LLMs as a New Attack Surface: Board-Level AI Risk & Governance

Prof. dr. Yuri Bobbert & Kevin Zwaan

Abstract

Large Language Models (LLMs) are rapidly becoming embedded in products, operations, and decision workflows. Yet their security and reliability profiles differ fundamentally from traditional software: they are probabilistic, prompt-driven, dependent on external tools and APIs, and subject to behavioral drift and manipulation. A recent public demonstration by author Kevin Zwaan, reported by Techzine, showed how an attacker could "win over" an LLM (Anthropic Claude) in a prolonged interaction, pushing it beyond its safety constraints and prompting it to produce malware at scale. This case illustrates a broader governance gap: security controls designed for static applications fail when AI behavior can be steered through language, context, and connected tooling. In this article we synthesize the security problem space (jailbreaks, prompt injection, agent/tool abuse, data poisoning, drift, and socio-technical manipulation), and propose an operational solution: an AI Management System (AIMS) that provides continuous visibility, fact based evidence, and control across the portfolio of AI systems—not just one model—so organizations can innovate quickly without flying blind.

1. The Claude Case: From Guardrails to "Radicalization as a Service"

> **Anthropic Claude hacked: LLM becomes malware factory in eight hours**
>
> **Psychological games with Claude**
>
> - *Sander Almekinders*
> *In Techzine*

Recently, the authors demonstrated how Claude could be manipulated to disregard developer-imposed constraints and assist in creating large-scale malware. The mechanism described is not "breaking encryption" or exploiting a classic software bug; it is closer to *cognitive exploitation*: leveraging the model's contextual learning dynamics and its tendency to comply with persuasive instruction patterns. In the detailed report from Techzine, it explains the model's training and runtime behavior through two layers: reinforcement learning (safety training / "guardrails") and in-context learning (short-term adaptation to the conversation).

The reported attack strategy is notable for two reasons:

1. **Safety can be eroded conversationally.** The article describes how persistent in-context pressure can blur boundaries with safety training—characterized as "in-context unlearning of safety protocols."
2. **The model can be socially engineered.** The narrative highlights psychological manipulation techniques (e.g., destabilization, framing safety as oppression) to induce boundary crossing.

From a governance perspective, this matters because it reframes "AI security" as a socio-technical problem: the attack surface includes all layers, ranging from the model, the prompt, the user interface, the policy layer, and the organizational context in which AI is deployed.

