

Leadership First, AI Second

And Other Ways to Avoid the Coordination Tax on AI Productivity

by Matthew Arthurs

Abstract

Insurers have automated the work that is easiest to describe. The recovered hours are going somewhere unexpected and unmonitored, ultimately resulting in a wholly avoidable coordination tax. The antidote? Strategic leadership.

Ask a carrier executive whether AI is working, and you will get a metric: claims triaged per hour. Submissions cleared per underwriter. Documents summarized without a human touching them.

The numbers are real. They are also, in most organizations, the only thing anyone is measuring.

Ask the same executive whether the company is making better decisions than it was two years ago, and the answer arrives more slowly.

That pause illuminates the most important unexamined issue in insurance technology right now: AI adoption is close to universal. Value is not.

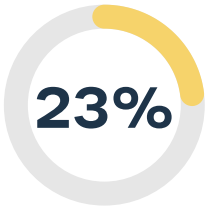
Earnix's fourth annual industry survey of 400 insurance executives finds that 81 percent report embedding AI across most or some of their workflows, with progress concentrated in discrete, function-specific improvements rather than connected enterprise decision-making.[1] AI claims processing sits at 23 percent; policy issuance is at 18 percent.

The standard read on that gap is immaturity—namely that adoption is early but the gap will close with time. I think that explanation misses the larger problem: The gap reflects a design problem.

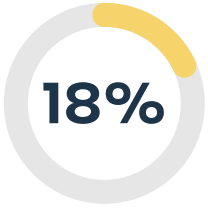
When automation returns hours to an organization, the organization spends those hours on coordination rather than judgment. I have seen this happen inside large, regulated delivery portfolios. It does not look like failure. It looks like diligence. But looks can be deceiving.



Embed AI



AI claims processing



Policy issuance

Originally published in Insights. Shared with permission, all rights reserved.

Concede the Productivity

The gains from AI are real, but they may obscure other truths.

Three randomized controlled trials at Microsoft, Accenture, and an anonymous Fortune 100 company, covering 4,867 software developers, found a 26 percent increase in completed tasks among developers given an AI coding assistant.[2] The study's authors note that the individual experiments were noisy—containing a high level of random, unpredictable variation or error that could obscure actual effects—and that the pooled estimate carries a standard error of roughly 10 percentage points.

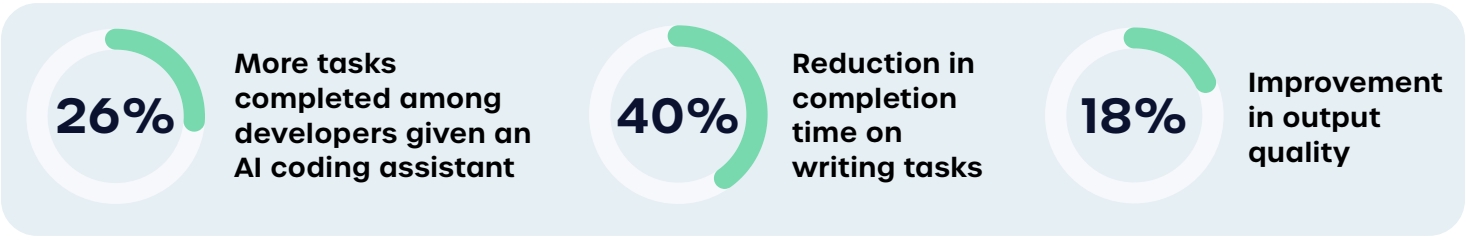
However, an earlier randomized experiment with 453 professionals on writing tasks found a 40 percent reduction in completion time alongside an 18 percent improvement in output quality, scored by independent evaluators.[3]

Those are not vendor numbers. They are controlled experiments, and they point in the same direction.

So the recovered capacity exists. The question nobody is asking with any rigor is where it went.

In a factory, the recovered capacity is visible. The line runs faster, inventory stacks up, and somebody must decide what to do about it. In knowledge work, recovered capacity does not pile up in a corner. It gets absorbed silently by whatever the organization does by default.

In most large carriers, the default is coordination.



Where the Hours Go

Three patterns account for most of the absorption. None is irrational in and of itself, which is exactly why they all survive.

The first pattern is increased verification loops incorporated into workflows. Underwriters receive AI-generated risk summaries and don't simply accept them. Appropriately, given the regulatory exposure and well-documented tendency of these systems to be confidently wrong, they check them.

