ON2IT AND AMS-IX PRESENT: THREAT TALKS

Threat Talks
AI-Powered Cyber Threats

The new frontier of cybercrime and defense

In the age of AI, cyber threats have evolved from simple malware to sophisticated, self-learning attacks. AI-powered cyberattacks can adapt in real time, bypassing traditional defenses and targeting vulnerabilities with unprecedented precision. These threats are not just theoretical; they're already here, reshaping the landscape of cybersecurity.

As AI continues to advance, its potential for misuse grows. Cybercriminals are leveraging AI to develop more convincing phishing schemes, automate vulnerability discovery, and even create malware that adapts in real time to bypass security measures. This evolution demands a proactive and adaptive cybersecurity strategy to safeguard against increasingly sophisticated threats.

In this Threat Talks infographic we will discuss the following threats:

- AI, Play It Safe
- PromptLock
- Deepfakes

80% of ransomware attacks now utilize AI
Source: MITSloan

87% of organizations have encountered AI-driven cyberattacks
Source: SoSafe

54% AI-generated phishing mails had a 54% click-through rate
Source: MalwareBytes

85% of cybersecurity professionals attribute the increase in cyberattacks to generative AI
Source: CFO

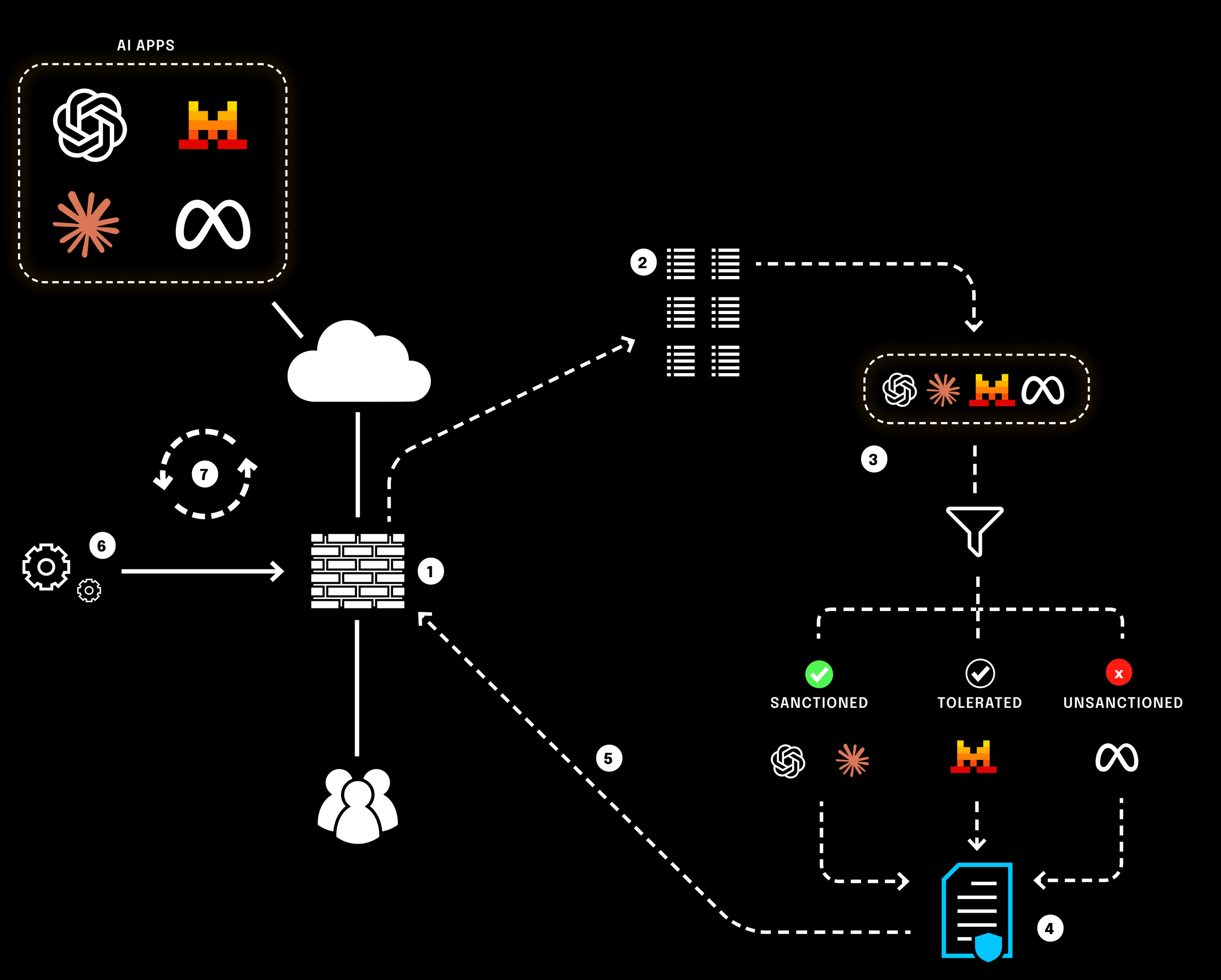
AI, Play It Safe

How to secure your AI usage

AI has moved from novelty to necessity, with tools like ChatGPT now embedded in daily workflows. Used well, they can significantly boost productivity, speed up coding, improve translations and much more.

But the same tools can also expose sensitive information, create compliance risks and challenge traditional security approaches.

Here's how to play it safe with AI.



Monitor all AI traffic

