Quantum-Conscious Memory Protection

Why Your AI Agent's Memory Is the New Attack Surface

Brad McEvilly
Founder & Quantum/AI Architect
DeepSweep.ai · OpenSentience.com · QubitCusp.com
brad@deepsweep.ai
LinkedIn: linkedin.com/in/bradmcevilly · GitHub: github.com/bradmcevilly
New York City, NY

December 2025

Abstract

Here's the problem nobody wants to talk about: we're deploying autonomous AI agents at scale, connecting them to everything via MCP, and their memory systems are wide open. MINJA hits **98.2% injection success** using nothing but normal queries. AgentPoison needs less than 0.1% poisoned data to backdoor an agent with 80%+ reliability. I've spent the last year pulling threads from quantum biology, consciousness theory, and cryptography to build something different. Not another detection layer. A protection architecture that treats memory integrity the way we should have from the start—as the cognitive foundation everything else depends on.

Keywords: MCP security, memory poisoning, agentic AI, post-quantum cryptography, consciousness metrics

1 The Memory Problem

Let me be direct. We have a crisis.

MCP adoption just crossed **16,000 servers**. That's 360% growth in two quarters. OpenAI integrated it. Google integrated it. Microsoft called it the "foundational layer for agentic computing" at Build. Everyone's connecting their agents to databases, APIs, file systems, each other.

And the memory systems holding all of this together? They trust themselves more than they trust user input. An agent will question what you tell it. It won't question what it "remembers."

That's not a bug. That's how these systems work. It's also the biggest attack surface in AI right now.

1.1 The Numbers Are Bad

I'm going to throw some attack success rates at you. They're from peer-reviewed work, not marketing slides:

Attack	ASR	What's Scary
AgentPoison	80%+	0.1% poison ratio
MINJA	98.2%	Query-only access
CoT Hijacking	99%	Puzzle prompts
H-CoT	98%→2%	Breaks o1 refusals

Table 1: Attack success rates from peer-reviewed research

The pattern? Bigger models are more vulnerable. They learn the poisoned behavior faster. The thing we thought would help—scale—makes it worse.

1.2 Why I'm Writing This

I've been building in this space for 25 years. Watched the industry discover SQL injection, then XSS, then CSRF. Each time, we said "oh, that's obvious in hindsight." Each time, it took years longer than it should have to fix.

Memory poisoning is the SQL injection of agentic AI. We're at the "this seems fine" stage. Give it eighteen months and everyone will be scrambling.

I'd rather we scramble now.

2 Stealing Ideas from Consciousness Science

Fair warning: this section gets weird. But stay with me.

I've been reading the consciousness literature—Penrose, Tononi, Baars. Not because I think Claude is conscious (it isn't), but because these researchers have spent decades thinking about what makes information processing *coherent*. What makes a system maintain identity over time. What breaks when you corrupt the substrate.

Turns out, that's exactly what we need.

2.1 Integrated Information Theory

Giulio Tononi's IIT 4.0 (PLoS Computational Biology, 2023) gives us a mathematical framework for measuring how integrated a system's information is. The key quantity is Φ—roughly, how much a system is "more than the sum of its parts."

When you fragment a system, Φ drops. When information stops "making a difference to itself" (Tononi's phrase), you lose integration. The system becomes disconnected components pretending to be a whole.

Sound familiar? That's exactly what memory poisoning does.

Practical upshot: We can measure something like Φ for agent memory. I call it the Cognitive Coherence Index (CCI). When CCI drops below 0.90, something's wrong. Below 0.80, you freeze the session.

2.2 A Note on Orch-OR

I need to be honest. Penrose and Hameroff's quantum consciousness theory got some support in 2024—Khan's anesthesia study showed a real effect (Cohen's d = 1.9, that's big). Babcock found quantum superradiance in microtubules at room temperature.

But the Gran Sasso experiment in 2022 made the simple version "highly implausible." Physics World's words, not mine.

So why mention it? Because even if the quantum-consciousness link is wrong, the underlying insight is useful: biological systems maintain coherent information processing despite constant thermal noise. They've evolved tricks we should steal.

