How AI Transforms Fragmented Public Data into Targeted Cyber Threats

Submitted by:

Sohith Vishnu Sai Yachamaneni
University of Zürich
Submitted for Cyber Law – Data protection, AI, and Cybersecurity

February 18, 2026

1 Introduction

Let me introduce a scenario: Assume yourself who is very active in the current digital space. As a individual has a "Standard" public digital footprint: a Linked-ln profile for networking, a X account for industry news, a public history reviews on amazon or technical forums. Individually, these are fragmented, harmless pieces of public information intended for connection and expression, How can this information can effect me, this harmless pieces of information which is not sensitive to me. I would argue to think again, to a human bad-actor this information is good enought to profile you. While a human bad-actor might read a few posts to guess a password, modern AI can ingest thousands of these public data points to build a Psychometric Profile.

AI does not simply identify who the target is; by analyzing personality traits, it tells the attacker exactly how to speak to them. (Imagine this efficiency, if companies applied this same level of psychological insight to customer care).Until recently, finding these exploitable elements was a slow, manual process. Today, automation allows attackers to scale this deep analysis instantly, making them significantly more dangerous.

The Attack Mechanism: The threat usually manifests as an AI-generated phishing email that creates a perfect mirror of a trusted connection's writing style. The target is 'hooked' because the message is hyper-personalized: knowing the subject has 'High Agreeableness' and is focused on a specific current topic, the AI frames the impersonation with extreme precision.

Does this really work ? Yes, the breach does not occur because of weak password, but because the AI utilized public data to bypass the human firewall by mirroring the target's own psychology. Here, we are discussing a combination of fields, Artificial Intelligence, Cyber Security, Data and Human Psychology working together to bypass the human firewall. Until now the actors who is able to pull such attacks are well trained and well resourceful groups, but with raise of AI people with minimal training and resources can initiate this type of attacks.

2 The Convergence of Risk: AI, Data, and Cyber Security

With our scenario, we established how the 3 three holy grails AI, Data and Cyber Security can bypass a human firewall. Now lets look at how traditional attacks has been overpowered with support of AI tools.

2.1 The Threat of Weaponized OSINT and Data Deanonymization

While Phishing is often the most visible vector of a cyber-attack, it is merely the final execution of a deeper, more pervasive threat: **Weaponized Open Source Intelligence (OSINT)**.

Traditionally, OSINT is a tool used by researchers and journalists for transparency. However, in the context of Artificial Intelligence, Data and Cyber Security , it possesses a dangerous dual nature. The same tools designed to expose corruption are now being exploited by malicious actors to construct high-fidelity target dossiers. This phenomenon represents a shift from "Passive Observation" to "Active Targeting."

2.1.1 The Mechanism of Weaponized OSINT

"Weaponized OSINT" refers to the systematic aggregation of publicly available data to identify human vulnerabilities. Unlike a technical exploit that scans for open ports on a server, Weaponized OSINT scans for psychological and operational "open ports" in people.

As noted in recent research on spear-phishing [1] susceptibility, attackers effectively gather "as much private information as possible" from sources ranging from "online databases containing leaks to social media platforms". This process has been industrialized by AI, which allows for:

- **Automated Aggregation:** An AI does not view data in isolation. It ingests thousands of fragmented data points—a Linked-In job update, a Twitter comment about a delayed flight, an Instagram photo of a workspace—and unifies them.
- **Psychometric Inference:** As discussed in the introduction, this aggregated data allows for the inference of personality traits. The attacker no longer sees a "User"; they see a "High Extra-version target" who is susceptible to social validation.

In a **organizational perspective:** Criminals use public data to map an organization's internal hierarchy. By identifying "High Value Targets" (such as system administrators or finance directors) and their assistants, they can plan precision strikes that bypass general security filters.

2.1.2 Deanonymization and The Mosaic Effect

The most critical data protection risk in this domain is **Deanonymization** (also known as Re-identification). This occurs when an attacker takes a "masked" or anonymized dataset—such as a public health survey or a customer sentiment dataset—and cross-references it with public social media data to reveal the identities of the subjects.

This phenomenon is often described as the **"Mosaic Effect"**:

> *"Individually, a single data point is a harmless ceramic tile. However, when aggregated, these tiles form a complete mosaic that reveals the target's identity, location, and habits."*

This renders traditional "Anonymization" techniques insufficient. If an employee's public digital footprint is distinct enough, no amount of stripping names from a dataset will protect them from being re-identified by a sufficiently powerful AI model.

