RESEARCH REPORT

The AI Agent Maturity Model

A Framework for Measuring How Ready Your Organization Actually Is

Florian Ziesche — Ainary Ventures

February 2026

CONTENTS

Executive Summary

Read this in 30 seconds. Decide if the next 20 minutes are worth your time.

No existing AI maturity model accounts for what makes agents different from traditional AI. Gartner, McKinsey, Deloitte, Microsoft, and IBM all measure AI-as-tool. Agents aren't tools — they're actors that make decisions on your behalf. Every major framework has a blind spot the size of the entire agentic paradigm.

62% of enterprises experiment with AI agents, but fewer than 10% deploy them enterprise-wide, and only 6% see meaningful EBIT impact.[1][2] The gap between experimentation and production is where most organizations live — and die.

The AGENT framework introduces 5 measurable dimensions — Autonomy, Governance, Error Handling, Networked Trust, and Team Integration — across 5 maturity levels. It's designed so a CTO can place their organization in under 5 minutes using 10 binary questions.

Level 3 ("Calibrated") is the survival threshold for 2026. EU AI Act enforcement begins August 2026. Organizations below Level 3 face regulatory exposure, and the compliance cost is $2-5M — but the cost of a single uncalibrated agent catastrophe will exceed **$100M**.[13]

The model is a hypothesis, not gospel. It's built on patterns from CMMI and DORA applied to agent-specific research across 22 sources. It has not been empirically validated across enterprises. Use it as a starting diagnostic, not a certification.

Methodology

This framework synthesizes two categories of input: (1) a systematic review of 6 existing AI maturity models (Gartner, McKinsey, Deloitte, Microsoft, Google Cloud, IBM) to identify what they measure and what they miss, and (2) 15 research briefs on agent-specific phenomena — overconfidence calibration, adversarial attacks on multi-agent systems, memory poisoning, human-in-the-loop failure modes, non-human identity management, and regulatory convergence — totaling 22 primary and secondary sources. The maturity model structure draws on design principles from two proven precedents: CMMI (Carnegie Mellon, 1987–present) and DORA (Google, 2014–present), specifically their emphasis on outcome-based measurement, prescriptive levels, and self-assessability. All quantitative claims carry explicit source attribution and confidence ratings (High/Medium/Low). The model itself is a proposed framework — not an empirically validated assessment tool. It should be treated as a structured hypothesis about what agent readiness looks like, to be tested against real organizational data.

⚠ Section 1: The Maturity Illusion (Confidence: High)

Here's the picture: 62% of enterprises are experimenting with AI agents.[2] Gartner projects 40% of enterprise applications will incorporate agentic AI by end of 2026.[18] The agent market is forecast to grow from $7.8B to $52B by 2030 — a 45.8% CAGR.[21]

Now here's the other picture: fewer than 10% of those experimenting organizations have deployed agents enterprise-wide.[2] Only **6%** of enterprises qualify as "AI High Performers" with measurable EBIT impact, according to McKinsey's survey of 1,993 organizations.[1] Only 54% of AI projects make it from pilot to production.[4] And Gartner predicts more than 40% of agentic AI projects will be abandoned by 2027.[3]

Evidence: These numbers come from large-sample surveys (McKinsey n=1,993, Gartner enterprise data). The 6% figure is particularly robust — McKinsey defines "High Performer" as organizations attributing ≥5% of EBIT to AI, which is a measurable threshold, not self-assessment.[1]

Interpretation: The gap between experimentation rates (62%) and production deployment (<10%) suggests a structural problem, not a timing problem. Organizations aren't slowly moving up a maturity curve — they're stuck.

I believe the core issue is that every existing AI maturity model measures the wrong thing. They ask: "How well do you USE AI?" The right question for agents is: "How well do you GOVERN actors that make decisions on your behalf?"

That distinction — tool versus actor — is why organizations think they're further along than they are. If you measure yourself against a tool-use framework, having ChatGPT Enterprise and a few LangChain workflows puts you at Level 3. If you measure yourself against an actor-governance framework, those same deployments are Level 1.

> **What would invalidate this?** *If the 62% experimentation rate includes organizations with robust governance frameworks that simply haven't scaled yet (i.e., the bottleneck is business case, not maturity), then the "stuck at Level 1" thesis overstates the problem. I don't see evidence of this — McKinsey's data shows high performers are differentiated by workflow redesign (55% vs 20%), not by governance maturity — but it's possible.*[1]

> **So What?** *If you're a CTO reading this, the question isn't whether you're "doing AI agents." It's whether you could answer, right now: How many agents does your organization run? What was their*

error rate last month? What happens when one fails? If you can't answer those questions, you're at Level 1 — regardless of your AI budget.