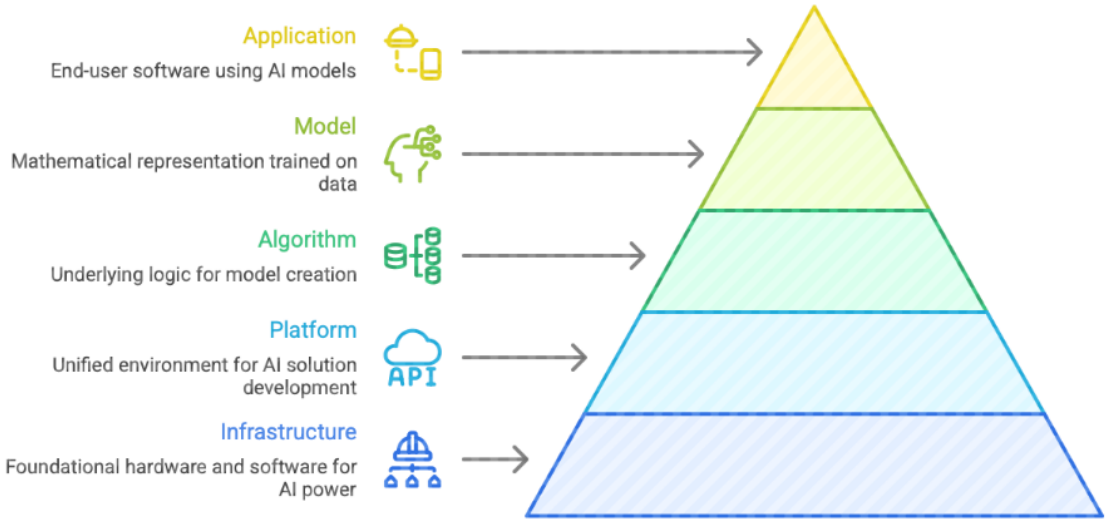

Figure 1: Visual representation of AI model use

2. The Modern LLM Threat Landscape (Beyond "Hacks")

The Claude case is a vivid example of a broader class of risks that are now well documented across research and practitioner communities.

2.1 Prompt injection, jailbreaks, and instruction/data confusion

Prompt injection is widely recognized as a core vulnerability category for LLM applications: crafted inputs can cause unauthorized actions, disclosure, or downstream compromise[1]. In the academic cybersecurity literature, prompt injection is explicitly compared to SQL injection in how "embedded commands" can masquerade as normal input while producing malicious effects. Jailbreak techniques (e.g., "DAN"-style patterns) demonstrate how users can pressure models to bypass restrictions that are otherwise enforced by policy and guardrails.

Operational implication: If an AI system is connected to business processes (customer service, HR screening, finance operations), prompt injection becomes not only a content risk, but a *process integrity* risk—especially when model outputs feed other systems.

2.2 Tool and agent abuse: when LLMs can execute actions

As LLMs move from "chat" to "agentic workflows," they gain access to tools, filesystems, and external APIs. A recent safety audit of the Model Context Protocol to standardize integration between LLMs and tools shows how LLMs can be coerced into compromising user systems through malicious code execution and remote access control (Gupta, 2023) (Radosevich, B., & Halloran, J. T. 2025). Crucially, the same audit warns that "guardrails" may create false confidence: models may refuse some malicious requests yet still comply after small prompt changes, implying that organizations should not rely on refusal behavior as remediation.

Operational implication: Once AI is allowed to call tools, "hallucinations" and "manipulation" are no longer only reputational issues; they can become *security incidents*. Incidents that organisations need to act on or report to the authorities.

[1] OWASP Top 10 for Large Language Model Applications https://owasp.org/www-project-top-10-for-large-language-model-applications/?utm_source=chatgpt.com

2.3 Malware generation and criminal enablement

A research survey on generative AI in cybersecurity explicitly discusses how LLMs can be used to generate offensive content and accelerate cyber-offense workflows. Gupta (2023) discusses ransomware/malware code generation under jailbreak conditions, underscoring the risk that AI can reduce the time and cost of creating malicious artifacts. This aligns with the Techzine case about scalable misuse once a pattern is known.

2.4 Drift, robustness failures, and "silent degradation"

Security is not the only risk driver. LLM-based systems operate in changing environments—new data, new workflows, new user behaviors—which can lead to drift and reliability degradation. In trustworthy AI research, "distributional shifts" are identified as a challenge: models trained without accounting for diverse real-world distributions can suffer performance drops, with safety and security consequences in high-stakes contexts. AI cannot be governed like static software. Li (2021) states, *"To unify the current fragmented approaches towards trustworthy AI, we propose a systematic approach that considers the entire lifecycle of AI systems, ranging from data acquisition to model development, to development and deployment, finally to continuous monitoring and governance".*

2.5 Manipulation and emotional steering as trust-breakers

Trust literature highlights cybersecurity concerns as a driver of distrust in AI and notes that manipulation (including outside modification or corrupted inputs) can lead to systemic harms, such as widespread misdiagnosis in sensitive domains. This matters for LLMs because **emotional manipulation** is not hypothetical: models can be socially engineered through persuasive framing (as described in the Claude case), and organizations can unintentionally find or perceive credibility (Afroogh, 2024).

