LLMs, Tools, Security

Presentation at RITE 2026

Bucharest, România

Bertram Dorn
Principal, Office of the CISO
AWS

Ionut Ionescu
Principal Security Solutions Architect
AWS



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Differences

Perspectives

Suggestions

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Differences

Perspectives

Suggestions

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

AI systems are not traditional software



Autonomous Decisions

The system decides what to do next - not bound by finite state machines



Adaptive Behaviour

Output varies based on context, history, and learned patterns - not deterministic code paths



Dynamic Data Access

The AI determines which data it needs at runtime - not predefined queries



Shifting Trust Boundaries

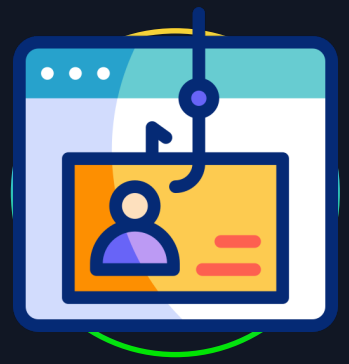
Where is the model hosted?

Self-hosted vs. provider changes scope

Every integration adds a boundary

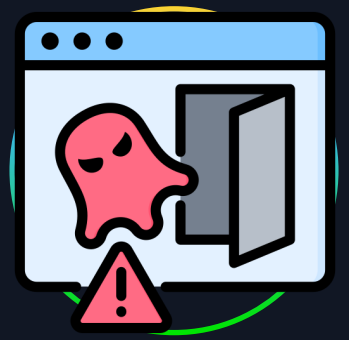
© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Can you stop AI?



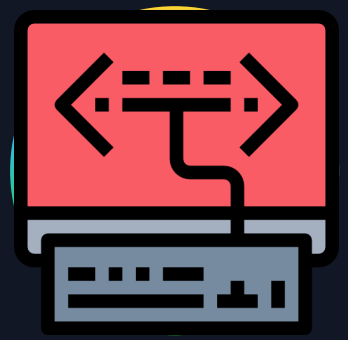
Sensitive Data in Public AI

A government officer pastes a **classified briefing into a public LLM** to get a summary. The data is now in a third-party system with no retention controls.



Unvetted Tools for Analysis

An analyst uses an **unapproved AI tool to process citizen records**, bypassing data protection controls and audit requirements.



Code Generation Without Review

A developer deploys **AI-generated code containing hardcoded credentials and insecure defaults** — no security review was performed.

