
ENDGAME
Can AI Detect AI?

Sami Laiho

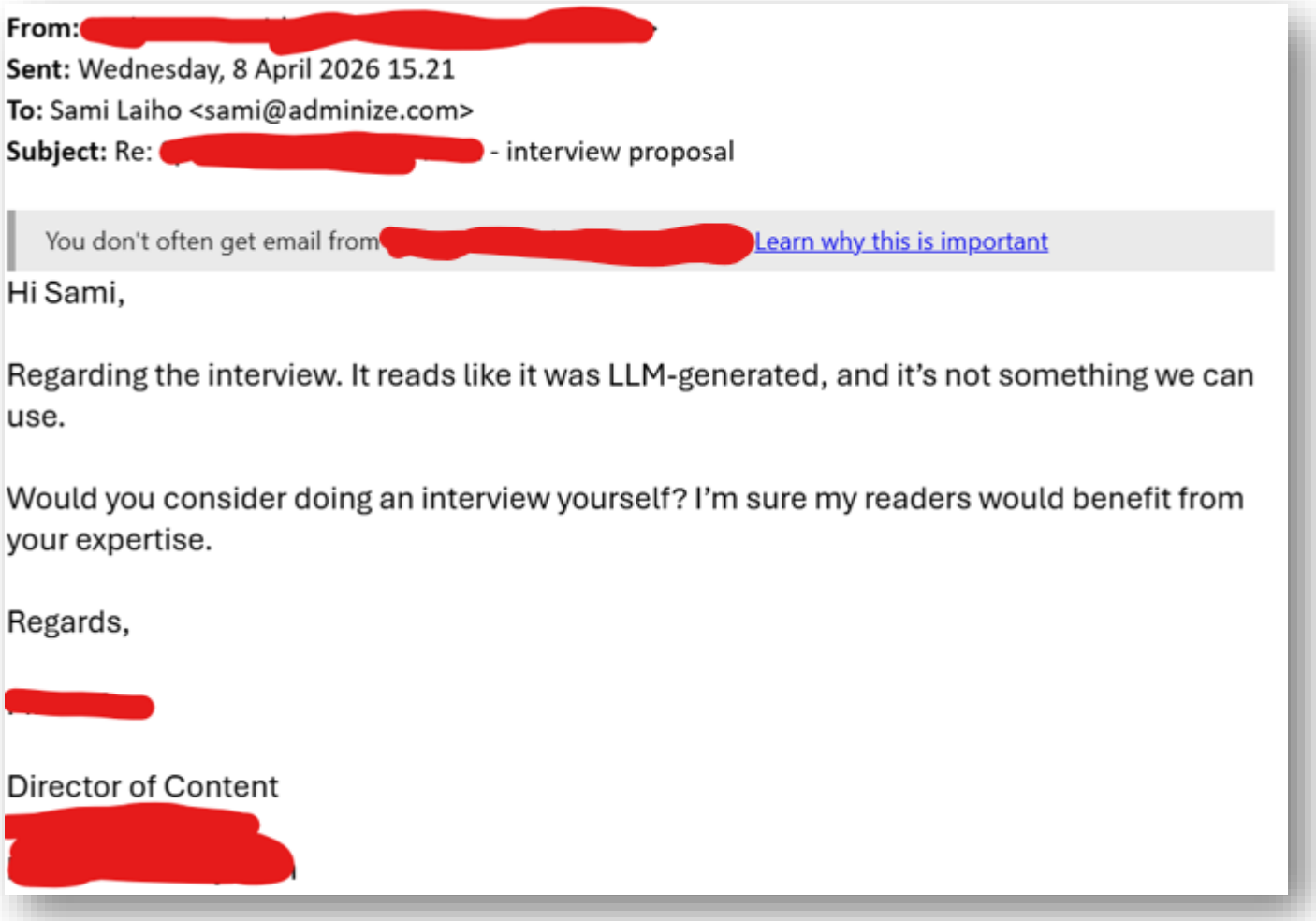
Chief Research Officer / Microsoft MVP, Adminize Oy