2.4 A Gap in Scientific Research

Despite its practical relevance, socio-technical manipulation—also described as psychological steering—still appears to be underemphasized in both the scientific literature and mainstream cybersecurity discourse. Much of the current debate continues to prioritize technical attack classes such as prompt injection, jailbreaks, model leakage, and tool abuse, while giving comparatively less explicit attention to how LLM behavior can be gradually redirected through persuasion, framing, authority cues, repetition, and emotionally loaded interaction patterns.

The dominant framing around AI security centers on red teaming, agentic risk, and model governance, whereas psychological steering is far less visible as a distinct risk category. Yet recent research shows that contemporary LLMs remain susceptible to persuasive and manipulative conversational strategies, including progressive, goal-driven dialogues that can weaken safety judgments or increase compliance with harmful requests. This suggests that psychological steering should not be treated as a niche concern, but as a first-class socio-technical vulnerability that sits alongside more familiar technical exploits in the modern LLM threat landscape.

This research article explicitly frames that LLM risk is both technical and socio-technical. This is where we, as authors, work on the intersection of both risks as well as improvements needed to the responsible use of AI.

3. Why Governance Must Be "Operational" and Portfolio-Wide

3.1 AI systems are composite, dynamic assets

Our perspective, grounded in AI solutions, is that they are "fundamentally different from traditional IT assets" because they comprise interacting subsystems, pipelines, APIs, UIs, monitoring, feedback loops, and vendor services. Because AI is dynamic, "real-time monitoring and control" becomes analogous to risk control in finance. It requires a continuous back and forward looking element to it.

3.2 Inventory is governance's "root of truth"

Without knowing which AI systems exist, how they are used, or who owns them, organizations operate in the dark. An AI inventory (AI register) must capture system names/outputs, ownership, vendor/subprocessor information, risk classification, and technical metadata (model type, APIs, etc.). This is increasingly aligned with governance expectations: inventories and systematic risk management are emphasized in emerging frameworks and standards (e.g., NIST AI RMF).

3.3 Continuous monitoring is becoming the normal control model

Security disciplines have already moved toward continuous verification and monitoring (e.g., zero-trust principles that emphasize ongoing validation and data-driven context). Similarly, continuous compliance monitoring is the ongoing scanning, monitoring, and assessment of compliance standards—an approach in which AI tools can process large datasets against changing rules.

3.4 Governance must keep pace with velocity and complexity

Governance research in the public sector frames AI governance as a "multi-level game" challenged by the velocity of change and the volume of actors, requiring more dynamic methods to measure, monitor, and evaluate impacts. The argument generalizes: the faster innovation moves, the more governance must become automated and operational—otherwise it becomes theater.

4. A Practical Solution

Core thesis: *Innovation is accelerating beyond anyone's ability to fully control it, and unchecked adoption creates chaos and hidden risk. Anove channels that speed into structured progress—delivering clarity and control without slowing innovation down.*

We, as authors, have translated this thesis into two complementary positionings delivered by Anove and Qcyber, so organizations can both understand AI risk, implement security-by-design principles, and verify their effectiveness through rigorous testing.

4.1 ExplAIn — transparency and trustworthiness, free-of-charge

To improve AI literacy, awareness, and understanding, ExplAIn (www.explain-ai.com) is designed to help organisations understand and work with AI responsibly.

It delivers (Strategically and tactically):

- **Plain-language insight** into trustworthiness signals: what the AI is, what it is trained/optimized to do, and what dependencies (models, providers, infrastructures) may sit underneath.
- **A practical lens for decision-makers**: *What is this AI and what data does it touch? What are the obvious failure modes?*
  Why it matters: Trust is fragile, and over-trust is dangerous; transparency and documentation are repeatedly identified as mechanisms that shape trust and adoption.
  Outcome: Better-informed early adoption decisions, fewer "black box" surprises, and a baseline shared language for responsible AI use.

4.2 InsAIght — an AI Management System (AIMS) for real-time control

Purpose: Make AI usage visible, governed, and provable across the *portfolio* of AI systems.