aws

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

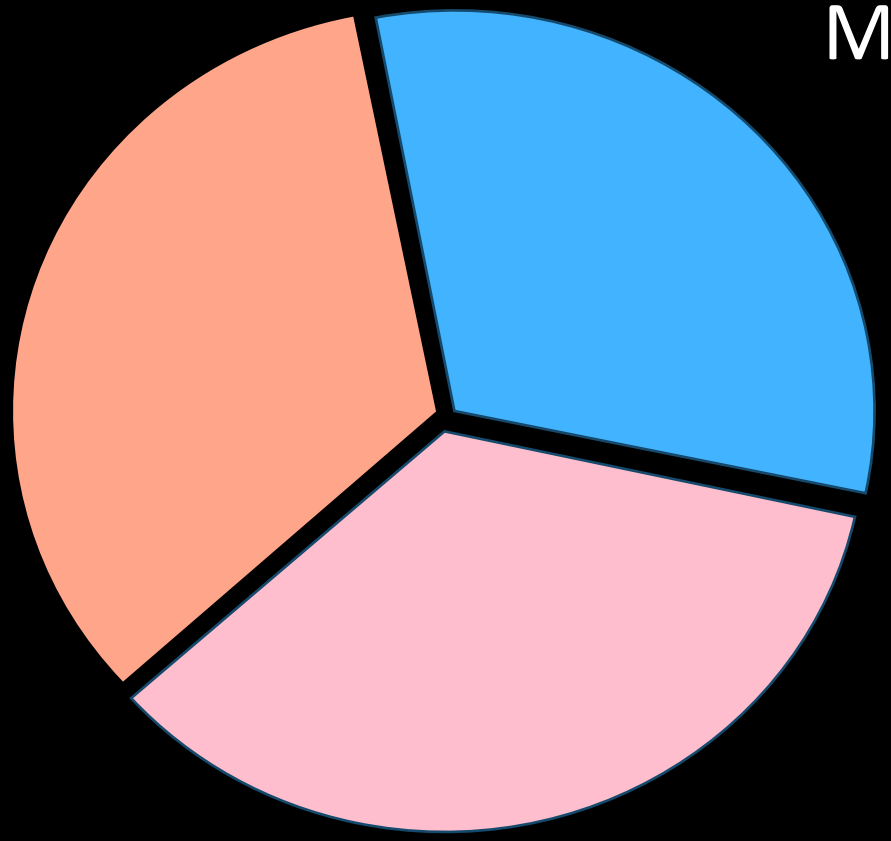
AI and the real world

IoT	ICS/OT	Medical	Critical National Infrastructure (CNI)
Smart buildings, sensors, actuators	Industrial control systems, SCADA	Diagnostic systems, robotics, dosing	
Connected devices controlled by AI	**Manufacturing, energy, utilities**	**Patient safety at stake**	**Transportation, water, power grids**

Key insight: AI security isn't just about data — when AI can actuate, security failures become safety failures.
The scope expands far beyond (mere) data protection.

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

AI Workload Trinity



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

What changes with AI?

Defense in depth is an area for traditional workloads and AI workloads. But, what changes with this compared to traditional workloads?

Functionality Non-Deterministic

Same prompt can produce different responses…due to settings with the model

Input Becomes Code

User prompts become instructions that the AI executes. Data and instructions are no longer separate

Autonomous Decision-Making

AI makes independent decisions and take actions without explicit programming (reasoning)