2.1.3 The Escalation: Why OSINT is a Critical Threat

It is crucial to understand that Weaponized OSINT is not merely a precursor to phishing; it is a distinct and often more harmful category of risk due to its **permanence**. While a successful phishing attack results in compromised credentials that can be remediated—passwords can be reset and systems patched—a **Deanonymized Profile is irreversible**. Once an individual's digital history is aggregated and linked to their real-world identity, their psychological triggers are exposed forever; an employee cannot "patch" their personality nor erase the historical footprint that the AI has already harvested.

Furthermore, Weaponized OSINT escalates the threat landscape from purely digital loss to **Physical and Reputational Harm**. Unlike phishing, OSINT enables **Doxxing** (publishing private addresses) and **Coercion** (blackmail via inferred private data). Ultimately, it serves as the **"Root Cause" enabler** for all advanced attacks; without the deep knowledge acquired through the "Mosaic Effect," attackers would be forced to revert to generic "Spray and Pray" tactics which are easily blocked. Therefore, mitigating OSINT is not just a data privacy exercise; it is a prerequisite for effective cybersecurity.

2.2 The Paradigm Shift: From Generic Phishing to AI-Precision

With the foundation of Weaponized OSINT laid, the attacker moves from "Passive Targeting" to "Active Execution." This marks the shift from generic nuisance to precision weapon. Usually a successful spear-phishing campaign follows a distinct five-phase life-cycle that moves from intelligence gathering to system compromise:

- **Phase 1: Preparation (The Intelligence Gather):** The attacker aggregates private information from online datasets, leaks and social media platforms. The goal is to curate specific details that make the email convincing enough to trigger a dangerous action, such as clicking a link or downloading an attachment.

- **Phase 2: Filtering Evasion:** To ensure the email reaches the target's inbox, the attacker must bypass technical spam filter and domain verification protocols. This requires the content to be "clean, clear, and balanced" to avoid automated detection.
- **Phase 3: The "Hook" (Opening the email):** Since the subject line is often the only visible element, it pays the essential role in convincing the target to open the message. Attackers may even send benign initial emails solely to establish trust before delivering the payload.
- **Phase 4: The payload (User Action):** This is the critical juncture where the attack shifts from passive to active. The target is manipulated into performing a dangerous activity, such as clicking a malicious link or running an executable file, often exploiting unpatched software vulnerabilities.
- **Phase 5: Compromise:** The final stage results in the theft of credentials or the compromise of the device and browser.

2.2.1 Traditional vs AI Enhancement

In this, AI primarily revolutionizes the first three stages by replacing manual effort with automated precision.

- **Phase 1:** Traditional method, an attacker manually searches online databases and social media to find "private information", AI tools automate this by scraping fragmented digital footprints to instantaneously build the "Psychometric Profile".
- **Phase 2:** Traditional method, attackers must format email to pass technical spam filters, AI can generate "clean, clear and balanced" content that mimics legitimate business or personal email.
- **Phase 3:** Traditional method, The attacker hopes the "Subject line" is relevant enough to get a click. AI leverages the target's personality traits (e.g., Neuroticism or Conscientiousness) to craft a subject line that triggers a specific psychological response, such as anxiety or curiosity.

2.3 The Other Side of the Coin: Cognitive Hacking and Consensus Manipulation

While the previous sections focused on the *extraction* of truth (OSINT) to target individuals, AI also enables the *manipulation* of truth to target groups. This phenomenon is known as **Cognitive Hacking**. In this context, public data is not just a resource to be harvested; it is a surface to be poisoned. Malicious actors utilize AI to inject synthetic data into the public discourse, manipulating the "census of reality" to stress-test or ideologically engineer specific demographics.

2.3.1 The Mechanism of Social Hacking

Just as AI can infer a single user's personality to phish them, it can infer a community's collective anxieties to destabilize them.

- **Consensus Manufacture:** AI agents can generate thousands of unique, human-sounding "public opinions" on social media. This creates a false consensus (Astroturfing) that tricks real users into believing a fringe viewpoint is the majority opinion.
- **Ideological Stress Testing:** By analyzing the "Agreeableness" and "Neuroticism" of a specific demographic , attackers can tailor disinformation campaigns designed specifically to trigger outrage or panic in that group.