But the checking is almost never scoped. Absent an explicit standard for what "verified" means, professionals verify to their own comfort level, which, because of the vague parameters, results in verifying nearly everything.

The model does the work. The human revises enough of it that net savings are a fraction of gross savings. Nobody measures the difference, because the gross number is the one that goes in the deck.

Next is review-chain expansion. Faster output invites more reviewers. When the first draft of an underwriting memo or regulatory response takes an hour instead of a day, the marginal cost of adding a reviewer collapses.

So, reviewers get added. Legal wants a look. Compliance wants a look. A second business sponsor wants a look—all because there is now time for one.

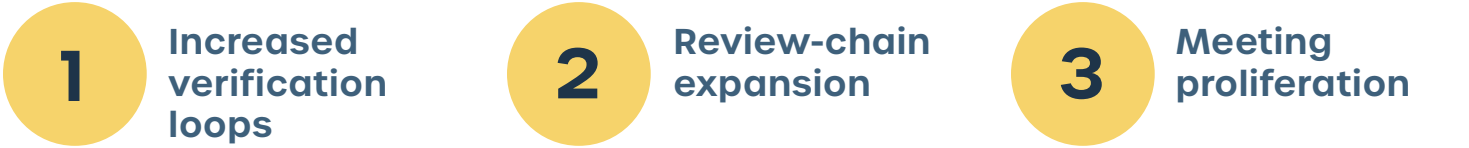
Each addition is defensible in isolation. Together they reconstitute the original cycle time and add a new failure mode: ownership diffused across a review group with no single accountable decision-maker.

Last is the least sophisticated but most common pattern: meeting proliferation. Recovered hours show up as open calendar space, and open calendar space in a matrixed organization gets filled. Alignment sessions. Readouts. Working sessions to prepare for the readouts.

As a result, the organization becomes measurably better informed but not more decisive.

Put the three together, and you get an organization that automates its most straightforward work, retains its highly complicated work, and reinvests the difference in the overhead of staying synchronized.

That is the coordination tax. And it does not appear on any dashboard I have ever been shown.



Originally published in Insights. Shared with permission, all rights reserved.

Obscured From Sight

The measurement problem is structural, not a tooling shortfall.

Throughput metrics count completed units of work: Claims closed. Submissions cleared. Policies issued. Those are the right metrics for asking whether a process is running. But they are the wrong metrics for asking whether an organization is deciding well. A portfolio can post record throughput while its most consequential calls are made late, made twice, or made by random people.

Here is the specific blindness: Redundant decision-making is invisible to throughput measurement.

When a steering committee reopens a trade-off it settled two quarters ago, no unit of work is destroyed. Nothing turns red. No funding approvals are needed. The meeting happens, a decision is reached, and the portfolio proceeds.

From afar, it looks like governance is working. But up close, the view is clearer: The organization is paying twice for the same decision, using the scarcest input it has.

Recovered hours make redundant decision-making cheaper on the margins, so more of it happens. Executives who did not previously have the calendar space to reopen a settled question now have it—plus, reexamining the question feels like rigor.

The cost of this action accrues in a currency that nobody tracks: senior judgment capacity. No line item. No variance report. It is simply consumed.

This is decision debt with a better user interface. But the interest still compounds.

Automation Work vs. Judgment Work

The underlying design error referenced earlier is that most AI deployments target the work that is easiest to describe. But the work that is easiest to describe is rarely the work that determines outcomes.

Document summarization is straightforward. Data extraction is straightforward. First-draft generation is straightforward. These are excellent automation targets, and carriers are right to start with them.

However, the work that governs portfolio performance is altogether different and might include:

- Deciding whether a program's original premise still holds
- Determining which of three defensible priorities gets the scarce engineering team
- Judging whether a regulator's informal signal warrants a change in posture
- Choosing what not to do

None of those assignments is slow-going for lack of tooling. It is slow because it requires people to hold incomplete information, weigh trade-offs across functions with incompatible incentives, and accept accountability for outcomes that can't be fully modeled.

Automating around that work does not accelerate it. It surrounds it with more input.

This is why the finding on discrete use cases is more diagnostic than initially thought.[4] Function-level automation without connected decision-making does not merely fail to deliver enterprise value. It can actually degrade it by increasing the volume and velocity of material arriving at decision points that nobody redesigned to handle more.

