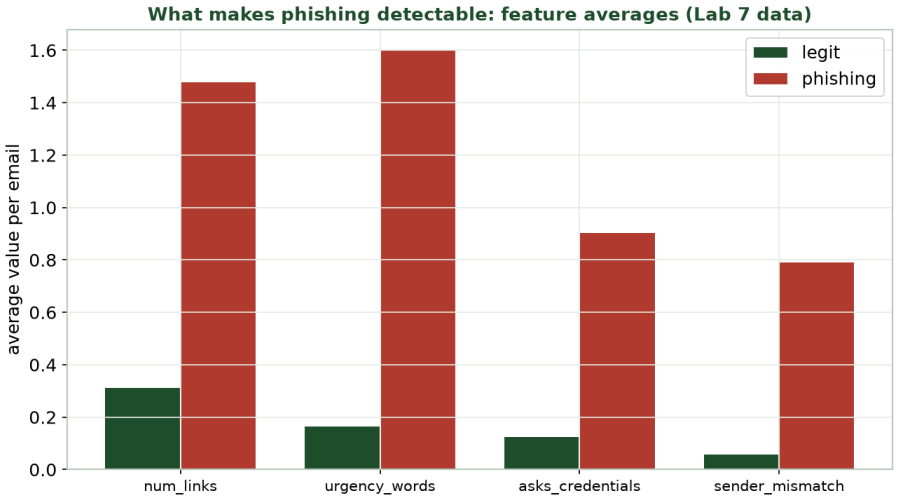

Figure 3. Phishing emails average more links, urgency words, credential asks, and sender mismatches than legit ones - so a simple detector can separate them (this is the Lab 7 data).

```
# Turn an email into features, then train a phishing DETECTOR (defense)
X = emails[['num_links','urgency_words','asks_credentials','sender_mismatch']]
y = emails['label']   # 'phishing' or 'legit'
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression(max_iter=1000).fit(X_train, y_train)
print(accuracy_score(y_test, clf.predict(X_test)))   # ~0.96 in Lab 7
```

But detection is not perfect: it has false positives (legit flagged) and false negatives (phishing missed), and as attackers improve, false negatives rise. Detection is one layer, not the whole answer.

5. Every Attack Has a Defense: Defense in Depth

No single control stops AI attacks. Strong security layers technology, process, and people so a gap in one is covered by another - and assumes some attacks get through, investing in detection and response.

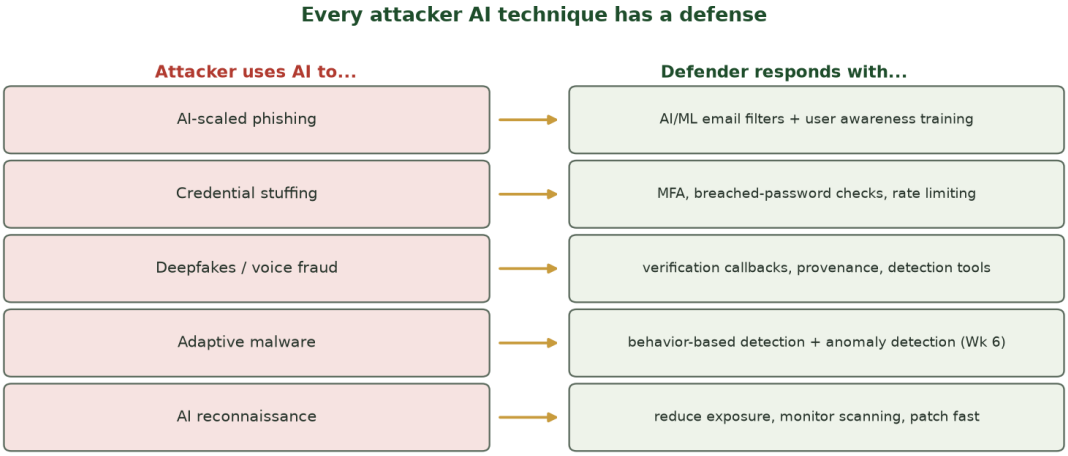

Figure 4. Each AI-enabled technique maps to a concrete countermeasure - most of which you already know.

The human layer is often decisive

Most AI attacks still target a person. Simple habits beat sophisticated fakes: verify through another channel, slow down when rushed, and report suspicious messages. Technology assists human judgment - it does not replace it.

Key Terms

- AI phishing - fluent, personalized lures produced at scale.
- Business Email Compromise (BEC) - impersonation emails targeting payments.
- Deepfake - AI-generated fake voice/video/image of a real person.
- Credential stuffing - reusing leaked passwords across sites with bots.
- Reconnaissance - automated scanning/profiling of targets.
- Adaptive malware - code that changes to evade signature detection.

- Multi-factor authentication (MFA) - a second factor beyond the password.
- Defense in depth - layered controls across people, process, and technology.

Review Questions (self-check for Quiz 7 and Lab 7)

- Name two ways attackers use AI, and a defense for each.
- Why is 'look for typos' weak advice for spotting phishing now?
- What is credential stuffing, and what is the best defense?
- How can defenders detect AI phishing, and why isn't detection enough?
- What is defense in depth? Give a people, a process, and a technology control.

Remember: AI tools are not permitted on quizzes and labs; the Week 8 material is for DEFENSE only.

References

- CISA, Phishing Guidance: Stopping the Attack Cycle at Phase One, 2023.
- MITRE ATLAS - Adversarial Threat Landscape for AI Systems; MITRE ATT&CK.
- NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023; SP 800-63B (authentication).
- FBI IC3, Internet Crime Report (business email compromise, deepfake fraud).
- Verizon, Data Breach Investigations Report (DBIR), annual.
- OWASP - phishing and credential-stuffing prevention cheat sheets.