IndustryContents.

READER LIBRARY · GOVERNANCE GUIDE

The Calibrated Autonomy Checklist

What to ask before you let an AI agent act on its own

An IndustryContents curated guide

Demos Aren't Deployment Decisions

Most AI adoption decisions get made around a demo, not around a task. Someone watches an agent complete a workflow cleanly a few times, and the conversation shifts from “should we” to “when do we ship it.” The research behind this guide argues that's the wrong sequence.

A 2026 academic study on agentic AI governance, published by Izmir Academy Association and written by researchers at Nevşehir Hacı Bektaş Veli University, Gazi University, and the Turkish Higher Education Quality Council, makes a case that's worth sitting with: adoption of an AI agent isn't a technical decision. It's a trust decision, and trust behaves differently once a system stops assisting and starts acting on its own.

Their term for getting this right is calibrated autonomy. Give a system room to act where the task is clear, the risk is contained, the outcome can be undone, and someone is actually watching. Pull that room back everywhere else. This guide walks through what that looks like in practice, using the source research's own case findings across customer service, energy, healthcare, and enterprise operations.

Four Things to Check Before You Hand Off a Task

Before any task gets handed to an agent instead of a person, the source research points to four conditions worth checking. None of these are yes/no in isolation. They're a combination.

Task definition

Agents perform predictably when the objective, the inputs, and the acceptable range of outputs are all specified in advance. Vague mandates, “handle customer complaints,” “optimize the schedule,” leave the agent filling gaps with its own judgment, and that's where things go sideways first.

Risk exposure

Bounded means: if the agent gets this wrong, how far does the damage travel? A miscategorized support ticket is bounded. An agent with write access to pricing or payroll is not.

Reversibility

Can a human undo this before it compounds? Sending an email is reversible in the sense that you can follow up. Approving a loan, shipping an order, or filing a report often isn't.

Monitoring

Not “is there a dashboard,” but: will a human actually see a problem before it becomes five problems? The research is specific here, calling for what it labels workflow explainability rather than model explainability. That means being able to trace which tools an agent used, what data it pulled, and where a person could have intervened, not just what the final output looked like.

Score a task low on more than one of these and the source research's recommendation is blunt: keep a human in the loop, or don't deploy the agent for that task yet.

Four Industries, Four Different Rules

The chapter grounds its framework in real domain differences rather than treating "AI adoption" as one uniform decision. Four sectors came up with distinct governance demands.

Sector	What the research found mattered most
Customer interaction	Disclosure that the user is talking to a non-human system, and a visible, easy path to escalate to a person
Energy systems	Fairness and affordability checks on automated optimization, not just efficiency gains
Healthcare and clinical decision support	Explicit responsibility structures, professional oversight, and stronger explainability given the stakes
Enterprise multi-agent systems	Workflow transparency, role clarity, audit trails, and clear accountability for agent-driven outputs

The pattern across all four: technical performance was never the limiting factor. In every case, adoption held or stalled based on whether the system was intelligible, contestable, and reversible. A highly accurate agent that nobody can audit or challenge still runs into resistance, and the research treats that resistance as reasonable, not as a training problem to be solved with better messaging.

The Cost of Skipping This

The chapter's framework is conceptual. It doesn't report outcome data of its own, and its authors say as much. So here's what the wider research base shows happens when organizations move to agentic deployment without doing this kind of calibration work first.

MIT's Project NANDA published a report in July 2025 titled The GenAI Divide: State of AI in Business 2025, based on a review of more than 300 public AI deployments, 52 structured interviews, and survey responses from 153 senior leaders. Its headline finding: roughly 95 percent of enterprise generative AI pilots showed no measurable effect on profit and loss, while about 5 percent extracted real, significant value. The researchers were explicit that this is a preliminary, non-peer-reviewed report built on a six-month observation window, not a settled finding, and that its sample may not represent every industry. The pattern they describe fits the calibration argument closely: pilots stalled where tools were dropped into workflows without memory, context, or a clear owner, exactly the conditions the chapter's four-question test is designed to catch before deployment, not after.