Dynamic Learning & Adaptation

Agents can learn from interactions and adapt behavior over time

aws

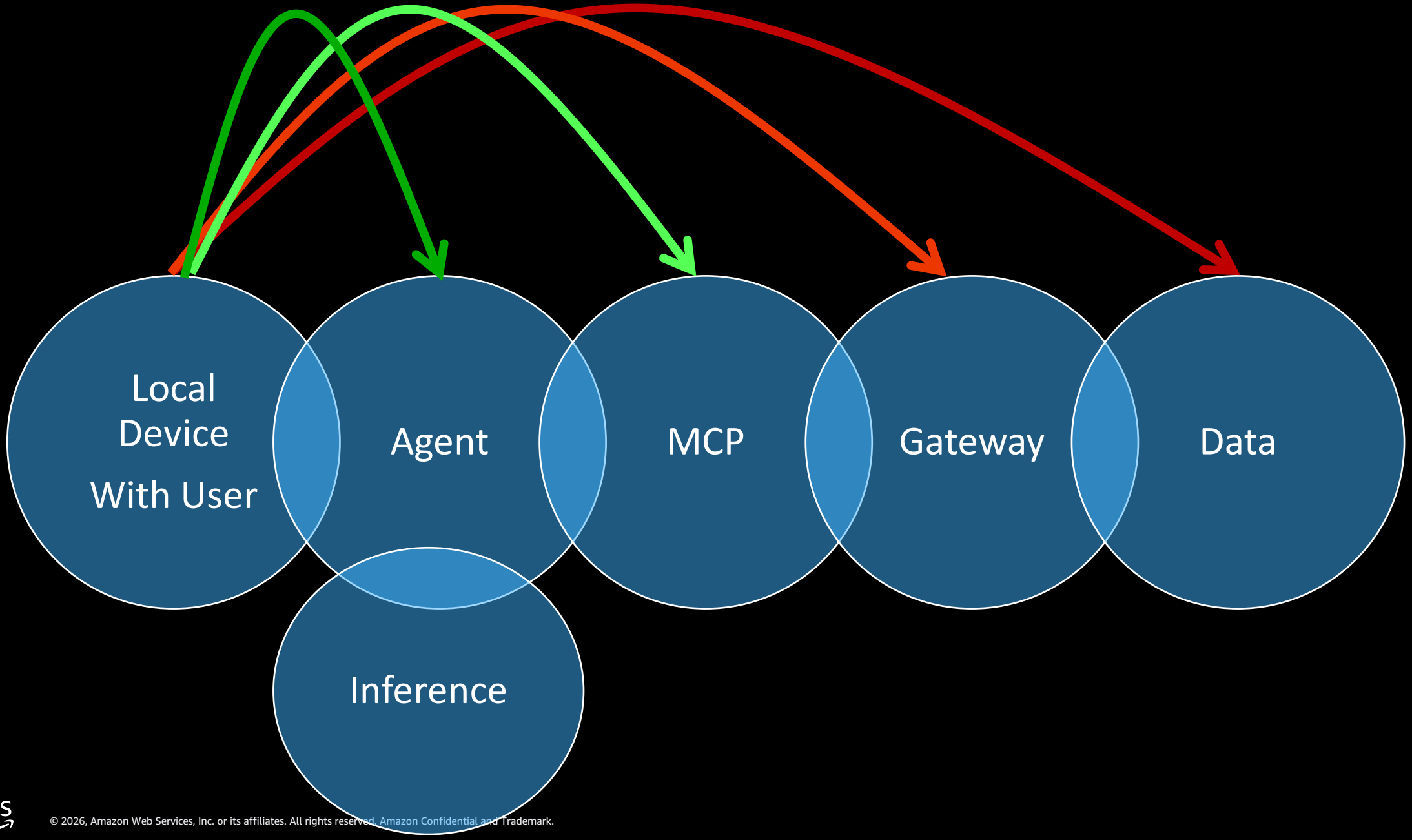
© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Differences

