Deployment Considerations

Risk Profile Assessment

Before configuring cognitive security mechanisms, assess your deployment risk profile:

Low Risk Profile

Characteristics: - Internal-only deployment - Non-sensitive data handling - Human-in-the-loop for all significant actions - Limited inter-agent communication

Recommended Configuration: - Firewall: Standard thresholds (accept: 0.3, reject: 0.7) - Trust decay: Moderate ($= 0.95$) - Consensus: Simple majority for coordination - Monitoring: Daily review sufficient

Medium Risk Profile

Characteristics: - Customer-facing but limited autonomy - Some sensitive data handling - Periodic human oversight - Moderate delegation chains

Recommended Configuration: - Firewall: Tighter thresholds (accept: 0.25, reject: 0.65) - Trust decay: Stricter ($= 0.9$) - Consensus: 2/3 majority with identity verification - Monitoring: Real-time alerts for critical events

Architecture-Specific Guidance

Hierarchical Architectures (Claude Code, AutoGPT)

Characteristics: Central orchestrator delegates to specialized workers

Key Risks: - Orchestrator compromise cascades to all workers - Worker escalation can influence orchestrator - Single point of failure

Mitigations: - Strong orchestrator protection (strictest thresholds) - Bounded upward influence from workers - Orchestrator tripwires for identity canaries - Consider multi-orchestrator redundancy for critical deployments

Peer-to-Peer Architectures (Camel)

Characteristics: Equal-authority agents with lateral communication

Key Risks: - Lateral movement attacks (compromise spreads horizontally) - Sybil attacks (injected fake agents) - Consensus manipulation

Mitigations: - Byzantine consensus for all multi-agent decisions - Strong agent authentication - Network topology monitoring

Scaling Considerations

Agent Count Scaling

Agents	Concerns	Recommendations
2-10	Individual agent security dominates	Standard CIF deployment
10-100	Coordination attacks become viable	Byzantine consensus required
100-1000	Emergent behavior security	Collective monitoring, quorum scaling
1000+	Colonial cognitive security	Stigmergic defense patterns (see Part 1 Appendix)

Latency Budget

CIF introduces overhead. Plan accordingly:

Integration Patterns

Pattern 1: Wrapper Integration

Wrap existing agent framework with CIF layer: - Input: Firewall classification before agent processing - Inter-agent: Trust verification on message passing - Output: Invariant checking before action execution

Pattern 2: Native Integration

Embed CIF into agent architecture: - Agent maintains own belief sandbox - Trust calculus integrated with delegation logic - Tripwires planted during agent initialization

Pattern 3: Sidecar Integration

Run CIF as separate monitoring service: - Asynchronous belief drift detection - Centralized trust matrix management - Aggregated alert dashboard