Separately, Gartner issued a widely cited prediction in June 2025 that over 40 percent of agentic AI projects will be canceled by the end of 2027, attributing this to escalating costs, unclear business value, and inadequate risk controls. Worth stating plainly: this is a forward-looking forecast from a single Gartner analyst, published as a “strategic planning assumption,” not a measured outcome backed by a disclosed survey or sample. Treat it as an informed prediction, not as data on par with the MIT findings above.

Neither of these studies is about the same book chapter this guide is built on. They're included here because they test the chapter's argument against what's actually happening in the market, and because a guide that only cites one source, however credible, is not the same thing as a guide grounded in the wider evidence base.

Overtrust Is a Risk Too

One distinction from the research is easy to miss and worth calling out directly: too much trust in an agent is treated as its own risk, not just too little.

The researchers describe algorithmic trust and algorithmic aversion as two separate failure modes that can show up in the same organization at the same time. Aversion looks like people quietly working around the agent, double-checking everything it does, or refusing to adopt it even when it's measurably more accurate. Overtrust looks like the opposite: people stop questioning outputs because the system has been right often enough, and stop noticing when it's wrong.

The research's recommendation isn't to push everyone toward confident adoption. It's to build in the conditions, explanation, contestability, and a visible way to interrupt the agent, that let people trust it the right amount for the specific decision in front of them. Skepticism in a high-stakes workflow isn't friction to eliminate. It's often doing its job.

The Checklist

Use this section on its own once you've read the reasoning above. Score each task before it moves from human-run to agent-run.

Task definition

☐ Objective, inputs, and acceptable outputs are written down, not assumed

☐ Edge cases outside the defined scope have a fallback (route to a human, don't guess)

Risk exposure

☐ The blast radius of a mistake is understood and limited

☐ Financial, legal, or safety-critical actions require a second check before execution

Reversibility

☐ The action can be paused, undone, or corrected before downstream effects compound

☐ There's a defined window during which reversal is still possible

Monitoring

☐ You can trace which tools, data, and steps the agent used for any given output

☐ A human actually reviews outputs on a set cadence, not only when something looks wrong

Governance

☐ Someone is accountable by name for this agent's actions, not “the AI team” generally

☐ Affected users or employees have a real path to contest or appeal an agent-driven decision

If a task clears every box, calibrated autonomy supports moving forward. If it doesn't, the gap tells you exactly what to fix before deployment, not just whether to wait.

Sources

This guide draws directly on the chapter “Agentic AI, Trust, and Social Acceptance: Governance for Autonomous Socio-Technical Systems” by Samet Ayık, Hafize Nurgül Durmuş Şenyapar, and Ramazan Bayındır, published in:

Aydın, Ö., Karaarslan, E., & Selim, A. (Eds.). (2026). *Agentic AI in the Age of Autonomy with Concepts and Case Studies* (pp. 25–65). Izmir Academy Association. DOI: 10.5281/zenodo.21461825. Licensed under Creative Commons Attribution 4.0 (CC-BY).

The original chapter is a conceptual, literature-based analysis. It doesn't report new experimental data, and its authors are explicit about that limitation. What follows in this guide is IndustryContents' own interpretation and application of that framework for an operator audience; the four-question test, sector table, and checklist above are our synthesis of the chapter's findings, not a reproduction of its text.

The full chapter and book are open access and worth reading directly for the underlying literature review and citations: izmirakademi.org.

No changes were made to the original research's findings. Any errors in interpretation here are ours, not the original authors'.

Additional citations

Two additional sources are cited in the “What Happens When Calibration Gets Skipped” section, independent of the book chapter above:

Challapally, A., et al. (2025). *The GenAI Divide: State of AI in Business 2025*. MIT Project NANDA. Preliminary report, not peer-reviewed.

Verma, A. (2025, June 25). Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027 [Press release]. Gartner, Inc.