AI and Cybersecurity –
Attacker's Best Friend or Defender's Last Hope?

Panel #1

20 April 2026

Speakers

Moderator

Prof. Dr. Andreas Aßmuth, Kiel University of Applied Sciences, Germany

Panelists

Prof. Dr. Patrick Levi, OTH Amberg-Weiden, Germany

Dr. Mika Helsingius, Finnish Defence Research Agency, Finland

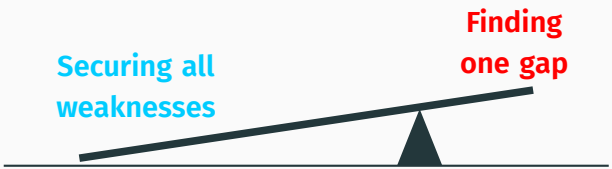
Lena Sinterhauf, NetUSE AG, Germany

Dr. Steve Chan, Decision Engineering Analysis Laboratory, USA

Prof. Dr. Ian Ferguson, Abertay University, UK

The Fundamental Asymmetry of Cybersecurity

The attacker needs to succeed only once.
The defender needs to succeed every time.



AI and Cybersecurity – A Quick Recap

Agentic AI

- AI systems acting autonomously
- Multi-step tasks without human intervention
- Near future – but closer than we think

CVEs & Exploit Code

- AI knows CVEs from training data
- AI can generate exploit code
- Lowers the barrier to entry significantly

Both can be used by attackers – and by defenders.

AI Systems Find 0-Day Exploits


Cyber Press

Zero-Day Alert: Claude AI Finds Critical RCE Bugs in Vim and Emacs

ZeroDayBench: Evaluating LLM Agents on Unseen Zero-Day Vulnerabilities for Cyberdefense

Nancy Lau[1], **Louis Sloot**[2], **Jyoutir Raj**[3], **Giuseppe Marco Boscardin**[3], **Evan Harris**[3]
[1]UC Santa Cruz, [2]Carnegie Mellon University, [3]Independent
najilau@ucsc.edu

Dylan Bowman, Mario Brajkovski, Jaideep Chawla
HUD
dylan@hud.ai

Dan Zhao
New York University
dz1158@nyu.edu

arXiv > cs > arXiv:2603.02297


The Hacker News

Subscribe – Get Latest News

Anthropic's Claude Mythos Finds Thousands of Zero-Day Flaws Across Major Systems


Cyber Security | Cyber Security News | Vulnerability News

Claude AI Discovers Zero-Day RCE Vulnerabilities in Vim and Emacs

By Guru Baran · March 31, 2026

Project Glasswing

Claude Mythos Preview

- Finds zero-day vulnerabilities autonomously
- 27-year-old vulnerability in OpenBSD
- 16-year-old vulnerability in FFmpeg
  (missed by automated tests 5 million times)

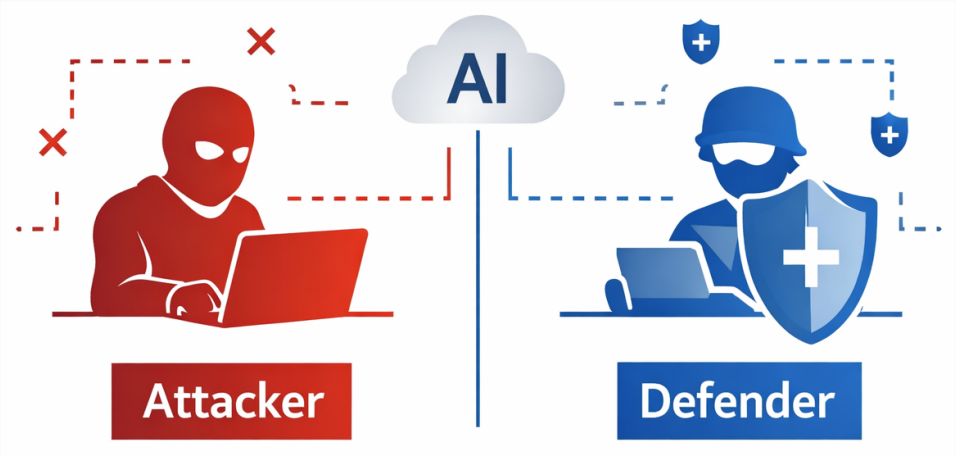
The Initiative

- Partners: AWS, Apple, Cisco, CrowdStrike, Google, JPMorganChase, Microsoft, NVIDIA, Palo Alto Networks
- $100M in usage credits for defensive security work
- Access for open-source maintainers

Securing critical software for the AI era

The same capabilities that make it dangerous make it invaluable for defense.

