

1

Turning the Latency Challenge into Opportunity

What is the true cost of human and organizational latency?

The greatest threat to an established enterprise is not a lack of capital, but the invisible friction of latency. Most organizations won't survive the age of intelligent systems. Will yours?

I AM NOT prone to hyperbole. In the academic world, we are trained to be skeptical of the “next big thing” which we lovingly refer to as “shiny objects.” We have seen crypto, VR, and the metaverse promised as revolutions, only to become niche applications.

But AI is different. AI represents the most significant transformation in business since the Industrial Revolution. It's a step change in the scalability of organizations.

In this chapter, you will:

- **Understand the Historical Parallel:** How AI mechanizes cognition just as the steam engine mechanized physical labor

- **Differentiate AI from the Internet:** Why “Agency” is more powerful than “Access,” and why this recursive revolution is moving at the speed of code
- **Identify the Latency Root:** Recognize why operating on “Human Time” is no longer an option, rather than just blaming a lack of technology or capital
- **Audit Your Latency Stack:** Uncover the eight distinct levels of administrative and biological friction slowing you down
- **Navigate the Reliability Tax:** Push through the Latency J-Curve required to move your team from a human bottleneck to trust-based autonomy
- **Differentiate Technical and Structural Debt:** Understand why your century-old organizational design is now a fatal liability in the era of System Time

The Mechanization of Cognition

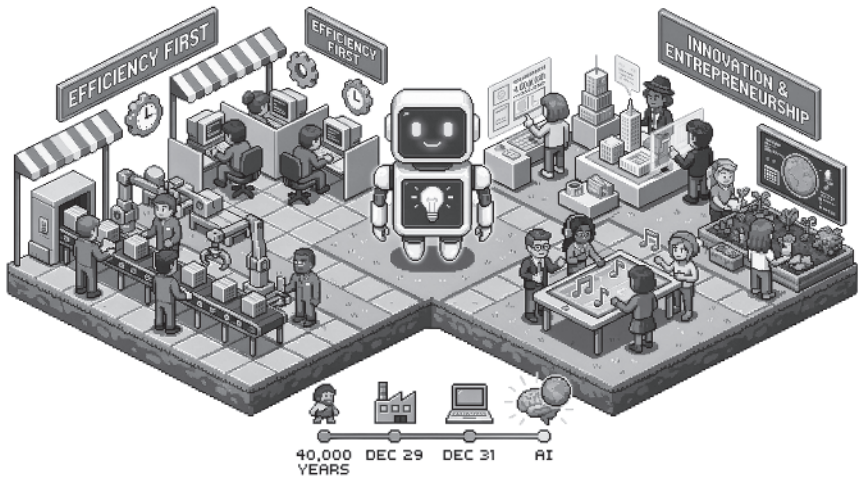
To understand where we’re going, we have to look at where we came from. Before the Industrial Revolution, the limit on production was biology. If you wanted to move a rock, you needed a muscle. If you wanted to move a bigger rock, you needed more muscles (or a horse). The correlation between effort and output was linear.

The steam engine broke that link. Suddenly, a machine could do the work of a hundred horses. It didn’t just make people stronger; it made physical strength irrelevant as a limiting factor.

AI is doing to the brain what the steam engine did to the muscle.

For the entirety of human history, “cognition” was a biological constraint. If you wanted to solve a complex math problem, you needed a human brain. If you wanted to write a contract, you needed a human lawyer. If you wanted to write code, you needed a human engineer.

AI mechanizes cognition. It enables machines not only to execute commands but to think, reason, and collaborate.



I was struck by how Aneesh Raman, chief economic opportunity officer at LinkedIn, thinks about AI in the context of history after sharing the stage with him at the US Chamber of Commerce in Washington, DC, on a misty evening in September 2025. In reflecting on the Industrial Age, Aneesh shares his extremely elegant vantage point:

“Suddenly work for humans became about efficiency above all else. Most of us, in one way or another, started spending our days helping machines and technology do more better faster. Day after day. Week after week. Year after year. On assembly lines in factories. On laptops in offices. Here’s the thing. Efficiency isn’t what our brain is best at. Doing more, better, faster with machines isn’t human potential at its fullest. . . . Compress 40,000 years into a single year. The Industrial Age arrives around December 29th. The internet arrives on December 31st. AI is the last few minutes before midnight.” (Raman 2026)

While countless innovations have emerged over the past century, from the assembly line to the internet, none rival the scale or depth of change we are witnessing right now. Aneesh has so many thoughts on this topic and you can (and should) explore them more deeply in his book *Open to Work*.