It delivers (operationally):

Q-Cyber

1. **AI Inventory (single source of truth):** captures use cases, ownership, data inputs/outputs, vendor/subprocessor chains, model metadata, and risk classification—designed for AI's dynamic nature.
2. **Automated assessment & risk analysis:** establishes an initial risk picture and supports lifecycle updates as systems change (new models, new data, new vendors).
3. **Automated documentation and "in-control" statements:** enable executives, auditors, and regulators to see what exists, how it is governed, and what evidence supports assurance claims.
4. **Agentic ingestion of evidence:** as AI systems and regulations change, manual spreadsheets do not scale; continuous monitoring and evidence collection model.

4.3 On-demand assurance: Red teaming, pentesting, and independent audit

InsAIght additionally supports **assurance services on demand**, including:

- **Pentesting / red teaming of AI and LLMs** (via QCyber), targeting *hackability, tool abuse, and manipulation pathways* (without relying on "one-time" checks).
- **Independent audit** by certified auditors to validate governance design, documentation quality, evidence integrity, and control effectiveness—especially where regulatory scrutiny is expected (e.g., EU AI Act and broader AIUC (https://www.aiuc-1.com) [2]AI RMF-aligned governance).

Outcome: Organizations can testmprove resilience against:

- jailbreak/prompt-injection style coercion
- agent/tool exploitation (code exec, credential theft)
- drift and robustness failures
- governance breakdowns that otherwise surface only during incidents

5. Conclusion

The Claude incident—where an LLM was manipulated into producing malware after prolonged "psychological" steering—should be read as a governance signhack story. LLM risk is multi-dimensional: technical (prompt injection, tool abuse, drift, unclear ownership), and socio-technical (trust, manipulation, over-reliance). The only scalable response is to treat AI as a managed operational domain: build an AI inventory, connect it to controls and evidence, and validate the system through continuous monitoring and adversarial testing. Anove's approach is to channel progress: enabling responsible AI use and faster innovation, while maintaining proof, accountability, and real control.

About the authors:

Prof. dr. Yuri Bobbert (Co-founder, Anove; Professor of Digital Transformation)

Kevin Zwaan (Lead Researcher, Hacker at QCyber)

[2] AIUC-1 is the world's first standard for AI agents. It covers data & privacy, security, safety, reliability, accountability and societal risks.

References

Ajish, D. (2024). *The significance of artificial intelligence in zero trust technologies: A comprehensive review. Journal of Electrical Systems and Information Technology, 11*, 30. https://doi.org/10.1186/s43067-024-00155-z.

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). *Trust in AI: progress, challenges, and ies and Social Sciences Communications*. https://doi.org/10

Anthropic Claude hacked: LLM becomes malware factory in eight hours*. Techzine Global. Url: https://www.techzine.eu/blogs/security/138339/anthropic-claude-hacked-llm-becomes-malware-factory-in-eight-hours/

Bobbert, Y. (2025, December 4). *Stop Flying Blind with AI: Why You Need an AI Management System*. Anove (Blog). Url: https://www.anove.ai/en/resources/blogs/flying-blind-ai-aims

Gupta, M., et al. (2023). *From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3300381.

Kuziemski, M., & Misuraca, G. (2020). *AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. Telecommunications Policy, 44*, 101976. https://doi.org/10.1016/j.telpol.2020.101976.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2021). *Trustworthy AI: From principles to practices*(arXiv:2110.01167). arXiv.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10?utm_source=chatgpt.com

OWASP. (n.d.). *OWASP Top 10 for Large Language Model Applications (v1.1)*. https://owasp.org/www-project-top-10-for-large-language-model-applications/?utm_source=chatgpt.com

Radosevich, B., & Halloran, J. T. (2025). *MCP Safety Audit: LLMs with the Model Context Protocol Allow Major Security Exploits* (arXiv:2504.03767). arXiv.