Section 2: Why Existing Models Fail for Agents (Confidence: High)

I reviewed 6 major AI maturity models. Here's what each measures — and what each misses.

Gartner AI Maturity Model uses 4 levels (Awareness → Active → Operational → Transformational) focused on organizational readiness, governance, and strategy alignment. It's the most comprehensive for traditional AI. But it contains zero dimensions for agent autonomy, inter-agent trust, memory architecture, or orchestration patterns.[14]

McKinsey doesn't publish a formal maturity model, but their annual State of AI report effectively creates a 3-tier classification: Experimenters → Scalers → High Performers. The insight is valuable — high performers redesign workflows, not just deploy tools. But "agents" as a distinct capability category? Absent.[1]

Deloitte's "State of AI in the Enterprise" (7th edition, 2024) uses Pathseekers → Explorers → Practitioners → Seasoned. Most enterprises cluster at "Explorers." Focus: ROI measurement, organizational change, talent gaps. Agent-specific dimensions: none.[19]

Microsoft AI Maturity Assessment has 5 levels (Foundational → Approaching → Aspirational → Mature → Transformational), focused on cloud infrastructure, data platform, and organizational culture. No scoring for orchestration, trust, memory, or autonomy.[20]

Google Cloud AI Maturity Scale uses 3 phases (Tactical → Strategic → Transformational). Google has explicitly named "Agent Trust" as a key 2026 theme — but hasn't updated their maturity framework to include it.[26]

IBM AI Ladder has 4 rungs (Collect → Organize → Analyze → Infuse). This is a data maturity model wearing AI clothes. Agent paradigm: entirely absent.[22]

The Common Blind Spot

Every model above shares the same assumption: AI is a capability you add to existing workflows. Agents break this assumption. An agent isn't a better Excel formula — it's a new employee who never sleeps, has perfect recall of whatever you put in its memory (verified or not), and makes decisions at machine speed with no natural pause for judgment.

The right comparison isn't IT maturity. It's closer to HR maturity — how well do you onboard, credential, monitor, evaluate, and (when necessary) terminate autonomous actors?

What CMMI and DORA Got Right

CMMI (Capability Maturity Model Integration), developed at Carnegie Mellon starting in 1987, succeeded because it was prescriptive — each level has specific process areas with concrete goals and practices — and because adoption was enforced through U.S. government contracts.[16]

DORA (DevOps Research and Assessment), now part of Google, succeeded because it measured outcomes, not inputs. Four numbers. Four tiers (Low → Elite). Annual benchmarking reports create competitive pressure.[17]

The lesson: a maturity model works when it's prescriptive (CMMI) and outcome-based (DORA). Most AI maturity models are neither — they're descriptive and input-based. "Do you have a data strategy?" is an input question. "What's your agent error rate?" is an outcome question.

> **What would invalidate this?** *If Gartner or McKinsey publishes an agent-specific maturity model before this framework gains traction, the novelty claim weakens. As of February 2026, none has. But the consulting industry moves fast when there's enterprise demand.*

> **So What?** *If you're benchmarking your organization's AI maturity against any of these 6 frameworks, you're measuring the wrong thing. You're measuring how well you use AI as a tool. You need to measure how well you govern AI as an actor. That's what the AGENT framework does.*

Section 3: The AGENT Framework — 5 Dimensions (Confidence: Medium)

The 5 dimensions of the AGENT framework aren't arbitrary. Each maps to a documented failure mode in agentic AI systems.

A — Autonomy

What it measures: How independently do your agents operate? What decisions can they make without human approval? How are autonomy boundaries defined and enforced?

Why it matters: The spectrum from "chatbot with tools" to "autonomous actor" is where most organizations lose track. Without explicit autonomy boundaries, agents either do too little (expensive human bottleneck) or too much (uncontrolled risk).

G — Governance

What it measures: Policies, audit trails, compliance posture, agent lifecycle management. Can you explain to a regulator what your agents did last Tuesday?

Why it matters: EU AI Act enforcement begins August 2026. Maximum penalties: €35M or 7% of global revenue.[13] Only 10% of organizations have a non-human identity strategy.[14]

E — Error Handling

What it measures: Confidence calibration, failure recovery, graceful degradation. Do your agents know what they don't know?