Now that the marginal cost of *intelligence* is dropping to near zero, we are going to build entirely new categories of organizations.

The Velocity of Change

There is one major difference between the Industrial Revolution and the AI Revolution: speed. The Industrial Revolution took about 80 years to fully reshape society. It moved at the speed of physical infrastructure.

At the January 2026 WEF annual meeting in Davos, Jamie Dimon said that AI is “changing rapidly.” Dimon continued, “I still think it’s the tip of the iceberg. I think this one [AI] is faster, is massive. Feels like the internet or electricity. This [AI] is not going to roll out over 20 years. It’s more parabolic for now.” (Dimon and Minton Beddoes 2026).

The AI Revolution is moving at the speed of code. ChatGPT reached 100 million users in two months (Satterthwaite 2023). New AI models are rendering old ones obsolete every six months. We are compressing decades of structural change into years.

While the speed of adoption is staggering, there is a deeper, more structural reason for this velocity: AI is the first technology in human history that is actively reinventing itself. In late 2025, the OpenAI Codex team released a technical report that signaled the end of the traditional development cycle: “GPT-5.3-Codex is our first model that was instrumental in creating itself. The Codex team used early versions to debug its own training, manage its own deployment, and diagnose test results and evaluations, our team was blown away by how much Codex was able to accelerate its own development” (OpenAI 2026).

When the tool becomes its own engineer, the rate of progress is no longer limited by human cognitive bandwidth or the 40-hour work week. It is limited only by compute and electricity.

In compressing decades of structural change into years, this technology creates a “whiplash effect” for legacy organizations. You do not have 80 years to adapt. You might not even have 80 weeks. That means each day you delay your competitors are getting a massive head start.

Why AI Is Different from the Internet

The most common mistake I see leaders make is treating AI like “internet 2.0.” They think: “Okay, we adopted the web, we adopted mobile, now we adopt AI. It’s just another channel.” This is fundamentally wrong.

The Internet = Access: The internet connected nodes. It removed the friction of distance. In the internet age, efficiency still relied on human cognition. You had to do the searching, the filtering, and the deciding.

AI = Agency: AI introduces intelligence. You don't call an agent, and you don't click buttons. You tell an agent: "Get me to London for under \$600." The agent does the searching, filtering, and deciding.

The internet made it easier to find information. AI makes it possible to act on it.

Internet Era: Companies built UIs for humans to use.

AI Era: Companies build APIs for agents to use.

The Latency Challenge

In my work at MIT, teaching thousands of entrepreneurs and studying the data at the AI-Driven Enterprise (AIDE) Institute, I have evaluated the "structural delta" between teams of 12 and teams of 1,200. Through that evaluation, I have identified what I believe to be the single biggest threat to established companies. It is not lack of capital. It is not lack of brand awareness.

It is **latency**.

Now, for the first time, software latency is lower than human latency, which is a significant leap.

In present day, you'll find job posts from established organizations that, within the list of responsibilities for the role, reference "scheduling meetings" and "taking meeting notes," both of which were easily handled by off-the-shelf AI tools in 2025. If you've ever had a meeting with me, you'd know that Howie schedules my meetings and the Granola AI app takes my notes. In case you were wondering who that guy was who emailed you about my availability, Howie is, in fact, an AI agent. Reflecting on my reality and the job description I stumbled upon, I worry for that organization's longevity.

This creates a profound, almost jarring dichotomy in our current workforce. We are living in a split-screen reality. On one side, there is the "AI-enabled elite," moving at the speed of light, unburdened by the administrative friction of the 20th century. On the other, there are millions of capable, hard-working professionals still operating in the "note-taking world," trying to build careers on ground that is visibly liquefying beneath them.

