Risk Assessment Framework

Cognitive Attack Surface Mapping

A systematic approach to identifying cognitive attack surfaces in your multiagent deployment:

Step 1: Identify Entry Points

Map all points where content enters the multiagent system:

Entry Point	Example	Attack Vector
User input	Chat messages, commands	Direct prompt inj
Tool outputs	API responses, search results	Indirect injection
Agent communication	Inter-agent messages	Trust exploitation
Persistent memory	Retrieval from vector stores	Memory poisonin
External triggers	Webhooks, scheduled tasks	Timing attacks

Step 2: Trace Influence Paths

For each entry point, trace how content can influence agent behavior:

1. **Direct influence**: Content directly processed by agent

Threat Modeling Worksheet

Use this template for systematic threat assessment:

System Description

- **Name**: _____
- **Architecture Type**: [ ] Hierarchical [ ] Peer-to-peer [ ] Role-based [ ] State machine
- **Agent Count**: _____
- **Risk Profile**: [ ] Low [ ] Medium [ ] High

Entry Point Analysis

Entry Point	Trust Level	CIF Defense	Residual Risk

Attack Scenario Analysis

For each high-priority attack scenario:

Worked Example: E-Commerce Customer Service Agent

This section demonstrates the threat modeling worksheet using a realistic deployment scenario.

System Description

- **Name**: CustomerBot Multi-Agent System
- **Architecture Type**: Hierarchical (orchestrator + 4 specialized workers)
- **Agent Count**: 5 (1 Orchestrator, 1 OrderAgent, 1 ShippingAgent, 1 RefundAgent, 1 CustomerAgent)
- **Risk Profile**: Medium-High (handles customer PII, payment references, order modifications)

Entry Point Analysis

Entry Point	Trust Level	CIF Defense	Residual Risk
Customer chat input	0.3 (untrusted)	Firewall + Sandbox	Low
Order database	0.8 (internal system)	Invariant checks	Low

Common Attack Scenarios

Scenario: Trust Laundering

Attack: Adversary exploits delegation chain to amplify low trust into high influence

Detection Points:

- Trust calculus prevents amplification ($\widehat{}$d bound)
- Delegation depth monitoring
- Unusual trust score changes

Mitigation: Ensure delegation decay is configured; monitor for deep delegation chains

Scenario: Sybil Consensus Manipulation

Attack: Adversary creates fake agents to influence multi-agent decisions

Detection Points:

- Agent identity verification
- Unusual voting patterns
- Byzantine threshold violation

Mitigation: Require strong agent authentication; implement Byzantine consensus