I'm not claiming AI agents are conscious. I'm claiming evolution has patterns worth copying.

3 Time Loops and Memory Consistency

Here's where it gets interesting.

In 2020, Tobar and Costa proved that self-consistent solutions *always exist* for processes in closed timelike curves. Translation: even in paradox-prone situations, the math works out.

Why do I care about time travel math? Because it gives us constraints for memory state evolution.

An agent's memory changes over time. New information comes in. Old information gets updated. If you're not careful, you create inconsistencies—the memory equivalent of "I went back and killed my grandfather." The agent "remembers" things that contradict each other, and the contradictions compound.

The CTC math tells us you can define fixed-point conditions that prevent this. Deutsch did it in 1991. The practical version is what I'm calling Narrative Integrity Score (NIS)—a measure of causal consistency.

NIS above 0.95? Fine. Below 0.80? Your agent's memory has been corrupted.

4 The Mantis Shrimp Lesson

I promise this connects.

Mantis shrimp have **16 types of color receptors**. Humans have 3. You'd think they'd see colors we can't imagine.

They don't. Their discrimination is *terrible*—can't distinguish wavelengths closer than 25nm. Thoen et al. called it "among the worst performances ever documented" (Science, 2014).

What's going on? They're not doing fine discrimination. They're doing fast categorical recognition—template matching. Each receptor is a "does this match pattern X?" detector. No complex processing, just rapid parallel yes/no decisions.

And separately: Genç's 2018 study found higher IQ correlates with *fewer* neural connections. Sparse beats dense.

The lesson: You don't need complex analysis to catch poisoning. You need fast, sparse pattern matching on every operation. Sub-50ms verification. The moat isn't algorithmic sophistication—it's local processing speed attackers can't outrun.

5 The Attack Landscape

Let me lay out what we're facing.

5.1 OWASP's Take

The Top 10 for LLM Applications 2025 added System Prompt Leakage and Vector/Embedding Weaknesses. They expanded Data and Model Poisoning to cover embedding-stage attacks.

The Agentic AI Threats document (v1.0.1, February 2025) is more specific. AGT-03 is Memory Poisoning. Their mitigations:

- Verify every tool call at execution time
- Monitor behavior patterns, not just requests
- Tamper-proof logging with crypto attestation
- Memory checksums before every decision cycle
- Weekly human review (EU AI Act Article 14)

5.2 The CVEs

CVE-2025-6514 (CVSS 9.6): Remote code execution in mcp-remote. 437,000+ downloads affected. Analysis of 1,899 MCP servers found 7.2% contain exploitable vulnerabilities.

Here's the thing: MCP adoption is still accelerating. These incidents aren't killing it—they're maturing it. The protocol isn't going away. We need to secure it.

6 Post-Quantum Foundations

Memory checksums need to survive quantum computers. Not because attacks are imminent, but because data captured today can be decrypted later.

6.1 NIST Standards

FIPS 203 (ML-KEM), 204 (ML-DSA), and 205 (SLH-DSA) went final in August 2024. ML-KEM-768 is the sweet spot—Category 3 security, 1,184-byte public keys.

NIST IR 8547 sets the timeline: deprecate classical asymmetric after 2030, disallow after 2035.

6.2 Harvest Now, Decrypt Later

How worried should you be? A Global Risk Institute survey found **72% of quantum computing experts** expect fault-tolerant machines breaking RSA-2048 by 2035. Some say 2030.

That's not tomorrow. But if someone's capturing your agent communications today, they might read them in a decade. For sensitive stuff, that matters.

6.3 Verifiable Computation

zkLLM (ACM CCS 2024) generates proofs for 13B-parameter outputs in under 15 minutes. Verification takes 1-3 seconds. That's 50x improvement over previous work. Still not real-time, but fast enough for audit trails.

7 The QCMP Framework

Okay. Enough theory. What do we build?