- IT Admin since 1995
- Microsoft MVP since 2011
- Erikoistunut:
  - Yritysten tietoturva-auditoinnit
  - Koulutus: IT, loppukäyttäjät ja johto
  - Tietoturva-arkkitehtuurit
- Julkiset esiintymiset, esim:
  - Best Session at Advanced Threat Summit 2020
  - Best Speaker at NIC, Oslo 2016, 2017, 2019, 2020, 2022, 2023, 2024 and **2025 (keynote)**
  - **Ignite 2018 – Session #1 and #2 (out of 1708)**
  - TechEd Europe and North America 2014 - Best session, Best speaker
  - TechEd Australia 2013 - Best session, Best speaker

Text with AI

Creation, correction, translation

VASTAAMO-case



From: [redacted]
Sent: Wednesday, 8 April 2026 15.21
To: Sami Laiho <sami@adminize.com>
Subject: Re: [redacted] - interview proposal

> You don't often get email from [redacted] Learn why this is important

Hi Sami,

Regarding the interview. It reads like it was LLM-generated, and it's not something we can use.

Would you consider doing an interview yourself? I'm sure my readers would benefit from your expertise.

Regards,

[redacted]

Director of Content

[redacted]

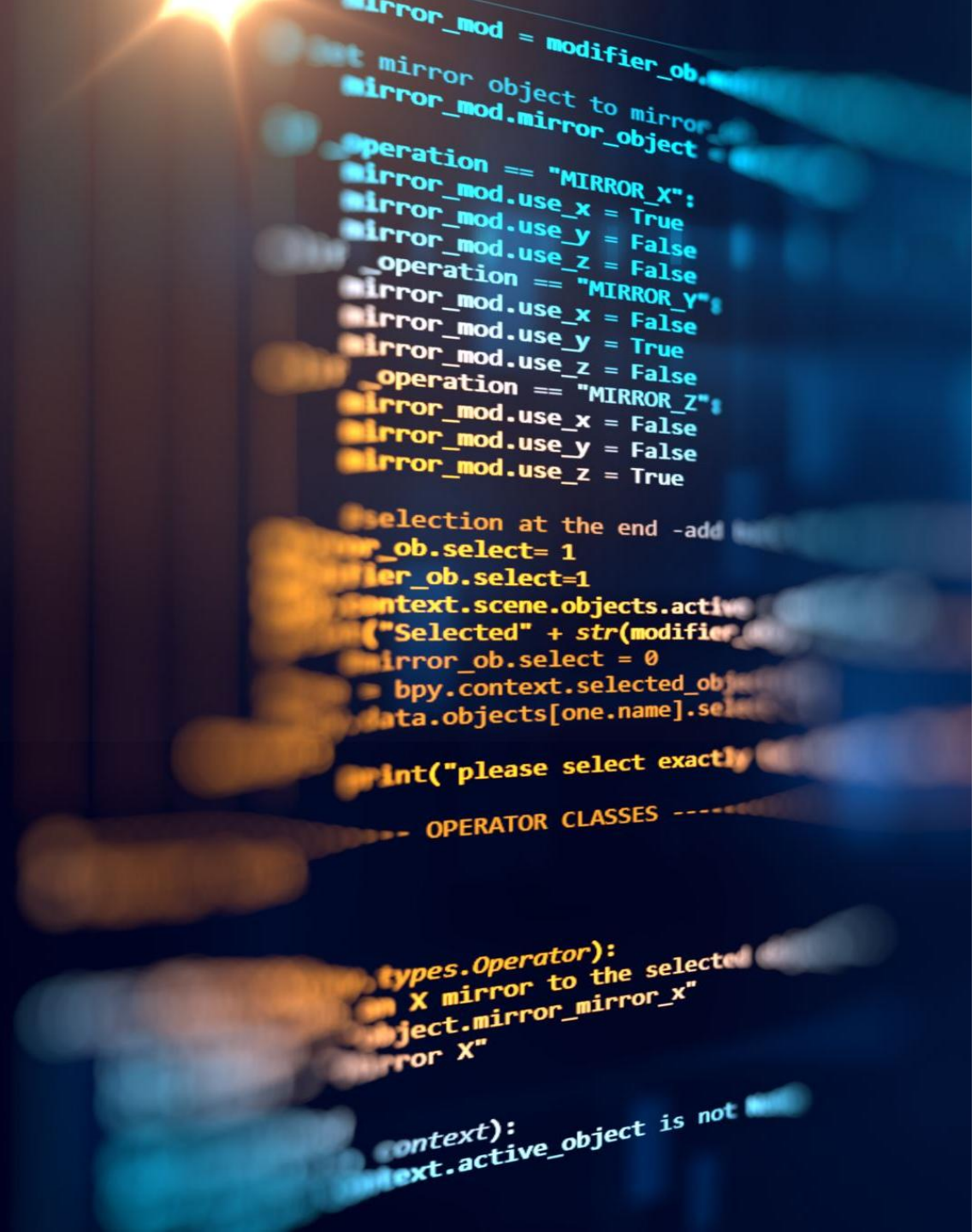
Viranomaisposti muuttuu pääosin sähköiseksi: tietoa ja tukea tarjolla kansalaisille

