AI Agent Risk Analysis Guide

When Built-In Controls Are Enough – and When They're Not

Not all AI deployments carry the same risk. This guide helps you determine when MCP 2.0's built-in controls are adequate and when you need additional measures.

THREE STEPS:

1. Identify your AI agent's highest privilege level (what it can do).
2. Identify the most sensitive data it can access.
3. Find your risk zone in the matrix below.

Each zone specifies required controls. Don't deploy until requirements are met.

THE RISK MATRIX

Tool Privilege	Public/Internal Data	Confidential Data	Regulated Data
Read-only	**GREEN** MCP 2.0 Sufficient	**ORANGE** Add monitoring	**ORANGE** Add monitoring
Write	**YELLOW** Add monitoring	**ORANGE** Add controls	**RED** Maximum controls
Execute/admin	**ORANGE** Add controls	**RED** Maximum controls	**RED** Maximum controls

For practical tools to help you evaluate your readiness for MCP 2.0, including an assessment and a decision-making guide, visit readiverse.com/mcp

READIVERSE

GREEN: MCP 2.0 Sufficient

When: Read-only access to public/internal data

Examples: Documentation search, knowledge base queries, internal analytics

Required controls:
- Authorization controls
- Structured schemas with validation
- Human-in-the-loop controls for unclear requests
- Standard logging (90-day retention)

Monitoring: Monthly log reviews, quarterly audits

YELLOW: Add Enhanced Monitoring

When: Write to non-sensitive data

Examples: CRM updates, internal content publishing, workflow automation on non-sensitive data

Required controls (beyond Green):
- Require human confirmation for ALL write operations
- Real-time alerting on unusual patterns
- Weekly security team log reviews
- Clear escalation procedures

Why more is needed: Write operations or sensitive data access requires early warning systems to catch issues before they become incidents.

ORANGE: Add Significant Controls

When: Execute privileges OR read confidential/regulated data OR write to confidential/regulated data

Examples: Infrastructure changes, financial transactions, database modifications, read-only access to confidential business data or regulated customer data

Required controls (beyond Yellow):
- Mandatory approval for every operation before execution
- Human review of AI-generated commands
- Dedicated infrastructure with strict network isolation
- Rate limiting to prevent rapid dangerous operations
- Daily security review of all activities

Human-in-the-loop process:
1. AI proposes action with full details.
2. Authorized user reviews and explicitly approves.
3. System validates user authorization.
4. Action logged before execution.
5. Outcome verified post-execution.

RED: Maximum Controls or Reconsider

When: Admin privileges on any data OR execute on confidential/regulated data

Examples: Permission changes, security controls, autonomous access to highly regulated systems

Critical question first: "Can we achieve this with traditional automation plus human decision-making instead?" If yes, that's often safer and more appropriate.

If proceeding (beyond Orange):
- Multi-person approval (separation of duties)
- Air-gapped environment with minimal connectivity
- 24/7 security monitoring with dedicated oversight
- Executive approval for deployment
- Tested rollback procedures

Alternative approaches:
- AI recommends, humans execute through existing systems.
- Split into multiple lower-privilege agents.
- Defer until additional MCP security features are available.

When Built-In Controls Are Enough – and When They're Not

QUICK EXAMPLES

Customer Service AI

Initial: Read customer records + write CRM + send emails = write + regulated = **RED**

Risk reduction:
- Split into Agent A (read-only analysis) → **YELLOW**
- Agent B (draft emails with approval) → **ORANGE**

Result: Two safer agents instead of one high-risk agent

Infrastructure Monitoring

Initial: Monitor + restart + modify configs = execute/admin + internal = **RED**

Risk reduction:
- Monitoring only (read-only) → **GREEN**
- Restart with approval (execute) → **ORANGE**
- Config changes: Human-driven, AI recommends

Result: Tiered approach, avoids Red Zone entirely

DEPLOYMENT CHECKLIST

Before deploying any AI agent:

- ☐ Risk zone identified using matrix
- ☐ All required controls implemented
- ☐ Monitoring and alerting operational
- ☐ Incident response procedures ready
- ☐ Rollback capability tested
- ☐ Appropriate approval obtained

WHEN TO REASSESS

- AI agent gains new tools or data access
- Tool privilege levels change
- Data classification updates
- Business use case expands
- After any AI-related incident
- Quarterly (minimum for Yellow/Orange/Red)

For practical tools to help you evaluate your readiness for MCP 2.0, including an assessment visit readiverse.com/mcp