Empower AI

From Client Zero to Governed Autonomy

How Empower AI builds, bounds, and proves digital teammates

Besma Abidi, PhD
Sr. Director, Applied AI & Modernization | Empower AI

Agent programs rarely fail because the model could not reason. They fail because no one decided, in advance and in writing, what the agent was permitted to do when it was wrong.

None of those are model problems.

After many years working in AI across defense, intelligence, civilian, and health missions, the thing I keep coming back to is that the model was almost never the hard part. The hard part is everything around it.

Gartner's June 2025 forecast put more than 40 percent of agentic AI projects on track for cancellation by the end of 2027, naming escalating cost, unclear business value, and inadequate risk controls. None of those are model problems. The same analysis estimated that only about 130 of the thousands of vendors claiming agentic capability were building anything that deserved the label.

So, the useful question in a federal environment is no longer whether an agent can perform a task. It is whether the organization deploying it can say, precisely, what it may touch, how anyone would know if it went wrong, and who is accountable when it does. This paper answers those three.

It is worth saying plainly why this matters. The teams we support are already carrying more than they can absorb: new requirements, security responsibilities, data calls, compliance reviews, backlogs that do not wait for the next hiring cycle. A digital teammate is not there to replace the people doing that work. It is there to give an already-stretched team more capacity and more consistent execution without giving up human oversight. That is why the gates matter as much as the capability does.

The order

The Technology Accelerator Framework comes first and leads to Empower AI serving as Client Zero. It is not a set of components; it is a lifecycle with hard stops in it. A capability moves through four phases, and three formal gates sit inside them. Each gate is a decision someone signs, with an evidence package behind it.

1 · IDENTIFY	2 · PRIORITIZE	3 · DEVELOP 1	3 · DEVELOP 2 (SHADOW)	4 · DEPLOY
Work with the mission to find use cases that are low risk and high value.	Rank against pain points and objectives. Feasibility and risk triage, scoring documented.	Build, train, bias-test. Every model documented.	Parallel operation against real work, under human review and evaluation.	Monitored deployment with human review and telemetry. Drift detection, retraining, ATO maintenance.
—	GATE 1 **AI Fitness Screen**	GATE 2 **Pilot Exit**	GATE 3 **Use Case Readiness**	*No exit gate. Findings return to Identify.*

The first is the one most often skipped.

All three gates test the same conditions: mission fit and whether AI is appropriate for the decision point at all, data availability and permissions, approved tools and access paths, accountable ownership, governance path, technical support, training, telemetry, security and ATO impact, and operational readiness. The first is the one most often skipped. A use case can be entirely feasible and still fail the fitness screen because AI is the wrong instrument for that decision.

The lifecycle is a loop rather than a line: what is learned in deployment returns to identification, which is what makes the next use case better chosen than the last. It is grounded in the NIST AI Risk Management Framework and NIST SP 800-53, aligned to OMB M-25-21, M-25-22, and M-26-04 (the last discharged by bias testing in Develop), and in production at CMS (CERT), GSA (DIGIT), DHA (NOVA), USCIS (STRATUS), and DIU (AIDE).

Risk management is continuous and auditable. It is not a one-time approval event.

Where Client Zero becomes a gate

It cannot clear the readiness gate without having done so.

Client Zero is a principle, and principles are easy to state. What makes it operational is that the framework contains a place where it is enforced. Inside Develop, a capability runs in shadow: parallel operation against real work, with people reviewing what it produces. It cannot clear the readiness gate without having done so.

That discipline has a boundary, and we would rather state it than let a client find it. Shadow running tells us how a capability fails, where integrations break, where a control needs to move earlier, and where users hesitate. It cannot tell us how it will behave inside your authorization boundary, at your data volumes. We are clear that we do not hand it over at that line. Because the framework defines a digital teammate by its role, its boundary, its controls, and its evidence rather than by the stack underneath it, what it produces is platform-agnostic: the capability moves to your environment rather than asking your environment to move to it. We stay through that integration — your identity model, your data, your tooling, your authorization boundary — until it runs where you need it to run. That is where the second half of the picture gets filled in, with your constraints rather than our assumptions.

The capability moves to your environment rather than asking your environment to move to it.

Autonomy is a cell, not a level

Authority asks what the agent may do. Reach asks how far the consequence travels.

Maturity ladders that number autonomy from one to four imply every agent should be climbing, and they collapse two independent questions into one axis. We separate them. Authority asks what the agent may do. Reach asks how far the consequence travels.

Authority ↓ Reach →	One record	One system	Cross-system	Mission-wide
Read / analyze	Default	Default	Logged, scoped	Logged, scoped
Recommend	Default	Default	Named owner	Named owner
Act — reversible	Auto, audited	Auto, audited	Approval gate	Approval gate
Act — irreversible	Approval gate	Two-person rule	Two-person rule	Not permitted

A boundary that is written down is a boundary a person can point to.

Two things follow. The bottom-right cell is marked not permitted, and that is a decision rather than a gap we intend to close: irreversible action at mission-wide reach is not a maturity level we are working toward. And a single agent can hold different cells for different actions, which no single maturity score can express.

