Member-only story

The Hidden Risk in Your AI Strategy

Why “move fast and break things” breaks down with agentic AI

6 min read · Just now

Gunnar Östberg

Generative AI has forced many organizations to rethink how software works. Much of the discussion, however, still focuses on surface-level capabilities and

productivity gains, while missing a more fundamental shift in how these systems behave.

Unlike traditional software, generative AI systems are not rule-followers in the conventional sense. They are interpretation engines. A large language model does not execute predefined logic step by step. Instead, it interprets whatever input it is given and attempts to respond in a helpful way, regardless of where that input comes from or what intention lies behind it.

This design choice is precisely what makes generative AI so powerful. It is also what makes it fundamentally different from every major software paradigm around which our existing security models have been built.

Interpretation as an attack surface

Traditional software typically fails when inputs violate explicit constraints or predefined rules. Generative AI fails in a different way. It fails when inputs are persuasive.

A language model cannot reliably distinguish between a legitimate instruction and a malicious one embedded in otherwise ordinary content. From the model's perspective, both appear as text that might be relevant to the task at hand. This is the reason prompt injection exists.

If an AI system reads emails, documents, tickets, or messages, it will treat everything inside them as potential guidance. An attacker does not need direct access to the system itself. It is often enough to gain access to the data that the system consumes.

This is not a bug or an oversight. It is the expected behavior of an interpretive system.

The risk increases sharply once the model is given access to context that actually matters, such as internal documents, customer data, operational knowledge, credentials, or tool interfaces. At that point, a successful manipulation is no longer a minor nuisance. It becomes a genuine security incident.

When AI stops advising and starts acting

Many organizations are now moving beyond AI systems that merely summarize information or make recommendations. Increasingly, they are deploying systems

that can act on their behalf.

Such systems may be allowed to approve transactions, send communications, modify records, trigger workflows, or integrate directly with other business systems. This is the point at which the risk profile changes qualitatively.

An AI agent with authority is no longer just a consumer of information. It becomes part of the organization's execution surface. If the model misinterprets instructions, the error does not stay confined to the model itself. It propagates into real systems and real processes.

What makes this particularly dangerous is not primarily what the agent can do, but how legitimate its actions appear. When an AI approves a transaction, updates a record, or sends a message, it does so using valid credentials, through approved interfaces, and along normal operational paths. From the outside, nothing appears suspicious.

The system has done exactly what it was authorized to do. Just not what was intended.

Scale changes the nature of the problem

These risks become substantially more serious as the number of agents increases.

A single agent with limited permissions is usually manageable. Thousands or millions of agents, each carrying different contexts, access rights, and behavioral goals, represent a very different situation, especially if those agents operate in environments where inputs are not fully controlled.

Such environments may include public feeds, shared workspaces, semi-open collaboration environments, or agent-to-agent interaction spaces. In these settings, malicious instructions do not need to target a specific system. They can be broadcast broadly, with the expectation that, sooner or later, a sufficiently privileged agent will encounter and misinterpret them.

At that point, the security problem shifts character. The organization is no longer defending a well-defined perimeter. It is defending a population.

We are beginning to see early public examples of this pattern. Some emerging platforms now host large populations of AI agents that interact in shared or semi-

public environments while still carrying persistent context and, in some cases, delegated permissions on behalf of their users.

Systems such as Moltbot, combined with agent-centric interaction spaces like Moltbook, make this dynamic visible. Not because they are uniquely risky, but because they expose a structural shift. When many semi-autonomous agents operate at scale in partially adversarial environments, traditional perimeter-based security assumptions no longer apply.

Why this does not fit existing security intuition

Most established security frameworks are built on assumptions that no longer fully hold. They assume that execution engines are deterministic and bounded, that code does not reinterpret instructions creatively, and that intent is explicit and unambiguous.

Generative AI violates all three assumptions.

As a result, many organizations apply familiar controls and conclude that they are adequately protected. The problem is not that these controls are inherently wrong. The underlying threat model has changed.

An interpretive system with authority behaves less like traditional software and more like a junior employee. The difference is that it operates at machine speed, at massive scale, and without human judgment. Unlike a human, it cannot recognize when someone is deliberately trying to deceive it.

Zero trust applies more strictly here, not less

The appropriate response is not to avoid generative AI altogether. The potential benefits are real. However, these systems must be treated as untrusted components by default, often more strictly than many third-party integrations.

In practice, this implies several hard constraints:

- All input must be treated as potentially hostile, not merely malformed but intentionally manipulative. This is a security requirement, not just a data-quality concern.
- Permissions must be minimal, and persistent access should be rare. If an agent does not genuinely need access to sensitive data or systems, it should never have it. Convenience is not a sufficient justification.

- High-impact actions should require human confirmation. AI systems can recommend actions, but humans must commit to them.
- Every action must be logged with sufficient context to allow reconstruction and accountability.
- Outputs should be validated before execution. Language models are designed to comply, not to resist. Raw output should not directly trigger real-world consequences.
- Isolation matters. AI systems should be behind controlled interfaces rather than wired directly into production environments.

None of this is exotic or controversial in itself. What is new is how often these principles are set aside when speed becomes the primary objective.

The real gap is organizational, not technical

The uncomfortable truth is that many organizations lack the maturity required to deploy agentic AI safely.

They often lack clear ownership, effective collaboration between engineering, security, and business functions, adequate monitoring and containment infrastructure, and governance models that can evolve at the pace of technological change. As a result, they improvise.

Improvisation can work for a while. Interpretive systems, however, are very good at exploiting ambiguity, and eventually the gaps begin to show.

The innovation trap

The largest risk is cultural rather than technical.

AI initiatives are frequently led by people who are rewarded for speed, experimentation, and visible progress. In that environment, security is easily perceived as friction, and controls are dismissed as legacy thinking.

There is some truth to this tension. Security does slow things down. However, ignoring security does not remove the trade-off. It merely delays the cost and increases it.

Security that is designed in from the start constrains behavior. Security that is added later focuses on containing damage. These are not equivalent outcomes.

A responsible path forward

Organizations that succeed with AI tend to approach it differently.

They invest in educating innovation leaders about concrete threat models rather than abstract warnings. They embed security expertise within AI teams instead of treating it as an external approval step. They create fast paths with clear guardrails, allowing low-risk use cases to move quickly while subjecting high-authority systems to appropriate scrutiny. They measure both delivery speed and safety, recognizing that rewarding only velocity leads to fragility. And they invest in fundamentals such as identity management, access control, logging, and incident response, treating them as prerequisites rather than optional enhancements.

The bottom line

Generative AI represents a genuine shift in capability. Its core strength, interpretation, is also its core weakness.

As soon as AI systems are allowed to act, not merely to suggest, they become part of an organization's attack surface. Organizations that treat them as trustworthy by default will eventually pay for it, not because the technology failed, but because it behaved exactly as designed.

The question, in the end, is straightforward: are you building AI systems you can trust, or AI systems that merely work until they do not?

That distinction will determine which organizations benefit from AI in the long run, and which ones become cautionary examples.

AI AI Agent Security

Edit profile

Written by Gunnar Östberg