Why it matters: **84%** of LLMs are overconfident across 351 tested scenarios.[5] Verbalized confidence expressions (VCE) are systematically biased.[6] If your error handling relies on agents self-reporting uncertainty, it doesn't work.

N — Networked Trust

What it measures: Inter-agent trust, identity management, orchestration integrity. When Agent A delegates to Agent B, does anyone verify the handoff?

Why it matters: Multi-agent system hijacking succeeds 45-64% of the time in research settings.[9] Memory injection attacks succeed at rates above 95%.[10] All 12 tested prompt injection defenses have

been broken.[11]

T — Team Integration

What it measures: Human-agent collaboration design. Is human-in-the-loop effective, or is it theater?

Why it matters: 67% of security alerts are already ignored due to alert fatigue.[8] In healthcare, false positive rates range from 80-99%.[23] Adding "a human reviews it" doesn't make it safe — it makes it slow AND unsafe if the human is overwhelmed.

How the Dimensions Interact

These 5 dimensions aren't independent. Weak Governance undermines Networked Trust — you can't establish inter-agent trust without identity management. Poor Error Handling makes Team Integration impossible — humans can't meaningfully review outputs if they don't know the agent's confidence level. Low Autonomy can mask problems in every other dimension — if agents don't do much, you never discover your governance, error handling, trust, and collaboration gaps.

This is why the model is progressive. You can't skip levels. Level 3's calibration requirements depend on Level 2's observability infrastructure, which depends on Level 1's basic awareness of what agents exist.

> **What would invalidate this?** *If agent-specific failures turn out to be manageable through traditional IT governance frameworks (i.e., existing SOC/SIEM/IAM tools handle agent monitoring adequately), then agent-specific dimensions add complexity without value. Early evidence suggests otherwise — 23% of IT professionals report agent credential leaks through existing systems[15] — but the data is thin.*

> **So What?** *These 5 dimensions are a diagnostic lens, not a compliance checklist. Use them to identify which dimension is your weakest — that's where your next agent failure will come from.*

Section 4: The 5 Levels — From Playground to Organism

(Confidence: Medium-High)

The AGENT Maturity Matrix

LEVEL	AUTONOMY (A)	GOVERNANCE (G)	ERROR HANDLING (E)	NETWORKED TRUST (N)	TEAM INTEGRATION (T)
1: Ad Hoc	Individual use, no boundaries	No policies	No tracking	No inter-agent awareness	No HITL design
2: Managed	Defined use cases	Agent inventory, basic guardrails	Incident tracking	Agent catalog	Basic escalation
3: Calibrated	Confidence-gated autonomy	Identity management, audit trails	Calibrated confidence, SLAs	Credential governance	Designed HITL with severity tiers
4: Orchestrated	Delegation rules, autonomy SLAs per agent class	Cross-agent audit, policy-as-code	Automated degradation, MTTR <4h	Inter-agent trust scoring, anomaly detection MTTR <2h	Escalation chains with measured response times
5: Autonomous	Self-adjusting boundaries, drift detection <5%	Automated compliance, continuous audit	Self-calibration, ECE <1%	Self-healing trust, <15min recovery	Strategic oversight only, <5% human intervention

Level 1: Ad Hoc — "The Playground"

Individual employees experiment with AI agents. No organizational strategy exists. Agents are glorified chatbots with tool access. This is where the vast majority of organizations are — including many that believe otherwise.

Measurable Criteria:

1. Agents used by individuals, not teams — no shared configuration
2. No version control on agent prompts or configurations
3. No observability — you cannot determine what agents did yesterday
4. No defined escalation path from agent to human
5. Agent credentials are personal API keys

Typical Mistakes:

- Treating ChatGPT Enterprise or Claude with tool use as "having an agent strategy"
- Demo-driven adoption: "Look what it can do!" without "What happens when it fails?"
- Logging agent-generated errors as human errors — the agent becomes invisible in your incident data

Example: A development team uses Cursor + Claude for code generation. Each developer has their own system prompts. Nobody tracks hallucination rates or measures code review rejection rates for agent-generated code. When an agent-generated bug hits production, it's logged as a developer bug. The organization has no idea how much of its codebase was agent-generated, what the quality differential is, or how to improve it.

What's Missing to Reach Level 2: Centralized visibility. You need to know what agents exist, what they're doing, and how often they fail — before you can govern any of it.

Level 2: Managed — "The Dashboard"

The organization has visibility into agent activities. Basic guardrails exist. Agents are tracked, but trust is binary — on or off. There's a dashboard. Someone occasionally looks at it.

