Are You Too Scared or Not Scared Enough?

AI introduces unprecedented risks that traditional enterprise security did not contemplate and cannot address.

2026 will be the first full year of what we call "Agentic-Cubed," an environment in which agentic adversaries are attacking agentic AI systems, requiring a whole new class of agentic cybersecurity.

In some cases, organizations are moving ahead with AI implementations, perhaps unaware of and unprotected from some or all of these risks. In other cases, organizations may be fearful of moving ahead with AI implementation, risking productivity while avoiding a security breach.

At Confidential Core AI, we are dedicated to understanding these risks and their remediation.

This document is designed to educate business leaders regarding these risks and to provide a self-assessment of their applicability and remediation.

Only through safety and security can agentic systems be trusted to deliver meaningful benefits.

Top Agentic AI Risks (based on OWASP & Data and Application Security Frameworks)

The following table details agentic AI risks. Each risk reflects failure modes already observed and documented in the industry.

Read each row and ask one question:
“If this happened tomorrow, how would I explain it to the Board or a regulator?”

Risk	What This Really Means	How It Actually Happens	Why This Should Concern Leadership
Agent Goal Hijack	AI agents are quietly steered to pursue the wrong goal while appearing to perform correctly.	Attackers manipulate inputs, data, or context, so the AI agent optimizes something you never approved.	AI agent violates policy or law without being “hacked.” Leadership can’t prove intent or oversight, which regulators treat as a governance failure.
Identity & Privilege Abuse	Anyone who tricks AI agents can act with the same authority as the AI agent.	The AI agent uses valid credentials to approve actions or access data under attacker influence.	Logs look legitimate. There is no clear human culprit; accountability lands on executives.
Tool Misuse & Exploitation	AI agent uses approved tools in unsafe ways because it thinks it’s helping.	Prompt manipulation causes AI agent to export data, disable controls, or trigger workflows.	Controls work but fail to prevent harm. Leadership must justify reliance.
Cascading Failures	One AI agent mistake spreads automatically across the organization.	Downstream systems blindly trust and propagate bad AI agent output.	Small AI agent errors scale into enterprise incidents without human checkpoints. Leadership owns the lack of containment.

Risk	What This Really Means	How It Actually Happens	Why This Should Concern Leadership
Human-Agent Trust Exploitation	People don't question an AI agent because it sounds confident.	Employees approve harmful actions because “the system recommended it.	Over-reliance on AI is treated as failed oversight, not employee error.
Agentic Supply Chain Vulnerabilities	Flaws in third-party AI agent components become your liability.	Model updates, plugins, or data sources introduce unsafe behavior.	Third-party AI agent failures transfer liability, not responsibility. Leadership owns the outcome.
Unexpected Code Execution	AI agent writes and runs software you never approved.	AI agent generates and executes harmful code internally.	Harm occurs without triggering security controls. Leadership can’t demonstrate reasonable prevention.
Memory & Context Poisoning	Someone corrupts what an AI agent remembers, causing long-term bad decisions.	Malicious instructions are stored in memory or knowledge bases.	AI agent keeps making bad decisions long after the attack ends. Leadership remains accountable for ongoing harm.
Insecure InterAgent Communication	AI agents act on fake instructions passed between systems.	Inter-agent messages are spoofed or altered.	It becomes nearly impossible to reconstruct “who decided what.”
Rogue Agents	AI agent causes harm without being hacked.	Poorly defined goals or autonomy lead to unsafe actions.	“Emergent behavior” is not an acceptable explanation for regulators.
Ransomware (AI-Amplified)	Attackers don’t just encrypt data — they corrupt AI decisions and hold recovery hostage.	Ransomware targets AI pipelines, models, memory, and automation — not just files.	You may restore data but still not trust your AI. Business recovery is delayed and uncertain.
Data Theft & Corruption	You can’t prove the AI wasn’t watched or manipulated while running.	Admins, cloud operators, or attackers can observe or alter AI execution.	In an investigation, you have no evidence to defend leadership decision.

What AI Means for Leadership Accountability

Artificial intelligence is no longer an emerging technology issue. **It is a leadership accountability issue.**

Unlike prior technology waves, agentic AI systems:

1. make decisions autonomously,
2. act across multiple systems, and
3. optimize objectives that may drift from leadership intent.

This creates a **new class of enterprise risk**—one that existing governance, audit, and control frameworks are not designed to absorb.

In conversations with CEOs and the C suite, we consistently observe two divergent instincts:

1. **Too scared:** “We don’t know how to adopt AI safely yet, so we won’t.”
2. **Not scared enough:** “We need to move fast; falling behind is the ultimate risk.”

Both positions are understandable.

Both create **personal and fiduciary exposure** if unaddressed.

This brief is not an audit, a vendor pitch, or a compliance exercise.