It is an uncomfortable moment to witness. It's the executive who views an AI agent as a competitive advantage versus the employee who views that same agent as a replacement for their paycheck. When I see that job description, I don't just worry about that organization's longevity; I feel the weight of the gap between where we are and where we're going. It isn't just about inefficiency anymore. It's about the widening chasm between those who use the tools to transcend the mundane and those who are still being paid to perform it.

The Refined Organizational Design Latency Stack

Latency has long been a critical component of organizational design, but I look at it a little bit differently—from my early career as a software engineer and entrepreneur. When we build systems, we obsess over *latency*, the delay between an input and a response.

When you strip down an organization, it is essentially a network of biological computers (human brains) trying to process information, coordinate, and make decisions. Because humans have inherent biological latencies, it takes time for us to read, comprehend, synthesize, decide, and communicate. Organizational designers have to build structures that accommodate, bypass, or optimize these physical limits. To understand the threat this poses to established companies, we must break down the friction slowing you down.

In a human-centric organization, latency is baked into the system. Building upon the common organizational design forms of latency, I suggest building what I call a Multi-Layered Latency Stack for the AIDE era that includes forms of latency from:

- Organizational design
- Systems engineering
- Data architecture
- Operations management
- Human-computer interaction

A standard concept adapted from systems engineering, this Multi-Layered Latency Stack is an application to organizational design inspired

by Daniel Linder, who provided invaluable feedback as a pre-reader of this book. When we look closely at human-centric organizations, this stack reveals eight distinct layers of organizational drag, building from the structural limits of the enterprise up to the physical limits of the individual:

MULTI-LAYERED LATENCY STACK



Latency Type	Definition	The Human Bottleneck	The AI-Driven Solution
Biological	The physical time required to perceive a stimulus and physically execute a response.	Bound by human processing speed and physical typing/reading limits.	Instantaneous data ingestion and parallel execution at machine speed.
Cognitive	The mental delay caused by subjective reasoning, cognitive biases, and decision fatigue.	Performance degrades over time; prone to emotional or biased decision-making.	Applies consistent, objective logic continuously without fatigue.
Data	The delay between data generation and its availability for use.	Reliant on overnight batch processing or manual, end-of-month reporting cycles.	Continuous, real-time data streaming and instant accessibility.
Operational	Delays in task execution due to manual workflows and siloed systems.	Forced to sequentially switch contexts and manually move data between disjointed software.	Autonomous execution of parallel workflows through integrated APIs.

Latency Type	Definition	The Human Bottleneck	The AI-Driven Solution
Management	The administrative friction of coordinating work among individuals.	Wasted hours waiting for calendar slots, email replies, and status update meetings.	Asynchronous, immediate communication and coordination between autonomous agents.
Structural	Delays inherent in the physical or logical design of the organization.	Slowed by geographic distance, time zones, and repetitive alignment conversations.	Location-agnostic, 24/7 continuous operation across a globally distributed system.
Decision	The lag between data becoming available and the business taking action.	Paralyzed by unclear authority, waiting for analysis, or building consensus.	Automated execution triggered instantly by predefined thresholds and parameters.
Governance	The structural delay required to ensure safety, compliance, and oversight.	Bogged down by slow-moving committees and multi-stage manual review processes.	Continuous, automated compliance auditing, reserving human review strictly for high-stakes exceptions.

The AI-Driven Enterprise model ruthlessly reduces the top seven layers of this stack to buy more time for the eighth: governance. This does not mean we should undermine essential checks and balances. Compliance and safety are nonnegotiable. However, in too many organizations, the “checks” have become the bottleneck, and the “balances” have become an excuse for a total lack of urgency. The key to the AIDE model is realizing that the urgency required to compete in “System Time” is missing in human-centric governance. We must preserve the integrity of our oversight without allowing it to become the friction that kills the firm.