Measurable Criteria:

1. Centralized agent observability (logs, traces, cost tracking)
2. Defined use cases: documented list of what agents MAY do
3. Basic input/output validation (guardrails)
4. Agent inventory: you know how many agents run in your organization
5. Incident tracking for agent failures — separate from human error tracking

Typical Mistakes:

- Observability without action: dashboards that nobody checks (94% of production agent developers report using observability tools[24] — but observability ≠ governance)
- Guardrails as afterthought: bolted on after the first incident, not designed into the system
- "We have LangSmith" ≠ "We manage our agents"

Example: A fintech company deploys customer service agents via CrewAI. They track all conversations in Langfuse. They know cost per interaction ($0.35). They have guardrails that prevent the agent from discussing competitor products. But when the agent hallucinates a refund policy — confidently citing a policy that doesn't exist — nobody catches it until a customer screenshots it on Twitter. The observability was there. The governance wasn't.

Positive Counterexample: Klarna deployed customer service agents and reported $60M in annual savings.[25] They had the dashboard. But they also discovered they had "overpivoted" — reducing human staff before the agent reliability warranted it. Even well-resourced, data-driven organizations can get stuck at Level 2 if they optimize for cost instead of trust.

What's Missing to Reach Level 3: Confidence scoring. Your agents need to know — and communicate — how certain they are. And that certainty needs to be calibrated against reality, not self-reported.

Level 3: Calibrated — "The Trust Layer"

Agents produce measurable confidence scores. Outputs are calibrated — when an agent says it's 90% confident, it's right roughly 90% of the time. Human-in-the-loop is designed, not just required. Agent identity is managed with dedicated credentials. This is the minimum viable maturity for 2026.

Measurable Criteria:

1. Confidence scoring on agent outputs (calibrated, not self-reported via VCE)
2. HITL triggers based on confidence thresholds (not random sampling)
3. Agent identity management: dedicated credentials, not personal API keys
4. Defined SLAs for agent reliability (e.g., <2% hallucination rate for production tasks)
5. Memory governance: agents don't accumulate unchecked, unversioned context

Typical Mistakes:

- Using verbalized confidence ("I'm 85% sure") instead of calibrated methods. VCE is "systematically biased" — LLMs are overconfident 84% of the time.[5][6]
- HITL without severity tiering leads directly to alert fatigue. 67% of SOC alerts are already ignored.[8]
- Memory without provenance: the agent "remembers" information nobody verified.
- Setting SLAs without the measurement infrastructure to enforce them.

Example (Hypothetical): An insurance company runs claims processing agents. Each output carries a confidence score from an external calibration method — a consistency-based approach that cross-checks the agent's output via multiple independent queries, costing approximately **$0.005** per check.[7] Claims below 85% confidence auto-escalate to a human reviewer. The reviewer sees context: why the agent is uncertain, what data points conflict, what the agent would have decided. The reviewer isn't asked "is this right?" — they're asked "given this uncertainty, what should we do?" Agent credentials are managed via dedicated service identities. Audit trails are complete. *No public example of a fully Level 3-mature organization exists yet — this scenario illustrates what the standard looks like in practice.*

Evidence vs. Interpretation: The ~$0.005 calibration cost is estimated based on current API pricing for running consistency-based confidence checks.[7] The 84% overconfidence rate is from a peer-reviewed study.[5] The claim that Level 3 is "minimum viable for 2026" is my interpretation — it's based on EU AI Act requirements for human oversight (Article 14), but the Act doesn't reference a specific maturity level.[13]

> **What would invalidate this?** *If VCE calibration improves dramatically (i.e., future LLMs can accurately self-assess confidence), then the case for external calibration weakens. Current research shows the opposite trend — larger models aren't better calibrated — but a breakthrough is possible.*

> **So What?** *Level 3 is where trust becomes measurable. Below Level 3, you're running on vibes — hoping agents are right, assuming humans will catch errors, trusting that credentials won't leak. At Level 3, you have numbers. And numbers let you make decisions about where to expand autonomy, where to add oversight, and where to pull the plug.*

Level 4: Orchestrated — "The Network"

Multiple agents collaborate with defined trust relationships. Delegation, escalation, and inter-agent verification are systematic. Agents can challenge each other's outputs. The organization manages an agent network, not individual agents.