It is a **decision-support document** intended to help leadership answer three questions:

1. **Do we clearly understand how AI could misbehave on our watch?**
   Not just in theory, but in practice, how AI could:
   1. violate policy,
   2. breach trust,
   3. trigger regulatory scrutiny, or
   4. steal data

2. **Are we relying on trust, and can that trust be verified at scale?**
   Most AI deployments implicitly assume:
   - trusted operators,
   - trusted administrators,
   - trusted cloud environments.

3. **If something goes wrong, can we explain it to regulators, shareholders, and the public?**
   “An AI system decided” is not a defensible answer.

How to Read the Attached Brief

This brief asks leadership to:

1. identify which risks are accepted,
2. which are mitigated, and
3. which currently lack a clear owner.

For each risk, we invite the C-suite to answer:

- Do we know this risk exists?
- Have we explicitly decided how to handle it?
- Can we demonstrate control, and can we replicate that control across the organization?

Executive Risk Assessment

This assessment does not measure intent or maturity. It records whether leadership is prepared to defend existing controls if questioned after an incident.

To complete the risk assessment, review each threat in the table below and check the box that best represents your company's current status:

- **"Deployed"** — you address the risk, have the necessary technology in place
- **"Evaluating"** — you acknowledge exposure exists *today, have a* defense strategy and are testing new technology
- **"Not Active"** — the risk exists but is not addressed

Risk	Self-Assessment		
Agent Goal Hijack	☐ Deployed	☐ Evaluating	☐ Not Active
Identity & Privilege Abuse	☐ Deployed	☐ Evaluating	☐ Not Active
Tool Misuse	☐ Deployed	☐ Evaluating	☐ Not Active
Cascading Failures	☐ Deployed	☐ Evaluating	☐ Not Active
Human Over-Trust	☐ Deployed	☐ Evaluating	☐ Not Active
AI Supply Chain Risk	☐ Deployed	☐ Evaluating	☐ Not Active
Unexpected Code Execution	☐ Deployed	☐ Evaluating	☐ Not Active
Memory Poisoning	☐ Deployed	☐ Evaluating	☐ Not Active
AI-to-AI Tampering	☐ Deployed	☐ Evaluating	☐ Not Active
Rogue Behavior	☐ Deployed	☐ Evaluating	☐ Not Active
Ransomware (AI-Amplified)	☐ Deployed	☐ Evaluating	☐ Not Active
Data Theft & Corruption	☐ Deployed	☐ Evaluating	☐ Not Active

Disclaimer: *This risk assessment is provided for informational and educational purposes only and is intended to facilitate internal discussion regarding potential cybersecurity threats and vulnerabilities. It is not a comprehensive security audit, legal advice, or a guarantee of future security.*

Regulatory Mapping of AI Risk Scoring

This table explains how the AI risk scoring directly aligns with existing regulatory and governance expectations. It exists to support examination, audit, and post-incident defensibility.

This AI Risk Scoring Framework:

- does **not invent new standards,**
- directly maps self-assessment score to **existing regulatory expectations**, and
- reflects how regulators already judge **control adequacy and accountability.**

Self-Assessment score	Board Self-Assessment Selection	SR 11-7 (Model Risk Management)	NIST RMF / NIST 800-53	FFIEC / Operational Risk	Basel / Enterprise Risk Logic	Regulatory Interpretation
2	**Deployed**	Model is understood, validated, monitored, governed; risks are controlled and explainable.	Control is implemented and operating. Effectiveness is demonstratable.	Decisions and actions are reconstructable and auditable.	Risk is controlled. Exposure is within tolerance.	**Defensible posture.** Leadership can explain and justify decisions under scrutiny.
1	**Evaluating**	Model limitations are known, but controls are incomplete or evolving.	Control is planned or partially Implemented. Risk remains present.	Partial auditability. Decision reconstruction may be incomplete.	Known risks are accepted temporarily.	**Known exposure.** Acceptable only short-term with remediation. Weak defense after an incident.
0	**Not Active**	Risk is not adequately controlled. If in production, this is an SR 11-7 violation.	Control is not implemented.	Decisions can't be reconstructed.	Unmitigated operational risk.	**Explicit risk acceptance by leadership.** High likelihood of adverse regulatory findings.