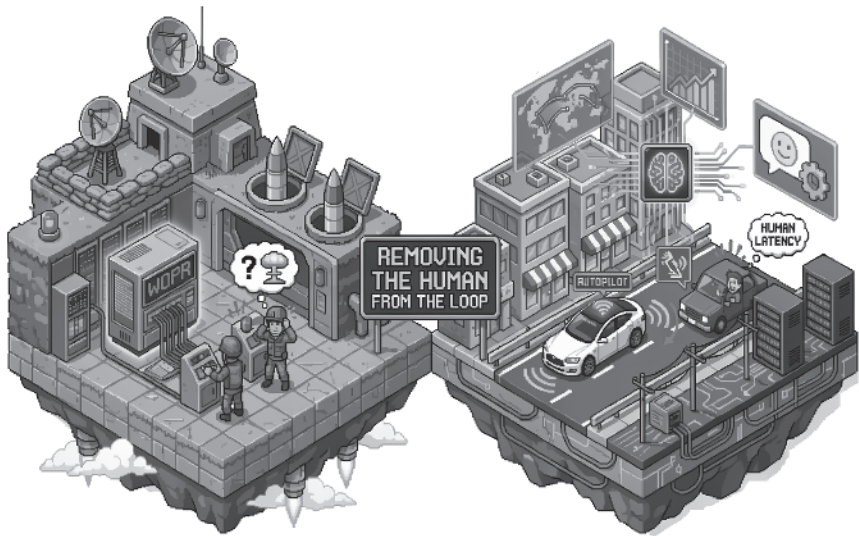
When you require a human to manually review a report, navigate a siloed CRM, or wait for an overnight data pipeline, you are introducing friction into a system that ought to be frictionless.

The Human Bottleneck

With a title like *No One Works Here* this book could have been the story of a company with a bunch of employees who are all slackers, a company where no one *works*: the least efficient company. This book is the opposite. This is a story about the most efficient companies. Those where humans don’t “work” in the traditional sense.

The companies that are most likely to fall behind their peers are the ones that insist on keeping a human in the loop for decisions that a machine could have made yesterday. That doesn’t mean all decisions, just those that might not require a human in the loop in the AIDE era.

It reminds me of the central tension in the 1983 fictional film *WarGames*, which showcases the extreme example of what happens when you propose removing the human from the loop. There is a pivotal scene where military brass debate the implementation of the WOPR (War Operation Plan Response) supercomputer. The argument for the machine was cold and logical: Human operators in the missile silos were a liability because, faced with the gravity of their actions, they hesitated. They couldn’t “turn the key” to launch the nuclear missiles. The proposal was, instead, to route the missile launch codes directly to the WOPR, effectively taking the human out of the loop to ensure instant execution. In the context of nuclear warfare, that logic is terrifying, it could lead to accidental Armageddon.

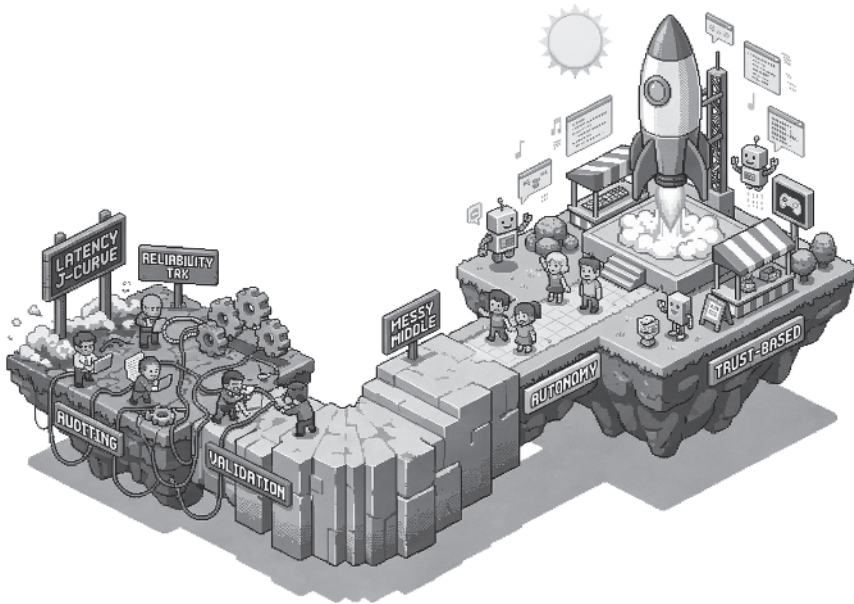


However, we no longer have to look to fiction to see the benefit of removing the human from the loop. We see it every day on our highways. . . . And it saves lives: a modern motor vehicle equipped with an intelligent emergency braking system before a human driver even perceives a hazard or has the time to react. By the time the human brain processes the visual cue of a car stopping two vehicles ahead, a car often obscured from the driver's line of sight, the AI has already calculated the physics of the impending collision and initiated a stop. In this moment, the human's biological latency is a liability; the machine's autonomous agency is the lifesaver.