Perspectives

Suggestions

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

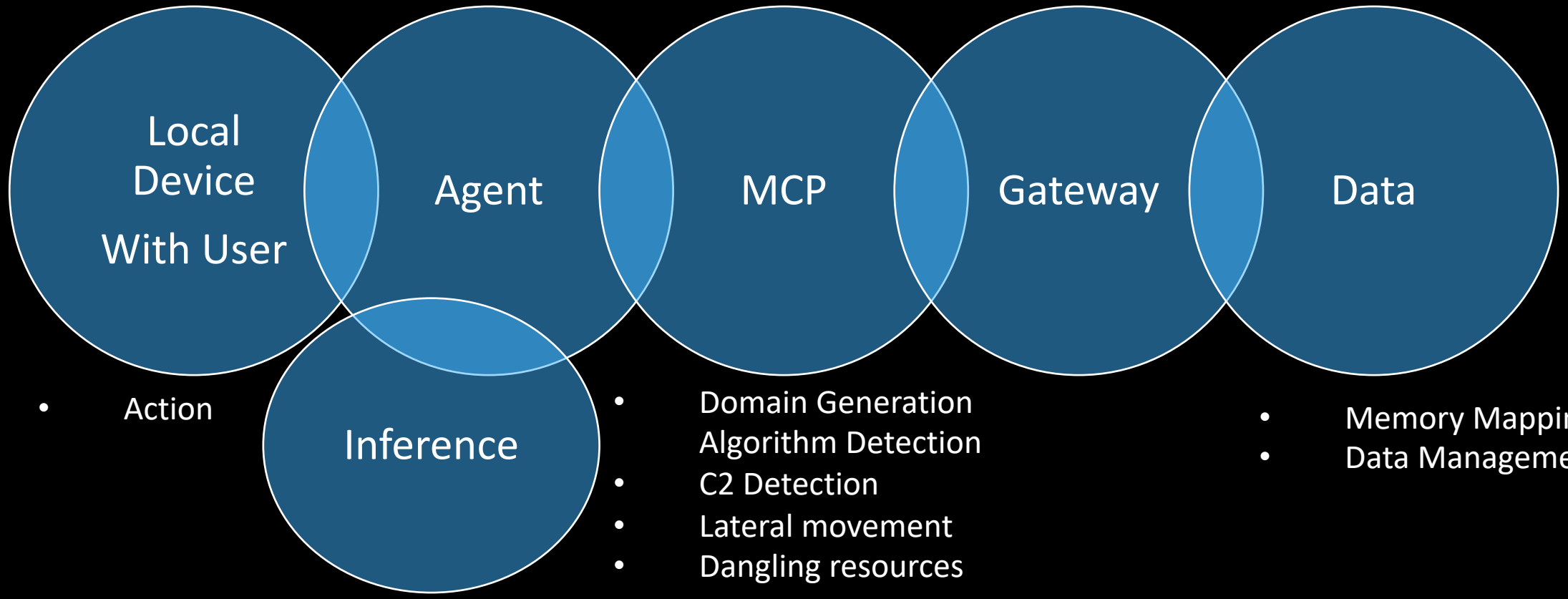


© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

The architect’s perspective

- Token Balance
- Summaries
- Follow-ups

- AuthN / AUthZ
- Scale
- Don’t Fail



- Action

- Domain Generation Algorithm Detection
- C2 Detection
- Lateral movement
- Dangling resources

- Memory Mapping
- Data Management

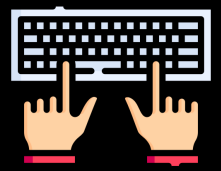
aws
© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

The "Lethal Trifecta": attributes



Access to Private Data

PII, credentials, financial data, health records, trade secrets accessible at runtime



Untrusted Inputs

User prompts, external API responses, third-party tool outputs — all potential attack vectors



External Communication

Ability to call APIs, send emails, modify databases, trigger workflows (actuation)

...while retaining Persistent Context in Memory

The system remembers and builds upon every interaction — including poisoned ones

When all three are present with persistent memory, the system can be weaponised against its own users

aws

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/

Three layers of AI security



Platform security

Compute, storage

Identity (roles, policies)

Networking, isolation

Data encryption at rest & in transit

Logging and audit trails



Model security

Model integrity and provenance

Training data poisoning risks

Model weights protection

Inference endpoint security

Guardrails and content filtering



Application security

Plugin/tool security (MCP)

User input handling

Orchestration logic

Output validation

Session and memory management

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Foundational Controls for AI Systems

Identity & Least Privilege

Every agent and plugin gets a unique identity with minimal permissions. No shared credentials. Agents operate with the narrowest possible scope — read-only where possible, action-specific where required.

Network Segmentation

AI components isolated in dedicated network segments. Agent-to-tool traffic restricted to approved paths. External API calls routed through controlled egress points with inspection.

Input & Output Validation

All inputs sanitized — user prompts, API responses, and MCP tool outputs alike. Content filtering on both ingress and egress. Structured output schemas to prevent injection through agent responses.

Secrets Management

API keys, tokens, and credentials stored in dedicated secrets management. Never embedded in prompts, configurations, or agent memory. Automatic rotation.

Monitoring & Observability

Every agent action logged with full context: who, what, when, why. Real-time alerting on anomalous behaviour patterns. Complete audit trail for compliance and incident response.

aws

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Human-in-the-loop: when must a human decide?



AI Acts Autonomously

LOW RISK • REVERSIBLE

The AI system takes actions independently without human intervention.

Example: TravelBot searches available flights and presents options to the user.



AI Proposes, Human Confirms

MEDIUM RISK

The AI recommends an action but waits for explicit human approval before executing.