This means we can no longer trust "Public Sentiment" as a raw metric. Decision-making based on unverified public data is now a vulnerability; we risk making strategic moves based on a mirage created by a "Social Hacking" campaign.

3 Ethical Concerns and Legal Blind Spots

While the technical mechanisms of these attacks are sophisticated, the true danger lies in the "Grey Zones" they exploit—areas where current legal frameworks and ethical guidelines have failed to keep pace with AI capabilities.

3.1 The "Public Data" Paradox (OSINT Blind Spots)

The Legal Blind Spot (GDPR Art. 9)[3]: The primary legal defense against Weaponized OSINT is the distinction between "Private" and "Public" data. However, this distinction is collapsing.

- **The Loophole:** Under GDPR Article 9(2)(e)[3], the prohibition on processing sensitive data (like political opinions or health data) does not apply if the data was "manifestly made public by the data subject."
- **The AI Exploitation:** Attackers—and even commercial data brokers—abuse this loophole. They argue that because an employee Tweeted about a "stressful week" (public), the AI's inference of "High Neuroticism" (sensitive psychological data) is fair game. This creates a massive blind spot where **Inferred Privacy** is unprotected.

The Ethical Concern (Contextual Integrity): This violates the principle of *Contextual Integrity*. Information shared in one context (a casual social post) is being weaponized in a completely different context (a cyber-attack). The ethical failure lies in treating "Accessibility" as "Consent."

3.2 The Asymmetry of Accountability (Phishing)

The Legal Blind Spot (Liability and Negligence): When an AI-Precision Phishing attack succeeds, who is liable?

- In an organizational perspective, Current legal frameworks often view the employee who clicked the link as "negligent."
- However, as the paper [1] demonstrate, specific personality traits (like Agreeableness) make susceptibility statistically predictable. If an AI uses "super-human" psychological profiling to trick an employee or individual, can the human reasonably be expected to resist?
- **The Gap:** The law currently lacks a "Defense of Superior Force" for social engineering. We treat a human falling for an AI script as a failure of training, rather than a failure of protection.

The Ethical Concern (Victim Blaming): Organizations often default to "Retraining" or punishing victims of phishing. Ethically, this is flawed in the AI era. If the attacker uses an automated tool to bypass human cognition (as proven by the "Non-Responder" profile in recent research, punishing the human is akin to punishing a user for a firewall failure.

3.3 The Truth Gap (Cognitive Hacking)

The Legal Blind Spot (The AI Act and Transparency): While the EU AI Act (Article 50)[4] mandates that AI-generated content be marked as such, this is technically unenforceable for text-based Cognitive Hacking.

- A "Social Hacking" bot farm can generate millions of unique comments to manipulate consensus.

- **The Blind Spot:** There is currently no legal mechanism to verify the "Humanity" of public discourse. We rely on platforms to police this, but they lack the incentive to delete engagement-driving content.

The Ethical Concern (Corporate Epistemic Duty): This creates an ethical crisis for decision-makers in AI. If we use public sentiment data to make business or policy decisions, we may be acting on "Poisoned" reality. The ethical question becomes: *Does an organization have a duty to verify that its data comes from humans before acting on it?*

4 Policy Recommendations: A Framework for Personal Cognitive Defense

As established in the previous sections, the "Human Firewall" is failing because AI has successfully weaponized our own psychology against us. Traditional organizational policies (like "change your password") are insufficient against these threats.

Therefore, this handbook proposes a shift from *System Security* to ***Personal Cognitive Defense***. The following guidelines are designed to empower the individual to protect their digital identity and psychological autonomy.

4.1 Defensive Compartmentalization (Countering Weaponized OSINT)

The Risk: As noted in the paper[1] study, attackers aggregate data from multiple sources—online databases, leaks, and social media—to build a target profile. An AI model requires a "rich" dataset to infer personality traits like Extraversion or Neuroticism.

Personal Policy Protocol:

- **Identity Segregation:** Do not allow your "Professional Persona" (LinkedIn/Work Email) and "Private Persona" (Instagram/X) to touch. Use different profile pictures, usernames, and email addresses for each. This creates a "Data Air-Gap" that prevents an AI from correlating your professional role with your personal psychological triggers.
- **The "Boring" Professional Rule:** Maintain a strictly factual professional footprint. By limiting emotional expression (rants, celebrations, personal opinions) on professional channels, you deprive the AI of the *Openness* and *Agreeableness* data points it needs to profile you.