Julkaistu 17.3.2026

Eduskunta hyväksyi 11.3.2026 lakimuutokset, joiden mukaan digitaalisesti asioivat saavat viranomaisten lähettämän postin jatkossa ensisijaisesti digitaalisesti. Paperinen viranomaisposti säilyy vaihtoehtona kaikille. Niiden, jotka eivät asioi digitaalisesti, ei tarvitse tehdä mitään. Muutos edellyttää vielä lakien vahvistamista ja tulisi voimaan 14.4.2026.

How does AI-detection work

- Modern AI detection tools don’t “Recognize AI” in a binary sense - they estimate the *likelihood* that text (or media) was generated by a machine based on statistical and structural signals.



Perplexity Analysis (Core Signal)

- Most detectors use a concept from NLP called *perplexity*.
- A language model assigns probabilities to sequences of words.
- **Human writing** → more irregular, less predictable
- **AI-generated text** → smoother, more predictable (lower perplexity)
- **Detection logic:**
  - Feed the text into a reference model
  - Measure how “surprising” the word choices are
    - Low perplexity → likely AI
    - High perplexity → likely human
- **Limitation:**
  Good human writers (or edited AI text) can also produce low perplexity → false positives.

Burstiness – variation in writing

- AI tends to produce:
  - Uniform sentence length
  - Consistent structure
  - Even tone
- Humans naturally show:
  - Short + long sentence mix
  - Irregular phrasing
  - Occasional “messiness”
- **Detection tools compute:**
  - Variance in sentence length
  - Entropy across phrases
  - Distribution of rare vs common words
    - Low burstiness = suspicious

Token Probability Patterns

- Detectors analyze **how likely each word is given the previous context**.
- AI models (like GPT variants) often:
  - Choose *high-probability tokens*
  - Avoid unusual phrasing unless prompted
- Detection models look for:
  - Overuse of statistically “safe” words
  - Lack of rare or unexpected token transitions

Classifier Models (Supervised Learning)

- Many tools train a **binary classifier**:
  - Input: text
  - Output: probability of AI vs human
- Training data:
  - Large corpora of human writing
  - Large corpora of AI-generated text
- These models learn subtle patterns such as:
  - Syntax regularity
  - Phrase repetition structures
  - Semantic predictability
- Examples include detectors inspired by work from:
  - OpenAI
  - Turnitin

Watermarking (Emerging Technique)

- Some AI systems embed hidden signals during generation.
- How it works:
  - The model slightly biases word choices using a secret key
  - This creates a detectable statistical “fingerprint”
- Detection:
  - Check if the text follows that hidden pattern
- Stronger signal than perplexity—but:
  - Only works if the generator used watermarking
  - Easy to break with paraphrasing

Stylometry (Authorship Analysis)

- Borrowed from forensic linguistics.
- Analyzes:
  - Writing style consistency
  - Grammar habits
  - Punctuation patterns
  - Function word usage (e.g., “and”, “the”, “but”)
- Used more in:
  - Academic integrity
  - Insider threat detection

Modern Detection

- Combines the previous
- Result: Probability

- Why is it unreliable
  - No absolute truth
  - Easy to bypass
  - False positives

Deepfakes are a real threat

“Cybersecurity firm DeepStrike estimated online deepfakes skyrocketed from roughly 500,000 in 2023 to about 8 million in 2025, with annual growth nearing 1,500%.”