Example: TravelBot selects the best flight and asks the user to confirm before booking.



Human Decides, AI Assists

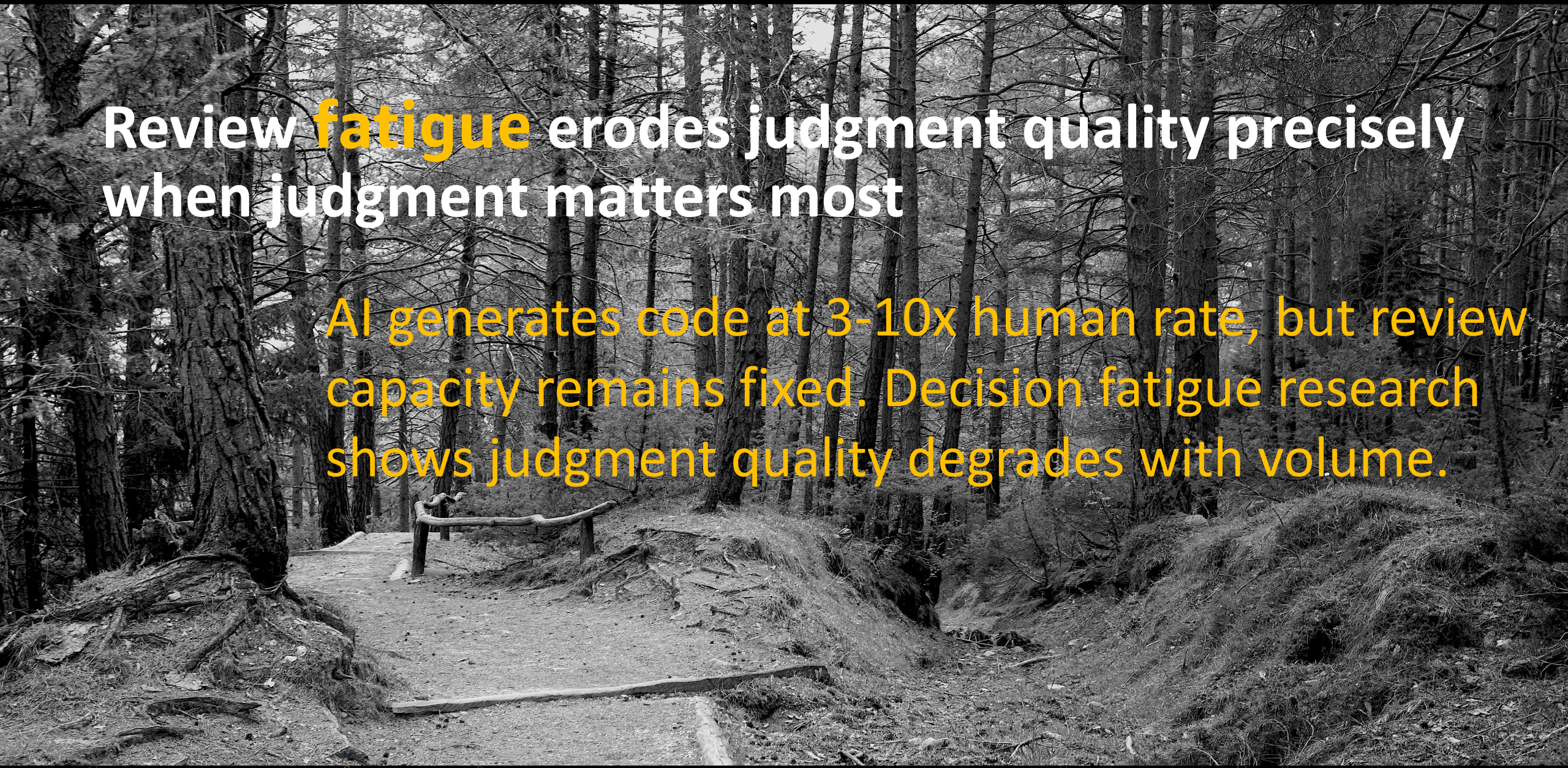
HIGH RISK • IRREVERSIBLE

A human makes the decision; the AI provides information and analysis to support it.