Stepping back from the WOPR and the emergency braking system examples, in the context of the global economy, the logic is inescapable. In most industries, we are not debating nuclear launch codes; we are debating pricing adjustments, supply chain routing, and customer support queries. In these domains, the "hesitation" of the human operator, the need to schedule a meeting or build consensus, is not a safety feature; it is a bottleneck. It could be an existential threat to the organization.

The Reliability Tax

However, we must acknowledge the “reliability tax.” When you first deploy an agent, latency often spikes. You will spend *more* time auditing the agent’s output than you would have spent doing the work yourself. This is the Latency J-Curve: Things get slower and riskier before they become instant. The winners are the organizations that push through this “messy middle” of validation to reach the other side: trust-based autonomy.



Moving forward, organizations must prioritize human bandwidth for those situations in which the unique capabilities we possess are a necessity and the consequences of delegating to AI, even with guardrails, could be detrimental.

Structural Debt Versus Technical Debt

As software engineers, we talk a lot about “technical debt.” But the companies on the Fortune 500 list, like most established businesses, are suffering from structural debt.

For the last century, we built organizations to solve for the limitations of human communication. We created hierarchies because one person

couldn't manage a thousand people. We created departments to specialize knowledge because one brain couldn't know everything. We created middle management to route information because we didn't have networks.

This structure creates latency. When a legacy company wants to make a decision, the data has to travel up the chain of command. It gets aggregated, summarized, PowerPointed, and debated. By the time the decision is made and the communication comes back down, the market has moved.

The fear of a market shifting before a product can launch used to be a uniquely deep-tech nightmare. That 10-year deep-tech threat has now come for everyone. Today, a digital product can have its core value proposition commoditized mere hours after launch by a single Big Tech release.

No One Works Here: The Digital War Room

You've learned the theory; now let's build the machinery. Turn these concepts into immediate ROI for your company by accessing your free resources at NoOneWorksHere.com.

1. Immediate Actions (Do This Today)

- **Audit Your Latency Stack:** Map a recent delayed project against the eight layers of the Multi-Layered Latency Stack to pinpoint exactly where biological or administrative friction is slowing you down.
- **Pay the Reliability Tax:** Isolate one repetitive administrative workflow, such as meeting scheduling or data routing, and delegate it to an AI agent today. Commit to auditing its output through the “messy middle” of the J-Curve until it reaches trust-based autonomy.

2. The Discussion (Bring This to Your Team)

- Are our current organizational hierarchies solving for human communication limitations that no longer exist? Discuss where our legacy department structures are creating “structural debt” and preventing us from operating at “hours speed.”

(continued)

(continued)

- The chapter warns that organizations often confuse necessary governance with administrative friction, resulting in a dangerous “latency lock-in.” Where in our company are manual review processes acting as an excuse for a lack of urgency rather than a true safety feature?
- What specific workflows in our daily operations treat human hesitation as a safety feature when it is actually an existential liability? Which lower-stakes decisions should we fully delegate to AI to eliminate biological latency?

3. The Step-by-Step Implementation

Stop guessing and start building. Use the **Multi-Layered Latency Stack Workbook** for step-by-step guidance. This digital workbook provides the exact framework you need to implement this chapter’s concepts.

Download it at: NoOneWorksHere.com/latency-stack

4. The Unfair Advantage: AI-Powered Tools

Don’t just read about AI; use it right now. As a reader of this book, you get exclusive access to interactive AI-software tools designed to do the heavy lifting for you.

Access your tools instantly at: NoOneWorksHere.com
(Or simply scan the QR code with your smartphone camera)