There is a human reason to draw it this way. A boundary that is written down is a boundary a person can point to. When people know exactly what a system may do and where their own judgment is required, they stop having to guess at their role in it. Governance that is legible to the people doing the work is what turns adoption into something real rather than something mandated.

What authorizes an action

If confidence is the only thing between an agent and a production system, there is no control.

The gates above are lifecycle decisions that a human signs. What follows is different: it is what stands between an agent and a system at the moment it acts, thousands of times a day, with no one watching.

The rule most often proposed is that the agent acts when its confidence is high enough. This does not survive scrutiny. Model confidence is not calibrated, and token probability is not a proxy for correctness, least of all on the long-tail inputs where an error matters most. If confidence is the only thing between an agent and a production system, there is no control. What holds is deterministic and independently verifiable:

1. Preconditions that must evaluate true before the action is offered, checked outside the model.
2. Typed actions drawn from an allow-list, not free-form calls generated at inference time.
3. Policy evaluated as code at a decision point between intent and action, so it is testable and versioned.
4. Proposing a change and committing it as two separate acts, so someone can see exactly what would happen first.
5. Idempotency and a defined way to undo, declared per action.
6. Hard caps on scope and rate, enforced outside the agent.
7. A second authorized human for anything that cannot be walked back.

Confidence still has a job. It can inform a person, and it can triage what gets looked at first. It cannot authorize anything.

A worked example: where the boundary actually sits

10 million
pages processed

1.5 million
bookmarks

30,000
records

200
pages per minute

0.90
weighted-average F1 across the 25 classes

Our *AI Medical Records Assistant* has run in a CMS production environment since March 2025, classifying OCR-derived medical-record pages into 25 bookmark classes for CERT reviewers. It has processed approximately 10 million pages, producing more than 1.5 million bookmarks across more than 30,000 records, at roughly 200 pages per minute. Weighted-average F1 across the 25 classes is 0.90.

The design decisions matter more than the throughput. The model's scope is deliberately narrow: medical-record pages in, bookmark classifications out. It does not act on its own, does not produce output beyond those classifications, and cannot reach data outside the pages in its defined workflow. There is no second function waiting to be enabled.

Human decision authority is unconditional. Every classification is reviewed by a person before it becomes an official bookmark, not a sampled subset and not only the uncertain cases. Per-bookmark confidence scores appear in the reviewer's application as context for that judgment. On the grid above, this sits at recommend, one record.

Confidence here is a transparency mechanism, not a gate. Nothing is applied on the strength of a score.

That last mechanism is the one worth copying.

Validation ran through every phase rather than as a launch event: model selection against held-out test sets through late 2024, with the validation set never used to train weights; an end-to-end pilot in February 2025 across roughly 10 reviewers and approximately 180 records; then continuous scoring against human-verified results from release onward.

That last mechanism is the one worth copying. Because a person reviews every classification, ground truth is generated as a by-product of the work itself. Per-class F1, precision, recall, and accuracy are computed against human-verified submissions continuously, so evaluation cannot go stale and drift surfaces as a measured decline rather than a user complaint. Drift triggers retraining, and every change is versioned, regression-tested, re-evaluated for bias, and released in a controlled window.

It also keeps an honest eye on the review itself. Human oversight is not free: when a model is right most of the time, approval becomes reflexive, and a reviewer who rubber-stamps does not add oversight so much as launder the decision. Scoring per class rather than in aggregate is what keeps that visible — an override rate that falls in classes we never changed is a signal about the review, not the model.

What federal leaders should be asking any agent vendor.

1	Which cell of the authority and reach grid does this agent occupy, for each action it can take?
2	What authorizes the action, specifically, and does any of it rest on model confidence?
3	Show me the evaluation, the per-class figures, and the date they were last computed.
4	How is an action reversed, by whom, and within what window?
5	What happens when the underlying model version changes, and who pays for revalidation?

Conclusion

The credible agentic AI programs will not be the ones with the most agents.

This is the direction federal AI has to move. Not isolated automation, and not agents counted like trophies, but mission-ready digital teammates that make the people already carrying the work more effective.

The credible agentic AI programs will not be the ones with the most agents. They will be the ones that can answer, for every agent in production, what it is allowed to touch, how anyone would know it went wrong, and who is accountable when it does. The accelerator framework is how we build the capability and engineer its boundary. Shadow running is how we find out what we got wrong and fix it, on our own time. The Agentic WorkForce Delivery Model is how what survives becomes capacity an agency can rely on.

> **The goal is not automation at scale. It is autonomy that can be explained, evidenced, and reversed.**

References: Client Zero series. NIST AI RMF 1.0, NIST AI 600-1, NIST SP 800-53. OMB M-25-21 and M-25-22 (3 April 2025, rescinding M-24-10 and M-24-18); OMB M-26-04 (11 December 2025, under E.O. 14319). Gartner, June 2025, on agentic AI cancellation and agent washing.

Empower AI

Empower AI is a premier provider of AI, IT modernization, cybersecurity, and mission support services to federal civilian, military services and fourth estate customers. Learn more at www.empower.ai.