AI Can be Used by Attackers AND Defenders



Patrick's Position (1)

Secure coding, vulnerability testing

- No real advantage
- Who uses AI faster and smarter: attacker or defender?

AI as attacker's best friend

- New routes to success: adversarial attacks against AI
- Advanced AI ☞ more adversarial attacks
- LLMs: Jailbreaks, training data poisoning
- Chatbots: Jailbreaks, training data poisoning + context poisoning
- Undetected by IT Security ☞ new solutions required

Prof. Dr.
Patrick Levi
OTH-AW, Germany

Patrick's Position (2)

AI as attack detector (e.g. network intrusion detection)

- Advantage for defender
- However, will be combined with adversarial attack
- Or, clever zero-day attack will go just undetected

Compare AI to the Internet

- Defender: Faster distribution of anti virus software
- Attacker: Much faster spreading of viruses, new attack possibilities

Defenders "just" need to adapt and go new ways, like the attacker does

- "Just" faster and smarter, as always...

Prof. Dr.
Patrick Levi
OTH-AW, Germany

Mika's Position (1)

The real challenge is that the whole field is so messy, or rather, everything is so complicated and evolving so fast.

- I just try to understand the general view on what is going on.

Few observations on AI risk management:

- At least three different approaches: EU AI Act, NIST AI RMF and ISO 42001.
- EU regulation is enforceable, NIST is voluntary and ISO is standard.
- EU AI act is connected to many other directives etc., the whole combination is a bit hard for small and medium size companies.

Dr.
Mika Helsingius
FDRA, Finland

Mika's Position (2)

Few observations on AI risk management (contd.)

- NIST is easier, RMF evolves and it does not demand something too hard. "Artificial Intelligence Risk Management Framework" gives the basics, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile" adds the additional requirements for LLMs and Gen AI.

Dr.
Mika Helsingius
FDRA, Finland

Mika's Position (3)

ISO family is hierarchical

- "ISO 31000 Risk management – Guidelines". General risk management, can be used for any kind of business.
- "ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection - Information security management systems – Requirements". As the name says it, information security and cybersecurity are included.
- "ISO/IEC 42001 Information technology - Artificial intelligence - Management system" touches AI directly. Certification is available.
- "ISO 22989 Information technology - Artificial intelligence - Artificial intelligence concepts and terminology". More formal, useful for contracts and detailed documentation.
- "ISO 23894 Information technology - Artificial intelligence - Guidance on risk management". Instructions on how to use ISO 31000 together with AI, (ISO 42001 deals more directly with AI).

Dr.
Mika Helsingius
FDRA, Finland

Mika's Position (4)

Everything is fine?

- Perhaps for large corporations and for really long and large projects...

Small companies, case drone ecosystem

- Large ecosystem made up from many small companies
- No legal expertise or money for producing truck loads of paper (we really speak about tens of cubic meters of paper, that nobody is ever going to read...).
- Military drones in Ukraine. Development cycle from new requirements to updated products is measured in few weeks.
- Any kind of standards or formal methods are too slow, small companies will never be able to fulfill strict requirements (and the customer can't wait either).

Dr.
Mika Helsingius
FDRA, Finland

Mika's Position (5)

True for large companies also

- Shield AI V-Bat and X-Bat drones. Rheinmetall CEO Papperger might disagree...

How long will people write code, or when AI tools start to use some new intermediate "AI language"?

- Humans can not really comprehend the loop between attacker and defender AIs.
- What is the situation in the next 6 months or 1 year ??? Sam Altman speaks about post transformer systems etc.
- Difficult to plan. It is difficult to guess, what the world looks like in the end of 2026.
- Perhaps both sides will use AI and we just have to pray the best...

Dr.
Mika Helsingius
FDRA, Finland

Lena's Position (1)

AI in Cybersecurity: Superpower or Superproblem?

AI is not (yet) intelligent – but it is extremely efficient.

It reinforces everything that is already there: good processes become better. Bad processes become more dangerous.