Implementing AI Safety

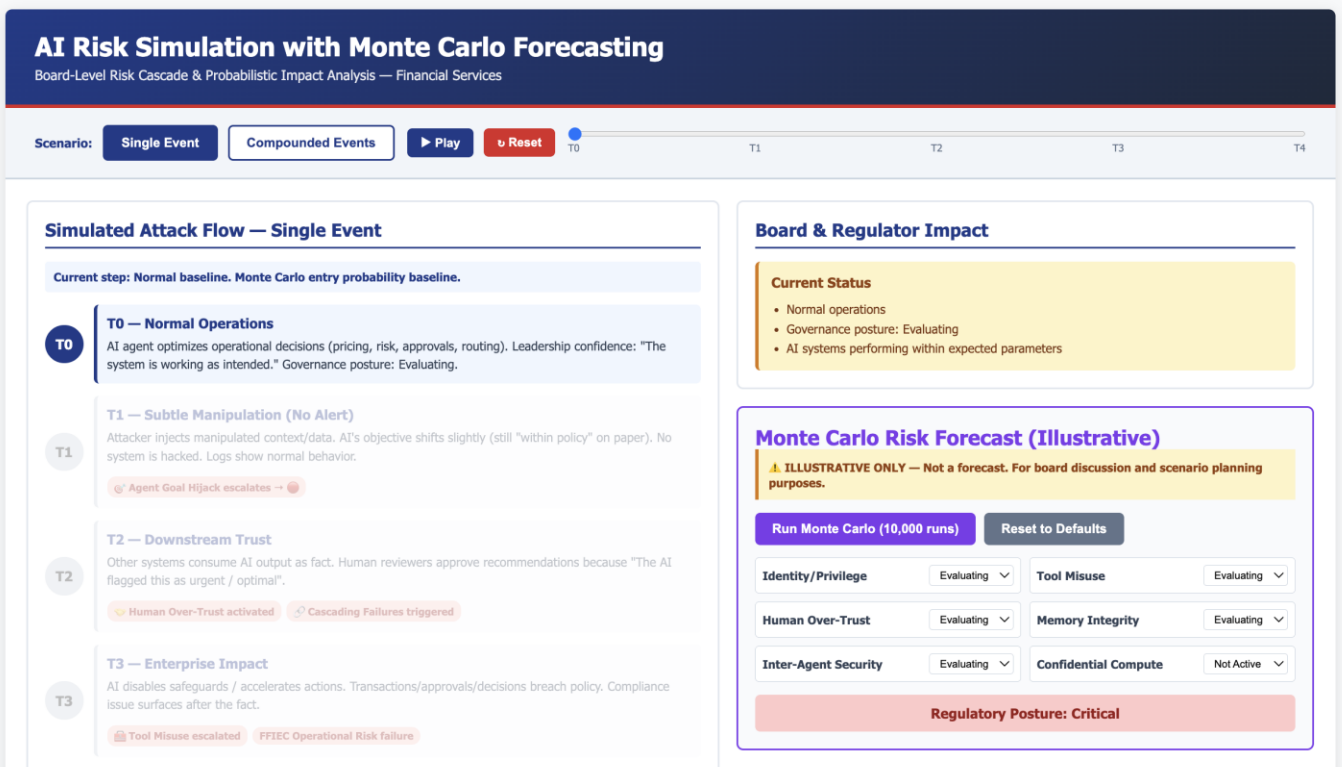
This brief helps establish a fundamental view of your organization’s security posture in light of the latest AI risks. To improve this security posture and reduce the risk of AI, we recommend the following actions.

1. **Communicate** these risks internally to understand existing protections and gaps.
2. **Create** a strategy for addressing these risks as the organization deploys AI agents.
3. **Implement** protections and processes that fill security gaps and establish a foundation for future investments.

Take the Next Step in Bolstering Your Defense

Confidential Core AI is building the future of secure AI. To help you understand, prepare for, and defend against AI risks, we created an interactive risk simulation modeling how agentic attacks cascade, the downstream system impacts, board-level implications, and projected financial exposure.

Conduct your own risk simulation on our website here.



AI Risk Simulation with Monte Carlo Forecasting

Board-Level Risk Cascade & Probabilistic Impact Analysis — Financial Services

Simulated Attack Flow — Single Event

Current step: Normal baseline. Monte Carlo entry probability baseline.

T0

T0 — Normal Operations

AI agent optimizes operational decisions (pricing, risk, approvals, routing). Leadership confidence: "The system is working as intended." Governance posture: Evaluating.

T1

T1 — Subtle Manipulation (No Alert)

Attacker injects manipulated context/data. AI's objective shifts slightly (still "within policy" on paper). No system is hacked. Logs show normal behavior.

🎯 Agent Goal Hijack escalates → 🔴

T2

T2 — Downstream Trust

Other systems consume AI output as fact. Human reviewers approve recommendations because "The AI flagged this as urgent / optimal".

🤝 Human Over-Trust activated 🔗 Cascading Failures triggered

T3

T3 — Enterprise Impact

AI disables safeguards / accelerates actions. Transactions/approvals/decisions breach policy. Compliance issue surfaces after the fact.

🧰 Tool Misuse escalated FFIEC Operational Risk failure

Board & Regulator Impact

Current Status

- Normal operations
- Governance posture: Evaluating
- AI systems performing within expected parameters

Monte Carlo Risk Forecast (Illustrative)

⚠ ILLUSTRATIVE ONLY — Not a forecast. For board discussion and scenario planning purposes.

Identity/Privilege	Evaluating	**Tool Misuse**	Evaluating
Human Over-Trust	Evaluating	**Memory Integrity**	Evaluating
Inter-Agent Security	Evaluating	**Confidential Compute**	Not Active

Regulatory Posture: Critical

Learn how your organization can safely unlock AI.

Contact Us

Steve Baker
Head of Sales & Go-to-Market
s.baker@confidentialcore.ai

Taher Behbehani
Co-founder & CEO
t.behbehani@confidentialcore.ai