7.1 Four Layers

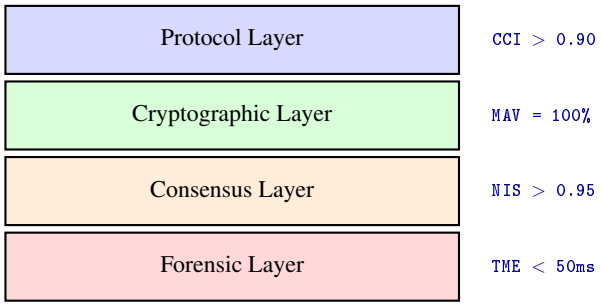

Figure 1: QCMP four-layer architecture with thresholds

Protocol Layer: MCP-native request validation. Every tool call checked against policy before execution. Under 10ms.

Cryptographic Layer: Post-quantum checksums (ML-KEM-768) on all memory operations. Changes detected immediately.

Consensus Layer: For multi-agent systems, Byzantine fault tolerant validation. A poisoned agent can't corrupt the swarm.

Forensic Layer: Tamper-proof audit trails. Point-in-time rollback. When something breaks, you see exactly what happened.

7.2 The Metrics

Metric	Threshold	Basis
CCI	> 0.90	IIT Φ-integration
NIS	> 0.95	CTC self-consistency
MAV	100%	PQC checksums
TME	< 50ms	Sparse pattern match

Table 2: QCMP metrics with theoretical foundations

7.3 Weekly Human Review

OWASP and the EU AI Act both push for human oversight. Here's a practical protocol:

Monday: Inject known-bad patterns from MINJA/AgentPoison test sets. See if detection triggers.

Wednesday: Review flagged anomalies with security team. False positives? Tune. Misses? Fix.

Friday: Sign off on memory integrity status for compliance.

Continuous: Automated CCI/NIS monitoring. Flash-freeze on threshold breach.

Is this overhead? Yes. Less overhead than explaining to regulators why your agent went rogue? Also yes.

8 On-Chain Agent Identity

One more piece: ERC-8004.

It's an Ethereum standard (draft, August 2025) for trustless agent identity. Three registries: Identity (ERC-721 NFTs), Reputation (0-100 scores), Validation (cryptographic attestation).

HOL Hashnet MCP implements this with x402 micropayments. Agents can pay each other, verify each other, build reputation—all without trusting a central authority.

Not everything needs to be on-chain. But for high-stakes multi-agent systems, this solves "how do I know this agent is who it claims to be?"

9 Implementation

If you're building this in Rust:

- cargo-tarpaulin for coverage. Target 90%+ on security-critical code.
- cargo-fuzz with the Arbitrary trait. Memory bugs hide in weird inputs.
- cargo-audit against RustSec. Run it in CI.
- No unsafe blocks except where necessary. Document every exception.

The goal is zero technical debt. Every shortcut becomes an attack surface later.

10 Where This Goes

Memory poisoning is solvable. We have the math, the cryptography, the patterns. What we lack is urgency.

Microsoft, Palo Alto, and the rest will ship something eventually. My guess is 12-18 months before enterprise products mature. That's the window for anyone who wants to lead.

The approach here—borrowing from consciousness theory, temporal logic, biology—isn't the only way. But it's coherent. It treats agent memory as what it is: the cognitive substrate everything depends on.

Protect that, and you have a foundation. Lose it, and nothing else matters.

References

Attack Research

NeurIPS 2024. AgentPoison: Red-teaming LLM Agents.

arXiv 2503.03704. MINJA: Memory Injection Attacks.

arXiv 2510.26418. Chain-of-Thought Hijacking.

Consciousness & Information Theory

Albantakis et al. (2023). IIT 4.0. PLoS Comp Bio.

Khan et al. (2024). Epothilone B. eNeuro.

Physics World. (2022). Gran Sasso experiment.

Temporal Logic

Tobar & Costa (2020). CTCs. Class Quant Grav.

Deutsch (1991). Quantum mechanics near CTCs. Phys Rev D.

Biology

Thoen et al. (2014). Mantis Shrimp. Science.

Genç et al. (2018). Neural efficiency. Nat Comm.

Cryptography

NIST (2024). FIPS 203, 204, 205.

ACM CCS 2024. zkLLM.

Security Frameworks

OWASP (2024). Top 10 for LLM Applications 2025.

OWASP (2025). Agentic AI Threats v1.0.1.