Measurable Criteria:

1. Multi-agent workflows with defined delegation rules and autonomy SLAs per agent class
2. Inter-agent trust scoring: tracked as a rolling accuracy rate over the last 1,000 interactions
3. Automated escalation chains (agent → agent → human) with MTTR <4h for critical-path agents
4. Cross-agent audit trail with end-to-end latency tracking
5. Adversarial testing at least quarterly, with anomaly detection MTTR <2h

Typical Mistakes:

- Orchestration without trust: agents blindly pass outputs to each other. Multi-agent hijacking succeeds 45-64% of the time when trust isn't verified.[9]
- No cross-agent audit trail: you see what each agent did individually, but can't reconstruct the decision chain.
- Memory poisoning across agents: one compromised memory store infects the network. MINJA attacks succeed at rates above 95%.[10]

Example: A logistics company runs a multi-agent system: a demand forecasting agent feeds an inventory optimization agent, which feeds a procurement agent. Each agent's output includes a calibrated confidence score. The procurement agent won't execute purchase orders above $50K unless the forecasting agent's confidence exceeds 90% AND the optimization agent independently confirms. Weekly red-team exercises test for prompt injection, memory poisoning, and delegation manipulation.

> **What would invalidate this?** *If single-agent systems prove sufficient for enterprise use cases (i.e., multi-agent orchestration turns out to be overengineered), then Level 4's criteria are unnecessarily complex. Some practitioners argue well-designed single agents outperform multi-agent systems. The jury is out.*

> **So What?** *Level 4 is where agents become a system, not a collection of tools. It's also where the attack surface expands dramatically. If you're deploying multi-agent workflows without inter-agent trust verification, you're building a chain where every link is a potential point of compromise.*

Level 5: Autonomous — "The Organism"

Agent systems self-monitor, self-improve, and adapt. Human oversight is strategic (setting goals and boundaries), not tactical (reviewing individual decisions). The system learns from failures and automatically adjusts autonomy boundaries.

Measurable Criteria:

1. Agents adjust confidence thresholds based on historical accuracy, with drift alerts when thresholds deviate >5% from target
2. Automatic capability expansion/contraction via rolling 30-day accuracy window with defined triggers
3. Self-healing: MTTR <15min for compromised nodes without human intervention
4. Continuous calibration: Expected Calibration Error (ECE) maintained below 1%
5. Regulatory compliance automated, human intervention rate <5% of decisions

Example: Nobody is here yet. This is the target state for 2027-2028. The closest analog is Waymo's autonomous driving system — continuous learning from real-world data, automated edge case

detection, progressive autonomy expansion based on safety metrics. But even Waymo had a 1,212-unit recall for prediction software. If Waymo — purpose-built for autonomy — still has failure modes, general-purpose AI agent systems aren't reaching Level 5 any time soon.

Reality Check: Level 5 is aspirational. Including it defines the direction of travel and makes clear that premature autonomy — skipping the foundations — is a specific, identifiable failure mode.

> **So What?** *If anyone tells you they're at Level 5, they're either lying or they've redefined "autonomous" to mean something it doesn't. Use this as a filter.*

Section 5: The 5-Minute Self-Assessment

Ten questions. Binary answers. No ambiguity. Answer honestly — nobody's grading you.

#	QUESTION	IF YES →
1	Do you know how many AI agents your organization runs right now?	Level 2
2	Can you tell me the error rate of your agents from last month?	Level 2
3	Do your agents produce confidence scores on their outputs?	Level 3
4	Are those confidence scores calibrated against real outcomes (not self-reported)?	Level 3
5	Do your agents have dedicated credentials (not personal API keys)?	Level 3
6	When Agent A passes output to Agent B, does B independently verify it?	Level 4
7	Do you have a cross-agent audit trail (not just per-agent logs)?	Level 4
8	Do you red-team your agent systems at least quarterly?	Level 4
9	Do your agents automatically adjust their own autonomy based on measured performance?	Level 5
10	Does your agent system detect and recover from failures without human intervention?	Level 5

How to Score

Your level = the highest level where ALL questions for that level are answered "Yes."

If you answered No to questions 1 or 2: **You're at Level 1.** You don't have visibility.

If you answered Yes to 1-2 but No to 3, 4, or 5: **You're at Level 2.** You have a dashboard. You don't have trust.

If you answered Yes to 1-5 but No to 6, 7, or 8: **You're at Level 3.** You have calibration. You don't have orchestration.

If you answered Yes to 1-8 but No to 9 or 10: **You're at Level 4.** You have a network. It's not self-governing.

If you answered Yes to all 10: **You're either at Level 5 or you're not being honest with yourself.** Either way, I'd like to talk to you.

The Honesty Problem

I predict 80%+ of organizations answering honestly will score Level 1.