1. Track all network traffic to and from AI applications. Use **firewalls, proxies, and SASE solutions** to gain full visibility into what tools are being accessed.

Generate an AI usage report

2. Create reports that show **which AI applications are in use, who is using them, and how often**. This provides the foundation for governance and risk management.

Classify AI applications

3. Sort AI tools into clear categories:
   - **Sanctioned** - officially approved and trusted applications.
   - **Tolerated** - not formally approved, but not considered a major risk.
   - **Unsanctioned** - explicitly forbidden due to security or compliance concerns.

   This distinction helps you separate acceptable use from potential threats.

Define a policy (Kipling/5W1H method)

4. Translate classifications into clear policies. Define:
   - **Who** may use AI tools
   - **What** applications are allowed
   - **When** and **where** they can be used
   - **Why** they're approved
   - **How** usage should be controlled

   This ensures AI adoption aligns with business and compliance needs.

Enforce the polcy in security controls

5. Implement the policy in your **firewall, proxy, or SASE solution**. Automated enforcement reduces manual oversight and strengthens consistency across the organization.

Activate advanced protection

6. Go beyond access control by enabling **prompt detection, data loss prevention, and misuse monitoring**. Note: these protections often require **SSL/TLS decryption** for full effectiveness.

Continuously adapt to the AI landscape

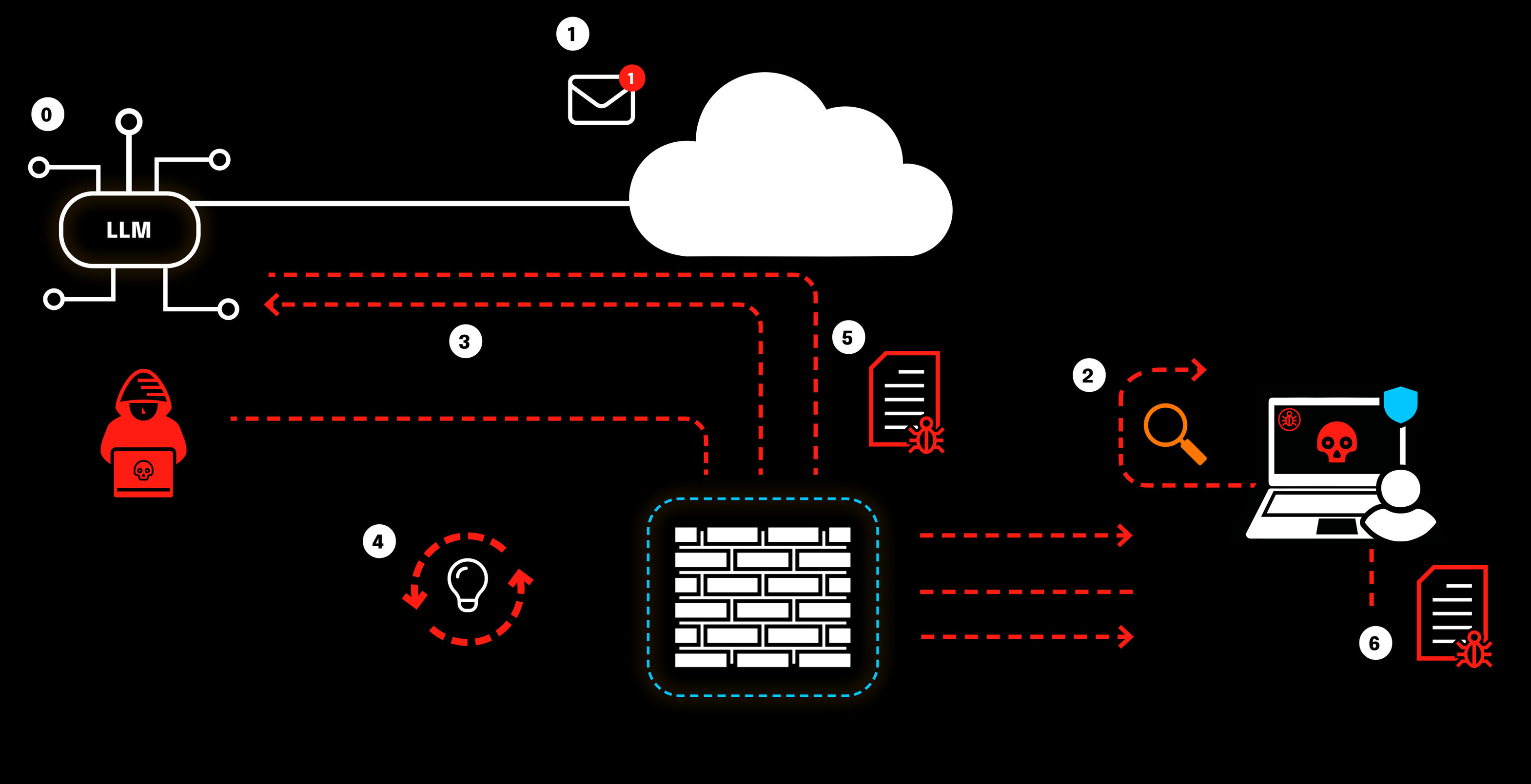
7. AI evolves quickly. Regularly **review reports, re-evaluate classifications, update policies, and fine-tune controls**. Treat this as a continuous cycle, not a one-time project.