4.2 The "Affective Gap" Protocol (Countering AI Phishing)

The Risk:Research confirms that individuals with high *Agreeableness* (desire to help) and high *Neuroticism* (anxiety) are statistically more susceptible to spear-phishing. AI attacks are designed specifically to trigger these emotional responses.

Personal Policy Protocol:

- **Emotion as an Indicator of Compromise (IoC):** We typically treat "bad grammar" as a sign of a scam. In the AI era, we must treat **"Emotional Urgency"** as a sign of a scam.
- **The Rule:** If a digital message triggers a sudden spike in emotion—whether it is fear (Neuroticism) or a strong desire to be helpful (Agreeableness)—you must effectively "pause" the interaction. Verify the sender through a secondary, non-digital channel (e.g., a phone call). Trust the voice, not the text.

4.3 3. Active Data Sovereignty (Countering Shadow Profiling)

The Risk: You are being profiled by third-party data brokers who feed this data to AI models, often without your explicit consent.

Personal Policy Protocol:

- **Exercise of Rights:** As a Cyber Law practitioner, you must proactively utilize **GDPR Article 15 (Right of Access)** and **Article 17 (Right to Erasure)**[3]. Regularly query major data aggregators to see what "Shadow Profile" exists on you and demand its deletion.
- **Pollution over Silence:** Where erasure is impossible, consider "Data Poisoning." Feeding conflicting or nonsensical data (e.g., browsing random products you don't need) into the public web can lower the confidence score of the AI trying to profile you, effectively "masking" your true psychological traits.

5 Conclusion

The convergence of AI, Data, and Cyber-security has fundamentally altered the threat landscape. We have moved from an era of "Hacking Systems" to an era of **"Hacking People."** As demonstrated in this handbook, the "harmless" public data we generate is no longer ephemeral; it is the raw material for **Weaponized OSINT** and **Psychometric Profiling**.AI tools have democratized the ability to launch precision social engineering attacks, scaling the capabilities of a state-sponsored actor to the level of a common criminal.

However, this does not require us to retreat from the digital world. Instead, it requires a new form of digital literacy—one that recognizes that **Privacy is no longer about secrecy, but about control.** By understanding how AI interprets our Openness, Agreeableness, and Neuroticism, we can construct a "Defensive Visibility" that allows us to participate in the digital economy without becoming its casualty.The policies outlined here—Identity Segregation, Emotional Verification, and Data Sovereignty—are not just security measures; they are an assertion of our right to remain **un-predicted** and **un-profiled** in an age of automated surveillance. This policy handbook has been inspired from the following papers [1] and articles [2]

6 AI Use Acknowledgment

I acknowledge the use of **Google Gemini** (Large Language Model), published by Google (URL: https://gemini.google.com/), to assist in the preparation of this assessment.

Specifically, I used the tool for the following purposes:

- **Ideation and Structuring:** To brainstorm the initial concept of "Weaponized OSINT" and develop a logical six-page structure that integrates Cybersecurity, AI, and Data Protection.
- **Polishing:** To contextualized certain text to fit the specific constraints of the assignment and remove mistakes.
- **Summarization:** To summarize key findings from the research paper *"Spear-Phishing Susceptibility Stemming From Personality Traits"* (Eftimie et al., 2022) to ensure accurate citation of the attack lifecycle and statistical results.
- **Technical Formatting:** To generate the LaTeX code used for the document's layout and bibliography.

References

[1] Sergiu Eftimie, Radu Moinescu, and Ciprian Răcuciu. "Spear-Phishing Susceptibility Stemming From Personality Traits". In: *IEEE Access* 10 (2022), pp. 73548–73561. DOI: 10.1109/ACCESS.2022.3190009.

[2] Elena Martynova. *The Dark Side of OSINT: How Extremists Exploit Open-Source Intelligence*. Accessed: 2025-12-11. Counter Extremism Project. Nov. 2025. URL: https://www.counterextremism.com/blog/dark-side-osint-how-extremists-exploit-open-source-intelligence.

[3] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. EN. Official Journal of the European Union, L 119, pp. 1–88. Current consolidated version: 04/05/2016. ELI: http://data.europa.eu/eli/reg/2016/679/oj. May 2016.

[4] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. EN. Official Journal of the European Union, L 2024/1689. In force. ELI: `http://data.europa.eu/eli/reg/2024/1689/oj`. July 2024.