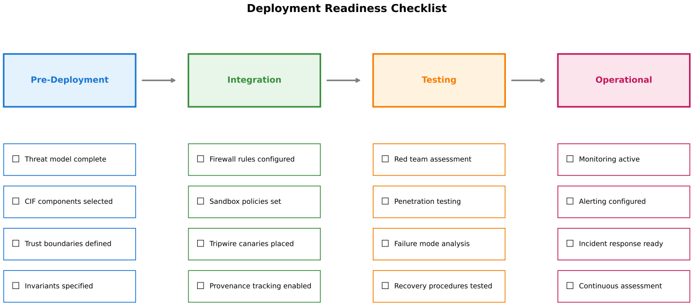
Human-Actionable Checklist

Pre-Deployment Checklist

Before deploying a multiagent system in production, verify the following. Figure 1 provides a visual overview of the deployment phases and their associated verification checkpoints.



Deployment Readiness Checklist

Pre-Deployment	Integration	Testing	Operational
□ Threat model complete	□ Firewall rules configured	□ Red team assessment	□ Monitoring active
□ CIF components selected	□ Sandbox policies set	□ Penetration testing	□ Alerting configured
□ Trust boundaries defined	□ Tripwire canaries placed	□ Failure mode analysis	□ Incident response ready
□ Invariants specified	□ Provenance tracking enabled	□ Recovery procedures tested	□ Continuous assessment

Figure 1: Deployment Readiness Checklist. The cognitive security deployment lifecycle consists of four phases: Pre-Deployment (threat model completion, CIF component selection, trust boundary definition), Integration (firewall configuration, sandbox policies, tripwire placement), Testing (red team assessment, penetration testing, failure mode analysis),

Operational Checklist (Daily/Weekly)

Daily Monitoring

- □ **Review drift alerts**: Check for unusual belief changes
- □ **Verify tripwire integrity**: Confirm canary beliefs unchanged
- □ **Check trust metrics**: Monitor for unexpected trust score changes
- □ **Review failed consensus**: Investigate any Byzantine fault indications

Weekly Review

- □ **Analyze attack patterns**: Review blocked injection attempts
- □ **Audit delegation chains**: Check for unusual delegation patterns
- □ **Verify invariant compliance**: Confirm no invariant violations
- □ **Update threat intel**: Incorporate new attack techniques into defenses

Incident Response Checklist

When a cognitive attack is suspected:

Immediate Actions (First 15 Minutes)

- □ **Preserve evidence**: Capture current cognitive state before any changes
- □ **Assess scope**: Identify which agents and beliefs may be affected
- □ **Contain spread**: Isolate affected agents from propagating beliefs
- □ **Notify stakeholders**: Alert security team and relevant operators

Investigation (First Hour)

- □ **Trace provenance**: Follow belief origins to identify injection point
- □ **Identify attack vector**: Determine how adversarial content entered
- □ **Assess impact**: Evaluate what decisions were influenced
- □ **Check for persistence**: Verify attack doesn’t survive agent

Configuration Quick Reference

Trust Calculus Parameters

Parameter	Recommended Value	When to Adjust
Base weight ( )	0.3	Increase for stable archite
Reputation weight ( )	0.4	Decrease for new deploym
Context weight ( )	0.3	Increase for specialized ta
Decay factor ( )	0.9	Decrease for security-criti

Firewall Thresholds

Threshold	Recommended Value	Risk Trade-off
Accept threshold	0.3	Lower = more strict, more false positives
Reject threshold	0.7	Higher = more permissive, more risk
Quarantine range	0.3-0.7	Narrower = faster decisions, less nuance