Designing the Human-AI Enterprise

Delegate AI With Boundaries

Curtiss J. Shupe

04

Delegate AI With Boundaries

How to define the limits, thresholds, and escalation rules that make machine delegation governable.

Autonomy typically doesn't arrive within an enterprise through some type of strategic announcement. It arrives through small operating changes that individually might seem insignificant. A process begins to automatically handle requests. A customer service system begins to draft and send responses under certain conditions. A finance system begins to automatically approve low-value actions based on a threshold. A case queue begins to get reprioritized without anyone manually adjusting each step. Collectively, however, it represents a significant change: the enterprise has now begun to delegate operational action to machine systems.

This shift is already happening at scale. For example, the Microsoft Work Trend Index for 2025 highlights the emergence of human agent teams, where an operating model is emerging with managers increasingly coordinating digital labor. In fact, if we look at the associated business reporting from Microsoft, it is clear that "50 percent of respondents reported their organization is already using AI agents to automate workstreams or business processes for entire teams or functions." At that point, the only question an enterprise will ask is where the AI is allowed to act.

The term Bounded Autonomy is used in this article to refer to the design discipline answering this question. This term is a synthetic term in this series and is not found in the source material. However, the requirement it describes is well-supported. "Once you start to grant AI systems permission to behave within a live workflow, you need to define limits, thresholds, escalation paths, and intervention rights. Delegating without those limits isn't maturity – it's just exposing yourself to risk with a fancy brand."

McKinsey's State of AI 2025, published in March, reaches the same conclusion from the value side. The organizations seeing measurable impact aren't just giving access to the models. They're redesigning the workflows, raising the governance bar, and addressing more risks in the deployment of gen AI. McKinsey also finds that workflow redesign has the strongest relationship to EBIT (Earnings Before Interest and Taxes) impact from gen AI. That's significant. Bounded autonomy is not just a model parameter or a product capability. It's a workflow design choice. It's not just whether the system is able to act. It's where the enterprise will allow it to act, under what constraints, and through what logic for transfer of control.

A critical bounded autonomy model can answer five hard questions. First, what type of action is being delegated? Classification, recommendation, routing, execution, and commitment are not interchangeable. Each has unique operating characteristics.

Curtiss J. Shupe

Second, what are the boundaries of system authority? Scope, financial limits, data domain, time window, risk tolerance, and context are the reality of system authority. Third, what events should cause escalation? Should a loss of confidence, a policy violation, an exception condition, or a financial threshold cause escalation or should the system continue to operate on questionable assumptions? Fourth, who authorized the delegation pattern? Delegation should not simply happen by default or by local convenience. Fifth, what evidence supports the boundaries being effective? If an enterprise can't measure override behavior, escalation rates, exception behavior, and boundary violations, then there is little substance to their autonomy model beyond slideware.

The National Institute of Standards and Technology's AI Risk Management Framework makes this argument an even better operating premise. The AI RMF Core is based on govern, map, measure, and manage, and NIST makes two key points. One is that the govern function is meant to permeate the other functions, not live over on the side as the control function that nobody wants to talk to. The other is that risk management is meant to be continuous throughout the entire AI process, not just at launch and then admired from a distance. NIST also says that processes for human oversight are meant to be defined, evaluated, and documented. This gets at the lazy enterprise tendency to say "we have a human in the loop" without ever explaining what the human can see, decide, or stop.

ISO/IEC 42001(AIMS) takes this same logic from the perspective of the management system. ISO's description of AI governance is that it's something that's established, implemented, maintained, and constantly improved. This is a pretty mundane description, but it's actually more important than it appears. This means that autonomy boundaries aren't just a one-time design artifact. They have owners, they have evaluations, they have actions, and they have redesign. This means that if a threshold is too loose, if an escalation route is ignored, or if a boundary is constantly failing in production, the organization is expected to improve the system, not treat it as a one-time weirdness.

The EU AI Act and the guidelines from the OECD make this concept of oversight even more tangible. Article 14 of the EU AI Act states, "For high-risk AI systems, the oversight shall ensure that a person tasked with it can, where applicable, understand the system's capabilities and limitations, identify irregularities and unexpected performances, avoid overreliance, understand the results of the AI, disregard or override the results, and, where applicable, intervene or stop the system." The guidelines from the OECD for the workplace also seem to be moving in this direction

by emphasizing trustworthy AI, transparency, explainability, accountability, and effective human intervention. The message here is very clear: if the concept of oversight is only present on a piece of paper, it's not really present.

That's why bounded autonomy is the right enterprise frame for delegated machine action. It rejects the fake binary between full manual control and unrestricted automation. Most operational value will sit somewhere in between. The real design job is to define the smallest safe unit of delegated action, place it inside explicit limits, attach escalation logic to edge conditions, and collect evidence that the control model works in practice. That isn't bureaucratic drag. It's the thing that makes automation governable at scale instead of merely exciting in a demo.

So the executive question isn't, "Should we let the AI act?" That's too vague to be useful. The better question is, "What action are we delegating, under what limits, with what escalation triggers, under whose authority, and with what proof?" The organizations that can answer those questions precisely are the ones that will scale machine action without giving up control. Everyone else will learn the expensive way that autonomy without boundaries doesn't create leverage. It creates instability.

Curtiss J. Shupe