I want myself standing on a pyramid. I should talk about myself.

I want myself standing on a pyramid. I should say "My name is Sami Laiho and I can't wait to get to Jyväskylä".

Me standing on the moon saying "My name is Sami Laiho and I can't wait to get to Jyväskylä"

Sami_smiling.jpg

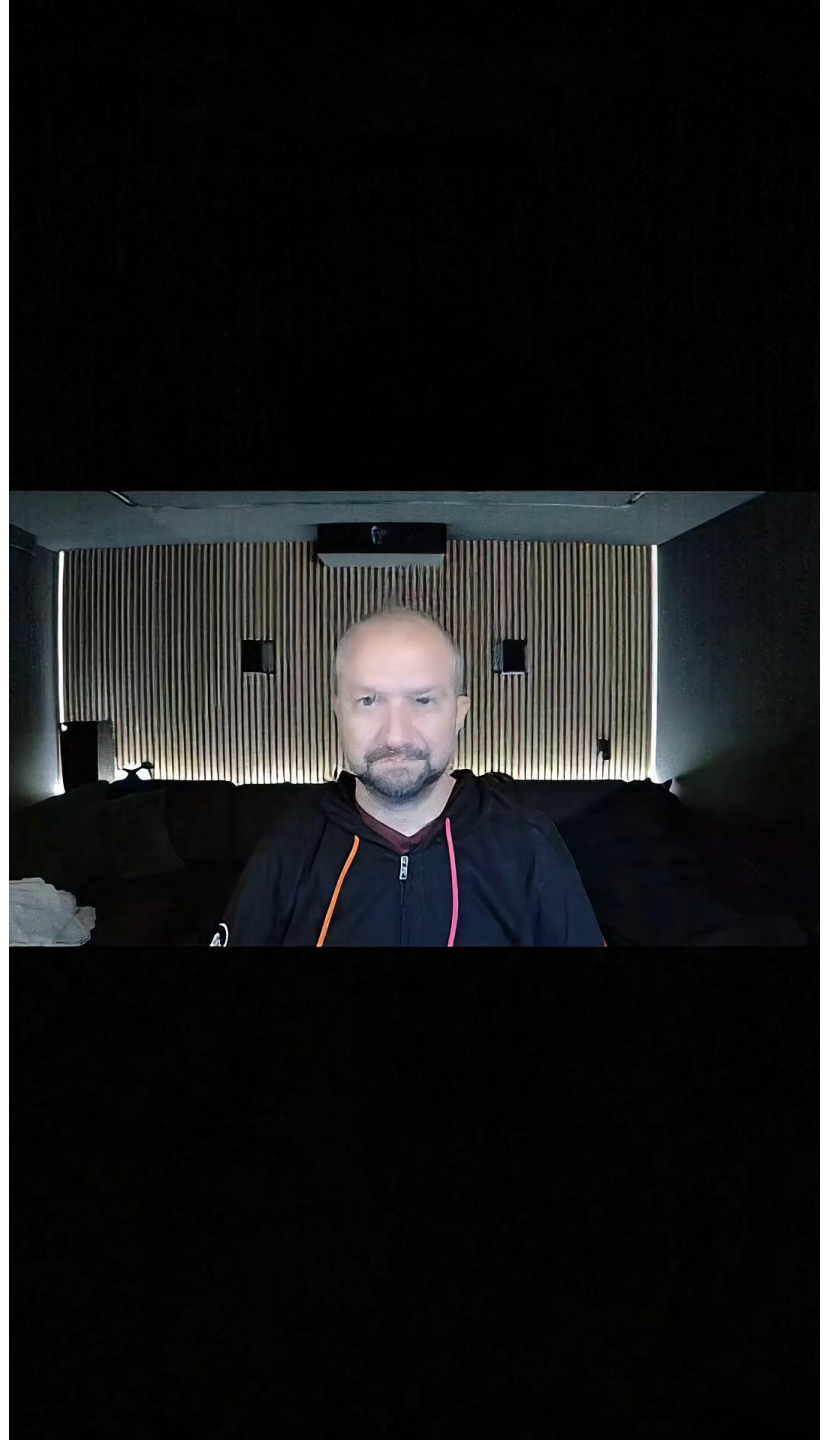
From a still image…

Disinformation

Geopolitics

Causing fear and terror

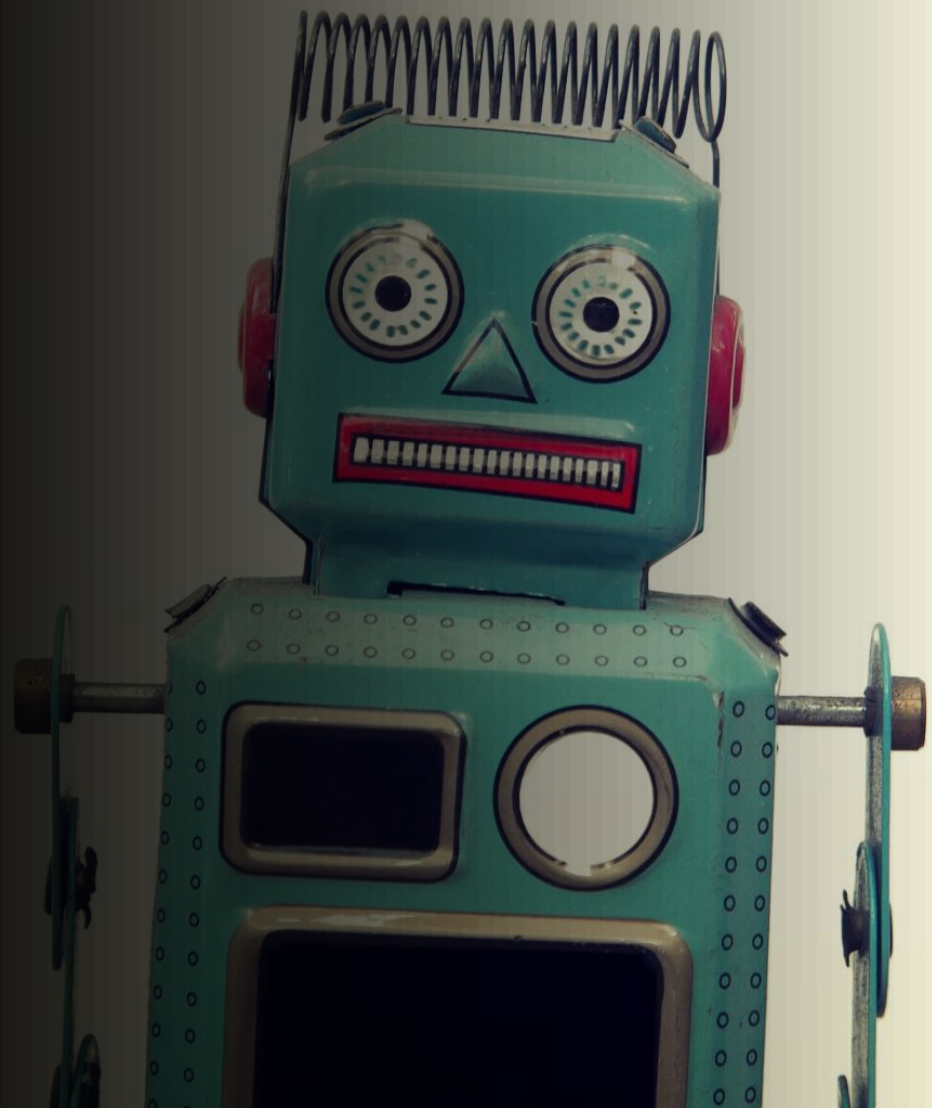

BREAKING NEWS:
FLAND INVADES FLORID



Florida Man

Recruting

North Korea



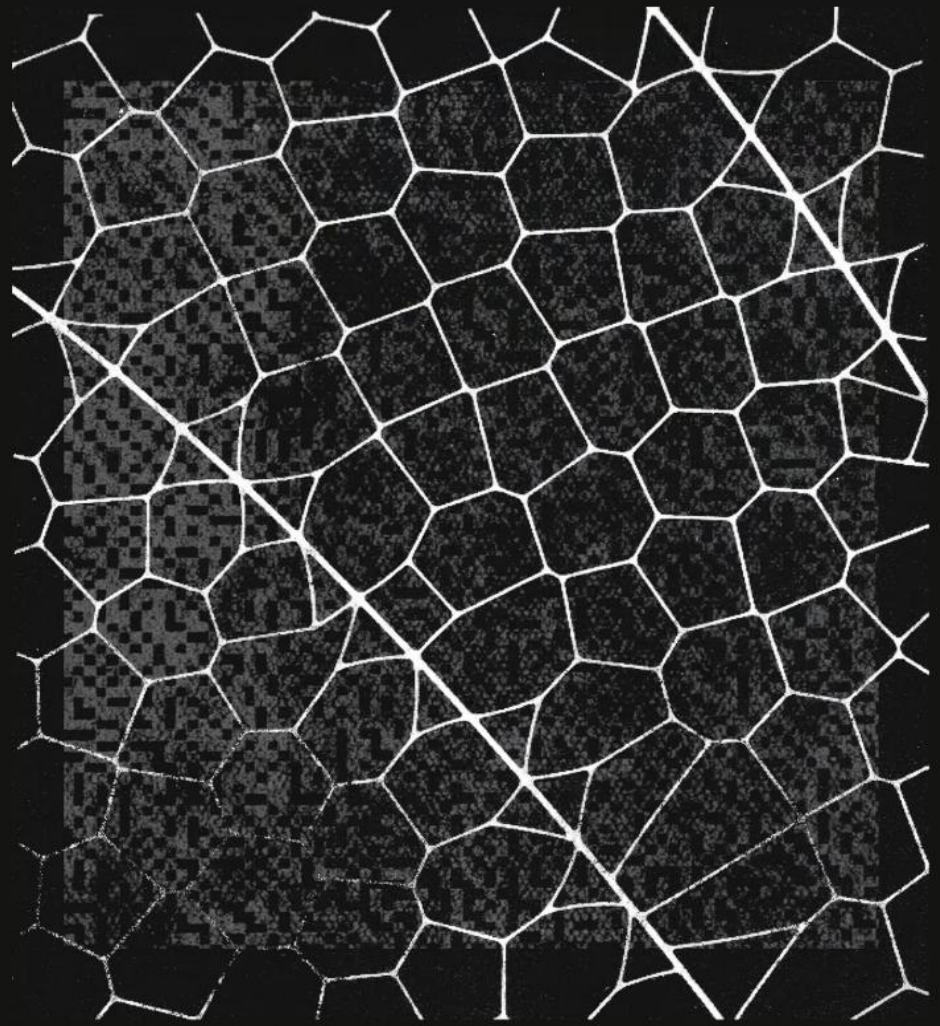
Mitigation

- MS identity check is around 95% (today)
- Turn your head, show your watch/phone, cover your face… (today)

Project Glasswing

Securing critical software for the AI era

Continue reading




BugBounty in the Future
Like Amazon with published books…

AI experiences no Fatigue

Old and new bugs are found - partly because the attack surface is older than most of the people hired to defend it.

Already an insane amount of work

Unfair Game

- **The attacker's velocity of adoption is gated only by compute and creativity.**
- **The defender's velocity of adoption is gated by enterprise policy, change control, production stability, compliance review, procurement, vendor consolidation, board appetite, insurance posture, and twelve people in four time zones agreeing on one Jira ticket**
- The attacker's cost of failure is low and getting lower.
- The defender's cost of failure is structurally civilization-scale.

- *Source: https://cje.io/2026/04/08/offense-scales-with-compute-defense-scales-with-committees/*

Endgame – Can AI detect AI

- Attacker has the advantage

- AI detection is a **probabilistic classifier under adversarial pressure**
- Detection without control of the generation process is inherently weak
- The only robust approach:
  - **Control the identity and origin (who/what generated it)**
  - Not just analyze the output

Thank you!

Contact on LinkedIn!

Follow via X/BSky @samilaiho / @samilaiho.com

Email: sami@adminize.com