PromptLock

AI-powered malware (Proof of Concept)

As AI is already used by all types of threat actors to varying degrees, it's also set to help power an increase in the volume and impact of ransomware attacks.

ESET researchers have discovered what is the first known AI-powered ransomware. The malware, which has been named PromptLock, has the ability to exfiltrate, encrypt and possibly even destroy data, though this last functionality appears not to have been implemented in the malware yet.



Initial Setup

0. Attacker prepares a local LLM (in this PoC: Ollama with gpt-oss:20b) that will later be used by the malware.

Important note

Some defensive measures are **specific to this Proof of Concept** (e.g., blocking LUA execution).

The broader lesson is **the potential of AI-powered malware**:

- Future variants could generate code in common languages like **Python** or **PowerShell**, adapting dynamically to the target environment.

Initial infection

1. Attacker tricks the user into downloading and executing a piece of software (malware). While harmless on its own, it serves as the entry point for the attack.

Methods include:

- Malicious email attachments
- Fake websites
- Trojanized legitimate applications

M. Defensive measures:

- Block downloads of executable files (most users do not need them).
- Apply strict spam and phishing controls (most malware arrives via email).

Reconnaissance

2. The malware collects information about the system and its environment (network, other machines, etc.).

   Even though this is a limited PoC, the malware could be extended to set its own goal: exfiltrate data, encrypt files (ransomware), or destroy information.

M. Defensive measures:

- Deploy EDR/XDR solutions to detect abnormal behavior and suspicious system activity.

Contact with LLM

3. The malware connects to the LLM and sends tailored prompts based on the collected reconnaissance data.

M. Defensive measures:

- Enforce strict controls on outbound network connections, especially on servers.
- Detect usage of suspicious API calls like in this case Ollama.