Example: Large payments, access to sensitive data, changes to critical systems.

aws

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.



Review fatigue erodes judgment quality precisely when judgment matters most

AI generates code at 3-10x human rate, but review capacity remains fixed. Decision fatigue research shows judgment quality degrades with volume.

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Defender’s perspective: exploit gap has collapsed

Mean Time-to-Exploit

2.3 years (2018) → 5 days (2024) → ~20 hours (2026)

Source: zerodayclock.com

Attackers exploit newly disclosed vulnerabilities in a median of 5 days.

Your team takes 32–38 days to patch.

For critical internet-facing systems, mass exploitation begins before most orgs finish reading the disclosure.

Ask yourself: *How quickly can we safely patch increasing numbers of zero-days?*

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why the landscape is different

Scale

AI finds vulnerabilities across entire codebases simultaneously, not one at a time

Speed

Discovery-to-exploit collapsed from weeks to hours

Accessibility

The skill floor for sophisticated attacks has dropped dramatically; threat actors are already using AI offensively

97%
of orgs with AI incidents lacked proper AI access controls

96%
of tech professionals see AI agents as growing risk

94%
of executives say securing AI before deployment is critical

Sources: IBM Cost of Data Breach 2025 • SailPoint AI Agents Report 2025 • IBM Unify Security 2024

aws

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The impact

Vulnerability overload

Massive surge in disclosed vulnerabilities - your team cannot patch fast enough.

Compliance

Compliance and audit frameworks are being rewritten in real-time.

Risk management

Board-level risk models built on pre-AI assumptions are now wrong.

CISOs must re-baseline. Mean time-to-exploit assumptions of weeks or months are obsolete.