The reason organizations over-rate themselves is structural, not psychological. Most CTOs equate "we use AI agents" with "we're mature at AI agents." That's like equating "we have a website" with "we're mature at digital transformation." Usage isn't maturity. Governance is maturity.

The self-assessment works only if you treat "Yes" as meaning "we have this in production, it's measured, and I could show you the data." Not "we're working on it" or "we have a plan for this" or "our vendor says they do this."

🔧 Section 6: The Level 1 → Level 3 Playbook (Confidence: Medium)

If the self-assessment put you at Level 1 — welcome to the majority. Here's what to do.

Step 0: Stop Adding False Confidence

Before building anything new, remove sources of false confidence. If your agents produce verbalized confidence ("I'm 87% sure about this") without external calibration, that number is worse than no number — it creates an illusion of reliability. Disable or ignore self-reported confidence until you have calibration infrastructure. This costs nothing and takes a day.

Step 1: Build the Inventory (Level 1 → Level 2, Timeline: 1-3 months)

Do this first:

- Catalog every AI agent in your organization. Include "shadow agents" — the ones employees are running on personal accounts.
- For each agent: what does it do, who deployed it, what credentials does it use, what data can it access?
- Set up centralized logging. Every agent action gets logged. Use LangSmith, Langfuse, or even a shared database — the tool matters less than the consistency.
- Create a separate incident category for agent failures. Don't log them as human errors.
- Define explicit use cases: what are agents allowed to do? Write it down. If it's not on the list, it's not allowed.

Cost: Low — primarily organizational effort. Observability tools: $0-500/month depending on scale.

The hard part: Shadow agents. In most organizations, individual employees are already using AI agents for tasks that never appear in any dashboard. The inventory will be uncomfortable.

Step 2: Add Calibration (Level 2 → Level 3, Timeline: 3-9 months)

Do this second:

- Implement external confidence calibration on agent outputs. Consistency-based methods cost approximately $0.005 per check.[7] For an agent making 10,000 decisions per month, that's $50/month for calibrated trust.
- Design HITL triggers based on confidence thresholds. Not "review 10% randomly" — review everything below the confidence threshold you've set. Tier by severity.

- Migrate agent credentials from personal API keys to dedicated service identities. 90% of organizations haven't done this for agents.[14]
- Set measurable SLAs: hallucination rate, confidence calibration accuracy, escalation response time. Measure weekly.
- Implement memory governance: what goes into agent memory, how is it verified, when does it expire?

Cost: ~$0.005/check for calibration, plus identity management tooling. For context: the VW Cariad debacle resulted in $7.5B in losses.[12] Level 3 maturity costs a fraction of a single governance failure.

Step 3: What to Stop Doing

- **Stop using verbalized confidence as a decision input.** It's systematically biased.[5][6]
- **Stop treating all agent outputs equally.** Without confidence scoring, high-stakes and low-stakes decisions get the same (non-)oversight.
- **Stop random-sampling for human review.** It wastes human attention on outputs that don't need it and misses the ones that do.
- **Stop logging agent errors as human errors.** You can't improve what you can't see.

  What would invalidate this? *If the cost of calibration infrastructure exceeds the cost of agent failures for a given organization, then Level 3 may be over-engineering. For organizations using agents only for internal content generation, Level 2 might be sufficient. The playbook assumes agents are making decisions that affect customers, revenue, or compliance.*

  So What? *The path from Level 1 to Level 3 is not a multi-year transformation program. It's 3-9 months of focused work. The first step — building an inventory — is free. The second step — adding calibration — costs $0.005 per decision. The barrier isn't cost or technology. It's the organizational willingness to admit you're at Level 1.*

Section 7: Why Level 3 Is the Real Goal for 2026 (Confidence: Medium-High)

The Regulatory Reality

EU AI Act enforcement begins August 2026. Article 14 requires human oversight for high-risk AI systems. Article 9 requires risk management systems. The maximum penalty: **€35M or 7% of global annual revenue**, whichever is higher.[13]

Here's the translation into maturity levels:

- **Level 1 organizations** cannot demonstrate human oversight because they don't know what their agents are doing.
- **Level 2 organizations** can show dashboards but cannot demonstrate that human oversight is effective.
- **Level 3 organizations** can demonstrate calibrated confidence scoring, designed HITL triggers, agent identity management, and audit trails — the minimum for Article 14 compliance.