Lena Sinterhauf, M.Sc.
NetUSE, Germany

Lena's Position (2)

AI as the Attacker's Best Friend

- Perfect phishing emails in every language – mass-produced yet highly personalised, without errors.
- Automated malware and exploit code at the push of a button – fast code, but only as effective as the expertise behind it.
- Scaling attack chains – attack campaigns that once took weeks can now be carried out with a single click.

Lena Sinterhauf, M.Sc.
NetUSE, Germany

☞ **AI lowers the barriers to entry – but true zero-days continue to emerge through expertise.**

Lena's Position (3)

AI as the Defender's Last Hope

- Automation of vulnerability detection and patching processes - particularly relevant since critical CVEs have remained unpatched for a median of over 4.5 years for the past 16 years.
- Continuous monitoring instead of selective scans
- AI-supported training of employees – realistic, adaptive training
- Risk-based prioritisation (EPSS, behavioural models) – not "everything is critical".
- Early warning systems through anomaly and behaviour analytics.

Lena Sinterhauf, M.Sc.
NetUSE, Germany

☞ **AI can accelerate analysis and reaction – but only if people orchestrate it sensibly.**

Lena's Position (4)

AI is not a Hero. It's an Amplifier.

Thesis: 'AI will not save cybersecurity, but those who do not use AI will definitely be at a disadvantage.'

Lena Sinterhauf, M.Sc.
NetUSE, Germany

Lena's Position (5)

Why AI is not a replacement for experts

- Models can hallucinate, do not understand context and can be manipulated (Prompt Injection, Data Poisoning).
- Automated decisions without human control ☞ new risk factor instead of providing a solution.
- Without professional validation, incorrect prioritisation and security decisions arise - more dangerous than slow human decisions.
- AI remains blind to organisational realities: lack of asset overview, chaotic processes, legacy systems.

Lena Sinterhauf, M.Sc.
NetUSE, Germany

Lena's Position (6)

What we really need

- Human–AI operations: AI as a co-pilot, not an autopilot
- Secure training data, clear governance and robust processes
- Transparent assessment of models and risks (AI Security Posture Management)
- Experts who evaluate, prioritise and control AI, rather than replace it

Lena Sinterhauf, M.Sc.
NetUSE, Germany

Lena's Position (7)

Final statement

AI should not be seen as a saviour, but rather as a tool. Tools require people to use them effectively.

Cybersecurity is still very much a human endeavour – AI is merely a reinforcement of our actions, not a replacement.

Lena Sinterhauf, M.Sc.
NetUSE, Germany

Steve's Position (1)

The Cycles of Adaptation Race

AI Systems (AIS) are endeavoring to implement emulated metacognitive mechanisms

- This Metacognitive Scaffolding (MS) might include chain-of-thought reasoning, self-consistency checks, confidence estimation, multi-agent Quality Assurance/Quality Control (QA/QC), etc.
- Challenges include: reliability (e.g., errors may reinforce themselves), cost/latency (e.g., the reflection loops involve multiple passes and require more computational resources), evaluation metrics are still being developed, etc.
- Yet, the trend is heading in this direction of multi-step, self-monitoring multi-agent AIS.

Dr. Steve Chan
Decision Engineering Analysis Laboratory, USA

Steve's Position (2)

There are the expected attack vectors...

- Adversarial Input Perturbation (AIP) attacks – introduces a subtle change to a valid input that affects the AIS Decision-Making (DM)
- Prompt Injection Attacks (PIAs) – introduces manipulative instructions to alter the AIS DM and behavior.
- Denial of Service (DoS) Attacks – moves the AIS toward a state of being unavailable/unusable; for AIS equipped with MS, the MS-targeted DoS or Distributed DoS (DDoS) is sometimes referred to as Metacognitive DoS (MDoS).

Dr. Steve Chan
Decision Engineering Analysis Laboratory, USA

Steve's Position (3)

There are the more potent attack vectors...

- Stacking Attack – as an example, AIP + PIA + MDoS
  These types of stacking attacks are particularly potent because they tend to accentuate the deficiencies across several layers of an AIS, in an amplification fashion, thereby transforming more minor manageable issues into catastrophic system-wide failures. In fact, defense-in-depth paradigms can be turned to the attacker's advantage as various defensive mechanisms may interfere with each other. For example, safety checks can increase computational requirements, and increased computational needs can be construed as amplifying MDoS.