aws

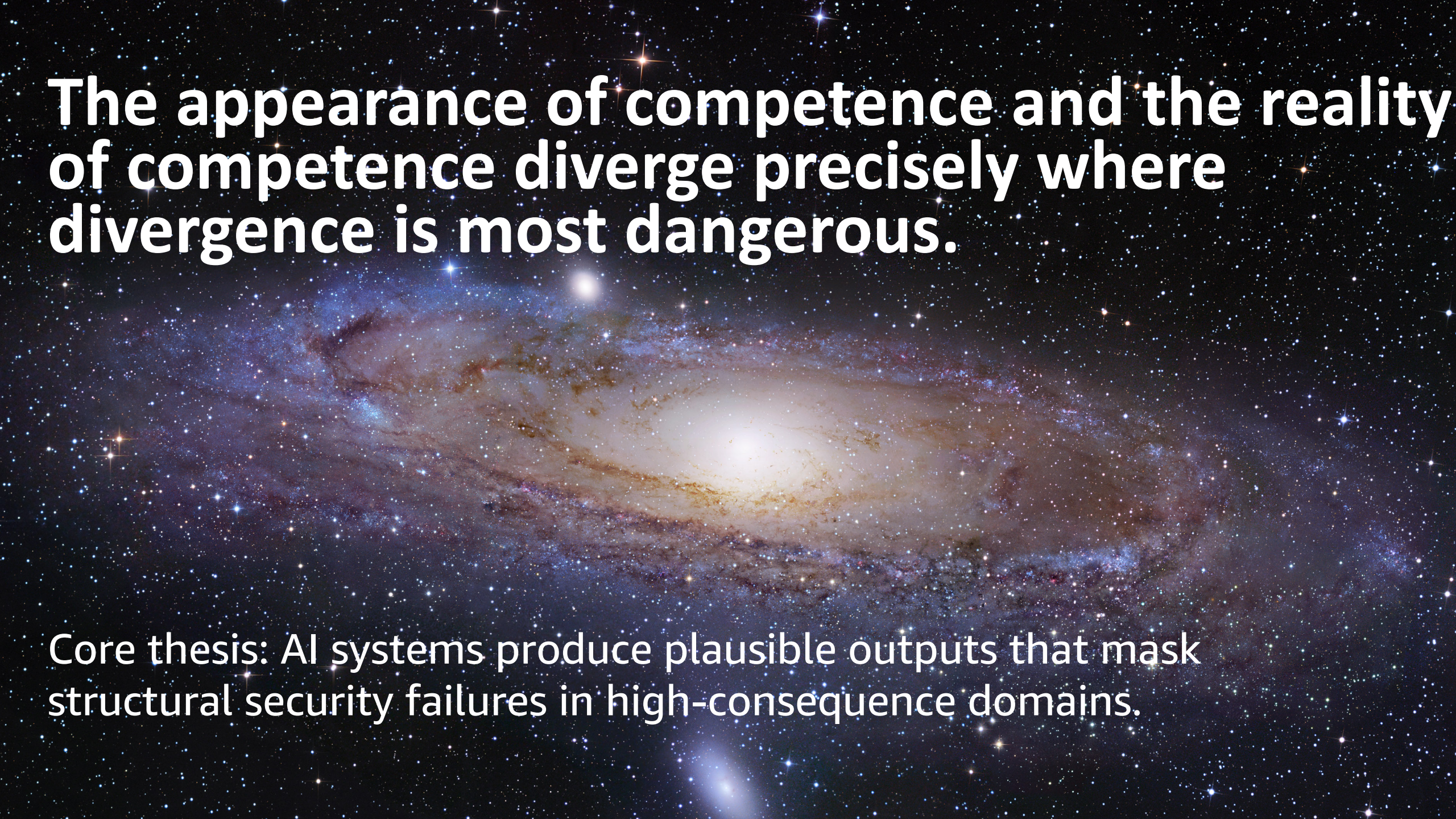
© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Differences

Perspectives

Suggestions

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.



The appearance of competence and the reality of competence diverge precisely where divergence is most dangerous.

Core thesis: AI systems produce plausible outputs that mask structural security failures in high-consequence domains.



AI has no stake in quality — no pride, no reputation, no fear of consequences.

The accountability vacuum: humans inherit full responsibility for code they did not write, without the authorial familiarity that ordinarily accompanies that responsibility.

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.



Documentation compression creates institutional amnesia — critical details do not survive summarization.

AI-mediated summarization loses the nuances, edge cases, and design rationale that matter most in future decisions.

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.



The engineers who can review AI code are not produced by AI tools.

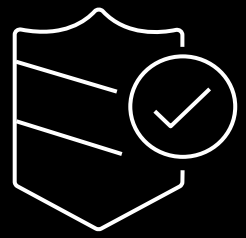
Deep expertise requires years of hands-on implementation. AI-augmented workflows risk eroding the talent pipeline that makes safe AI adoption possible.

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Key takeaways

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

How to strategically think about AI security



1) AI will become embedded everywhere.
AI drives economic and productivity gains. It will be embedded throughout our customers' infrastructure, applications, enterprises, and supply chains.

2) Security needs to be embedded everywhere AI is integrated.
AI connects to everything. Defense-in-depth security approach must be applied to all connected systems: from infrastructure to application layer.

3) Securing AI can drive enterprise-wide security improvements.
By building secure-by-design and secure-by-default AI, we can help customers improve security postures across workloads and their enterprise everywhere AI is embedded.