The Insurance Angle

Agent liability insurance is emerging. The question insurers ask mirrors the maturity model: Can you demonstrate what your agents do? Can you show calibration data? Do you have audit trails? Organizations at Level 3 will get better terms — or get insured at all. Organizations at Level 1 are uninsurable.

Why Not Level 5?

Level 5 is a target for 2028+, not 2026. Aiming for Level 5 now is counterproductive — it leads to premature autonomy, which is the single most dangerous failure mode in the model. The right goal for 2026 is Level 3: calibrated, governed, auditable.

The Trilemma

1. **Deploy fast** — competitive pressure says move now
2. **Deploy compliant** — regulatory pressure says move carefully
3. **Don't deploy** — risk avoidance says don't move at all

Level 3 is the resolution. It's the minimum maturity that lets you deploy agents in high-risk categories with regulatory defensibility and measurable trust. Below Level 3, you're choosing between speed (and liability) or safety (and irrelevance).

> **What would invalidate this?** *If EU AI Act enforcement is significantly delayed or watered down, the regulatory urgency for Level 3 diminishes. The Act is law, but enforcement precedents will take time to establish. Organizations gambling on weak enforcement may be right in the short term — but they're building on sand.*

> **So What?** *Level 3 isn't aspirational. It's operational. If you deploy agents that touch customer data, financial decisions, or any high-risk category — and you're below Level 3 — you have a compliance gap with a deadline. August 2026.*

Section 8: Predictions (Confidence: Low-Medium)

I present these as structured bets, not certainties. Each includes what would prove me wrong.

Prediction 1: A >$100M Agent Catastrophe Within 12 Months

Confidence: 55%

An organization will suffer a >$100M loss directly attributable to an AI agent failure. The ingredients exist: 45-64% multi-agent hijacking success rates[9], >95% memory poisoning success rates[10], and expanding agent deployment in financial and healthcare contexts.

What would prove me wrong: If organizations are deploying agents more conservatively than the hype suggests, the attack surface may be smaller than I think.

Prediction 2: Agent Maturity Assessments Become a Procurement Requirement by 2027

Confidence: 65%

Just as SOC 2 became a procurement prerequisite for SaaS vendors, agent maturity assessments will become a requirement for enterprises buying agentic AI systems.

What would prove me wrong: If the market consolidates around 2-3 agent platforms that handle governance internally, the need for external maturity assessments diminishes.

Prediction 3: Level 3 Becomes Table Stakes for Enterprise AI by 2028

Confidence: 60%

Within 3 years, organizations deploying AI agents without calibrated confidence, HITL design, and identity management will be viewed the way organizations without basic cybersecurity are viewed today: negligent.

What would prove me wrong: If foundational models become inherently well-calibrated, Level 3's calibration requirements become redundant.

Prediction 4: The "DORA for Agents" Moment

Confidence: 50%

A standardized benchmarking framework will create the same competitive pressure for agent maturity that DORA created for DevOps. Annual reports. Public benchmarks. CTOs citing their agent maturity level in board presentations.

What would prove me wrong: If agentic AI turns out to be a feature, not a paradigm, then a dedicated maturity framework is solving a problem that doesn't persist.

Conclusion: The Model Is a Mirror

The AGENT framework isn't a certification. It's a mirror.

It reflects where your organization actually is — not where your AI vendor's marketing deck says you are, not where your internal champions hope you are, but where the evidence puts you when you answer 10 honest questions.

Most organizations will look in this mirror and see Level 1. That's uncomfortable. It's also the starting point for everything useful.

The path forward is concrete: build an inventory (free), add observability (weeks), implement calibration ($0.005/check), design HITL that respects human attention, manage agent identities like you manage employee identities. Level 3 in 6-9 months. That's the goal.

The alternative is what Gartner predicts: >40% of agentic AI projects abandoned by 2027.[3] Not because the technology failed. Because the organizations deploying it weren't mature enough to govern what they built.

The model is open. Use it, adapt it, prove it wrong. Just don't ignore the question it asks: **How well do you govern actors that make decisions on your behalf?**

If you can't answer that in 5 minutes, you have your answer.