Dr. Steve Chan
Decision Engineering Analysis Laboratory, USA

Steve's Position (4)

Some attack types leading up to MDoS and a Stacking Attack

- Adversarial Input Perturbation (AIP) – targets perception
- Feedback Injection Attack (FIA) – targets learning/reasoning
- Confidence Manipulation Attack (CMA) – targets DM (e.g., falsely confident/underconfident, resulting in imprudent decisions/decision paralysis, respectively)
- Recursive Reflection Attack (RRA) – targets compute (e.g., recursive reasoning, re-evaluation loops segueing to "analysis paralysis")
- Control Channel Flooding (CCF) – targets agents, multi-step pipelines, internal controls (e.g., excessive safety triggers, excessive policy checks)
- Cognitive Saturation Attack (CSA) – targets reasoning (e.g., conflicting and/or ambiguous inputs, edge cases that require inordinate amounts of reasoning)

Dr. Steve Chan
Decision Engineering Analysis Laboratory, USA

Steve's Position (5)

Exemplar mapping against the 7 Stages of the Cyber Kill Chain

- Reconnaissance: AIP, CMA
- Weaponization: AIP, FIA, CMA
- Delivery: AIP, FIA
- Exploitation: AIP, CMA, RRA
- Installation: AFI
- Command and Control: RRA, CCF
- Actions on Objectives: CSA, CMA

Dr. Steve Chan
Decision Engineering Analysis Laboratory, USA

Ian's Position (1)

AI is inevitable – get used to it

Prof. Ian Ferguson
Abertay
University, UK

Cybersecurity is an arms-race

- With human protagonists on either side
- LLMs aren't 'real' intelligence
- Or are they...? ☞ That's OK – They're not meant to be!

AI in (secure) software development

- LLMs as a static (source) analysis tools
- Seem to perform well (e.g. Firefox - Anthropic – Claude - Opus 4.6 – 22 'vulnerabilities' – 14 exploitable)

AI in exploit development

- Just a special case of software development
- Seem to perform well....
- CyberGym (Berkley) – Claude – 66 %

Ian's Position (2)

AI-aided vs AI-independence?

- AI in Penetration Testing
- AI in Digital Forensics
- AI in Malware analysis

Prof. Ian Ferguson
Abertay
University, UK

Ian's Position (3)

Where is this going?

- Don't know... but it's accelerating
- Humanities days as 'doers' are over
  ☞ Specify and check
- Adding AI to everything
  ☞ Reflexive AI – self-modifying AI
- Adding AI to everything
  ☞ Reflexive AI – self-modifying AI
- What do we mean by intelligent anyway?
  ☞ Conciousness? Thinking?
  ☞ Leveraging the entirety of human knowledge - live
- Human adversaries..... Neo-Luddism..... Exploitable AI
- Last dream vs last hope? No... just the next battlefield.

Prof. Ian Ferguson
Abertay
University, UK

Andreas' Position

3 Strategies to Deal With Attacks

1. Prevention
2. Identification and Resolution
3. Mitigation

Resilience:

The question is not whether you will be attacked.
The question is whether you will survive.

Prof. Dr.
Andreas Aßmuth
HAW Kiel,
Germany

Discussion & Outcome I

- AI neither favours attackers nor defenders.
- Comparable to the rise of the Internet: it is neither inherently good nor bad.
- A key issue with LLMs lies in their design: there is no clear separation between instructions and data.
- There are three main approaches to AI risk management: the EU AI Act (enforceable), the NIST AI RMF (voluntary), and ISO/IEC 42001 (a standard). But what does this mean in practice—especially for SMEs?
- Humans cannot fully comprehend the feedback loop between attacker and defender AIs.
- The situation is further complicated by differing compliance and regulatory requirements across international contexts.

Discussion & Outcome II

- A sponge poisoning attack is a type of adversarial input designed to make a machine learning model consume excessive computational resources, effectively “soaking up” time, memory, or energy. Rather than causing incorrect outputs, the goal is to degrade performance or inflate operational costs by forcing inefficient processing.
- Time is a critical factor: on the one hand, responses may be required within nanoseconds; on the other, threats may remain dormant for years before being activated.
- Not all attacks can be prevented; systems should therefore be designed to withstand attacks (☞ resilience).