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Defense in Depth Security

THREAT DETECTION & INCIDENT RESPONSE

APPLICATION PROTECTION

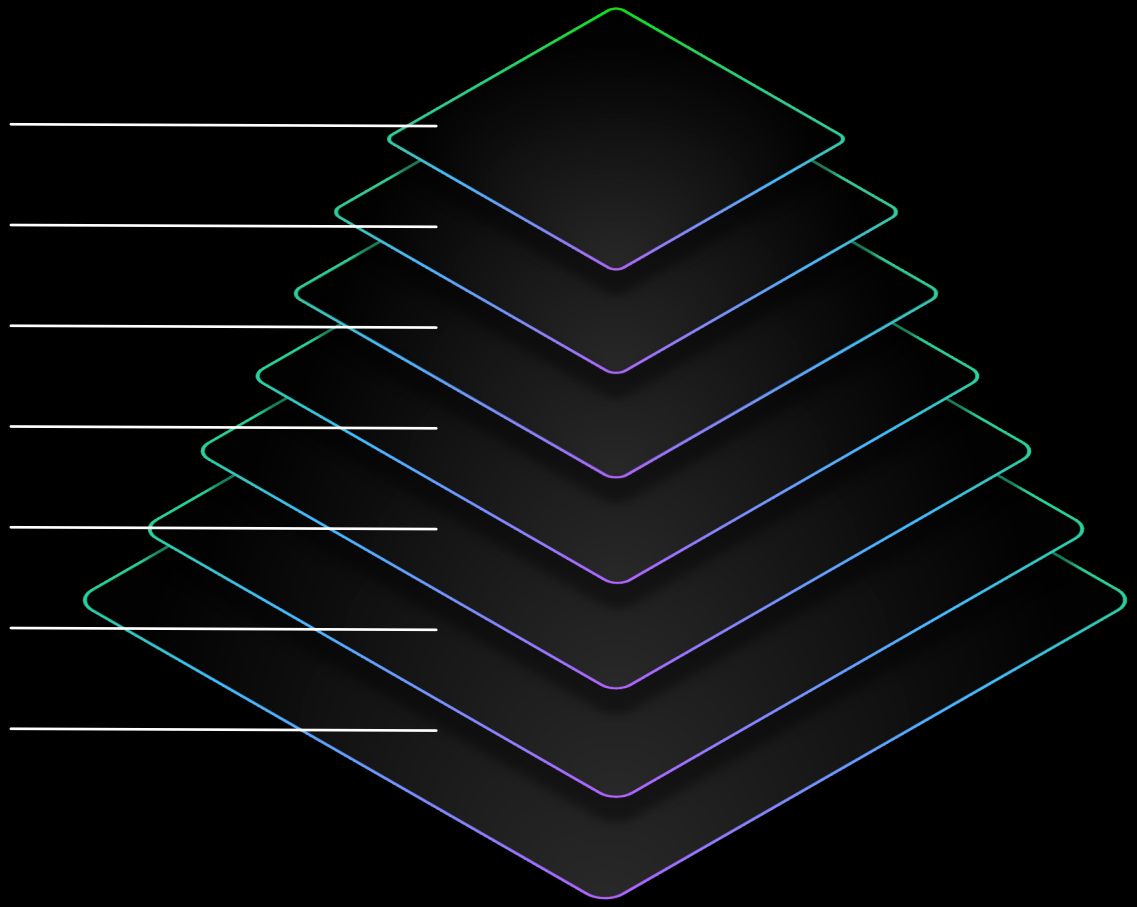
NETWORK & EDGE PROTECTION

DATA PROTECTION

IDENTITY & ACCESS MANAGEMENT

INFRASTRUCTURE PROTECTION

POLICIES, PROCEDURES, AND AWARENESS



aws

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Security and Responsible AI

Focuses on protecting the systems, data, and infrastructure from malicious attacks and unauthorized access.

Ensures AI systems are developed and used ethically and safely, preventing issues like bias, misinformation, and other societal harms

© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.

Thank you

ionutrei@amazon.co.uk

bedorn@amazon.de



© 2026, Amazon Web Services, Inc. or its affiliates. All rights reserved. Amazon Confidential and Trademark.