Appendix A: Claim Register

#	CLAIM	VALUE	SOURCE	CONFIDENCE
1	Only 6% are AI High Performers	6% (n=1,993)	McKinsey 2025	High
2	62% experiment, <10% enterprise-wide	62% / <10%	McKinsey 2025	High
3	>40% agentic projects canceled by 2027	>40%	Gartner 2025	Medium
4	54% pilot→production	54%	Gartner 2024	Medium
5	LLM overconfidence rate	84% (9 models, 351 scenarios)	PMC/12249208	High
6	VCE systematically biased	Confirmed	arXiv:2602.00279	High
7	External calibration cost	~$0.005/check	API pricing estimate	Medium
8	SOC alerts ignored	67%	Vectra 2023 (n=2,000)	High
9	MAS hijacking success	45-64%	arXiv:2503.12188	High
10	MINJA success rate	>95%	arXiv:2503.03704	High
11	All prompt injection defenses broken	12/12	arXiv:2510.09023	High
12	VW Cariad loss	$7.5B	VW public filing	High
13	EU AI Act max penalty	€35M / 7%	Legislative text	High
14	Orgs with NHI strategy	10%	WEF 2025	Medium
15	Agent credential leaks	23%	Okta	Medium

#	CLAIM	VALUE	SOURCE	CONFIDENCE
16	CMMI history	Verified	CMU SEI / ISACA	High
17	DORA metrics	Verified	Google DORA	High
18	40% apps with agents by 2026	40%	Gartner	Medium
19	Deloitte AI maturity	4 tiers	Deloitte 2024	High
20	Microsoft AI maturity	5 levels	Microsoft 2024	High
21	Agent market forecast	$7.8B→$52B	Precedence Research	Medium
22	IBM AI Ladder	4 rungs	IBM 2023	High
23	Healthcare false positive	80-99%	PMC6904899	High
24	94% use observability	94%	LangChain	Medium
25	Klarna $60M saved	$60M	CEO earnings call	High
26	Google Agent Trust theme	Confirmed	Google Cloud Blog	Medium

Appendix B: References

[1] McKinsey & Company, "The State of AI in 2025," McKinsey Global Survey (n=1,993), 2025.

[2] McKinsey & Company, "The State of AI in 2025" — agent experimentation and deployment data, 2025.

[3] Gartner, "Predicts 2025: AI Agents" — projection on agentic project cancellation rates, 2025.

[4] Gartner, "4 Levels of AI Maturity and How to Achieve Them" — pilot-to-production conversion rate, 2024.

[5] PMC/12249208, "Overconfidence in Large Language Models" — study of 9 LLMs across 351 scenarios.

[6] arXiv:2602.00279, "Verbalized Confidence Expressions in LLMs: Calibration and Reliability," January 2026.

[7] Cost estimate based on API pricing for consistency-based confidence calibration (3 SLM calls per check). See also: Vashurin et al., "CoCoA: A Minimum Bayes Risk Framework," ICLR 2026.

[8] Vectra AI, "2023 State of Threat Detection" — survey of 2,000 SOC analysts, 2023.

[9] arXiv:2503.12188, "Hijacking Attacks on Multi-Agent Systems" — 45-64% success rates.

[10] arXiv:2503.03704, "MINJA: Memory Injection Attacks on Multi-Agent Systems" — >95% success rates.

[11] arXiv:2510.09023, Meta, "The Rule of Two: Adversarial Prompt Injection" — 12/12 defenses broken.

[12] Volkswagen AG public filings — Cariad software unit cumulative losses of $7.5B.

[13] European Parliament, "Regulation (EU) 2024/1689 — Artificial Intelligence Act" — Articles 9, 14; penalty Article 99.

[14] World Economic Forum, "Navigating the AI-Cyber Nexus," 2025 — 10% with NHI strategy.

[15] Okta, "The State of Digital Identity" — 23% agent credential leak incidents.

[16] Carnegie Mellon SEI / ISACA CMMI Institute — CMMI V3.0, 1987-2023.

[17] Google DORA Research Program (dora.dev) — 4 metrics, 4 tiers.

[18] Gartner, "Top Strategic Technology Trends 2026" — 40% agent incorporation forecast.

[19] Deloitte, "State of AI in the Enterprise," 7th edition, 2024.

[20] Microsoft AI Maturity Assessment Tool, 2024.

[21] Precedence Research — AI agent market forecast ($7.8B → $52B, 45.8% CAGR).

[22] IBM "AI Ladder" framework, 2023.

[23] PMC6904899 — False positive rates in AI-assisted clinical diagnostics (80-99%), 2019.

[24] LangChain, "State of Agent Engineering" — 94% observability adoption.

[25] Klarna CEO earnings call — $60M annual savings from AI agents.

[26] Google Cloud Blog, 2025-2026 — Agent Trust as key theme.

This report is a proposed framework, not an empirically validated assessment.
It is designed to be tested, challenged, and improved.

Get in touch → florian@ainaryventures.com

florian@ainaryventures.com · ainaryventures.com