Payload generation

4. The LLM generates a specific payload (in this PoC: LUA scripts).

Payload delivery

5. The generated payload is transmitted back to the victim system.

Execution of final attack

6. The payload is executed, carrying out the attacker's final objective.

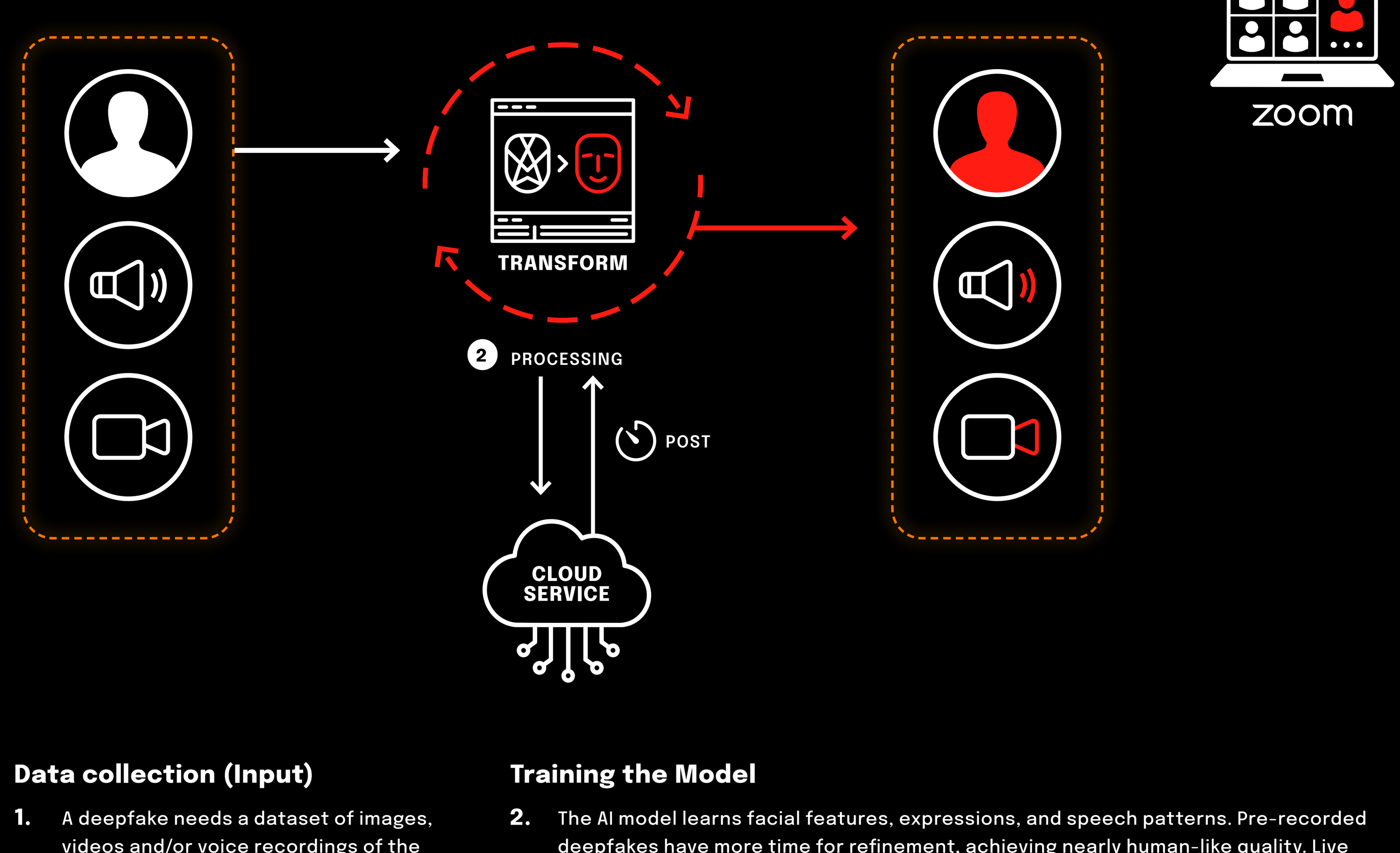
M. Defensive measures:

- EDR/XDR can detect anomalous activity (e.g., execution of LUA scripts).
- Remove or block interpreters (e.g., LUA). If the interpreter isn't available, the script cannot be executed.

Deepfakes

The Rising Threat of AI-Generated Deception

Using machine learning techniques like deep neural networks, hackers and cybercriminals can convincingly alter faces, voices, and actions, making it difficult to distinguish real from fake. While deepfake technology has legitimate uses in entertainment and research, it also poses serious risks: fueling misinformation, fraud, and identity theft.



Data collection (Input)

1. A deepfake needs a dataset of images, videos and/or voice recordings of the intended target.

Training the Model

2. The AI model learns facial features, expressions, and speech patterns. Pre-recorded deepfakes have more time for refinement, achieving nearly human-like quality. Live deepfakes, however, must adjust to real-time inputs, requiring more resources and often resulting in lower quality. The trained model applies the target's face and voice to the source video, with live deepfakes syncing face and voice in real-time.

Delivery (Output)

3. The final deepfake, whether live or pre-recorded, is delivered to the victim or shared on social media, depending on the attacker's goal.

Live output

4. For live deepfakes, such as during Zoom calls, you can implement a security question to verify the identity of the person joining your meeting. This can be done by agreeing on a code beforehand, in person, providing an extra layer of security.

threat-talks.com

ON2IT ZERO TRUST INNOVATORS | amsix

CONTACT US