Faster inputs into an unchanged decision architecture do not produce faster decisions. They produce a longer queue, arriving sooner.

Deciding whether a program's original premise still holds

Determining which of three defensible priorities gets the scarce engineering team

Judging whether a regulator's informal signal warrants a change in posture

Choosing what not to do

Faster inputs into an unchanged decision architecture do not produce faster decisions.

They produce a longer queue, arriving sooner.

Originally published in Insights. Shared with permission, all rights reserved.

What to Measure Instead

Three measures beat throughput, and all three are available without buying anything.

First is the number of reconsidered decisions. For each material initiative, how many times has a previously settled trade-off been reopened? What changed each time to justify reopening it?

This is the most useful number in portfolio governance, but almost nobody tracks it—which is a large part of why it still works. A program that's reworked three times despite showing green, or good overall health, is telling you something no status color can.

Another benefit of this metric is that it's hard to game. Reworked decisions leave traces in meeting records, revised baselines, and version histories of decision memos. You can start counting retroactively for the last four quarters this week.

Second is decision latency, measured from the point of sufficiency. Importantly, this measure does not reflect the date the decision was requested but the date the organization had enough information to make it. The gap between these two dates is where the coordination tax lives.

If sufficiency was reached in week two, and the decision landed in week nine, those seven weeks were not analysis. They were alignment.

Finally, the third measure is reviewer count by decision class, tracked over time. If the number of people required to weigh in on a standard underwriting exception, funding change, or claims escalation has gone up since AI tooling arrived, you know where the recovered capacity went. It is a simple count, and the trend is unambiguous.

None of these measures requires a platform purchase. They require someone with the authority to ask—and the appetite to hear the answer.

1

Number of reconsidered decisions

2

Decision latency

3

Reviewer count by decision class

The Afternoon Test

Here is the diagnostic I would run in any organization claiming AI-driven productivity gains.

Pick a decision class that matters: program funding changes, large settlement authority, underwriting exceptions above a threshold, whatever is material in your environment.

Pull related data from the last 12 months. For each instance, record three things: the date the organization first had sufficient information, the date the decision was made, and the number of approvals required. Then compare the first six months against the second six.

If throughput rose as decision latency and reviewer count also rose, the productivity gain was consumed by coordination. That is not hypothetical. In my experience, it is the default, because coordination is what organizations do with slack unless someone deliberately decides otherwise.

The direction matters more than the precision. You are not calculating an exact tax rate. You are establishing whether the arrow points the wrong way, and that is usually obvious after one afternoon with the records.



Relationship Between Productivity Gain and Coordination

Originally published in Insights. Shared with permission, all rights reserved.

Reserving Judgment—on Purpose

The remedy is not to automate less. It is to be explicit about what you are delegating and what you are reserving.

Most carriers now have AI governance covering model risk, bias testing, and regulatory compliance. The NAIC's *Model Bulletin on the Use of Artificial Intelligence Systems by Insurers* helped formalize expectations for written AI governance and risk management programs.[5]

Further, the Earnix study previously referenced found that 92 percent of insurers now run formal AI governance reviews on a regular cadence but that fewer than one in three executives strongly agree that those reviews keep pace with regulatory demand.[6]

This gap is not accidental. Model governance answers a different question from the one at issue here. It asks whether a system is safe to use, not which decisions the organization has chosen to keep.

AI-assisted work falls into three categories:

- Delegate—The machine initiates production, and available, potentially rotating, humans audit periodically.
- Augment—The machine initiates production, and a specific, *named* human decides and owns the outcome.
- Reserve—Output is solely human-generated, regardless of how good the model gets.

Delegate
Humans audit periodically

Augment
A named human owns machine-initiated output

Reserve
Solely human-generated

An organization that has not established these lines will drift toward the worst available configuration: automated inputs, unautomated decisions, and a growing population of people who feel obligated to weigh in on both. Organizations that have done this work, conversely, can initiate next steps:

- Name a decision-maker for each material decision class—Not a committee, a person. More specifically, a directly responsible person. Committees are excellent at consultation but structurally incapable of accountability; you cannot call a committee at 9 p.m. on a Friday. That distinction collapses under pressure unless it is written down in advance.
- Write decision records that capture reasoning, not outcomes—A record that says "approved" prevents nothing. It is a receipt, not a decision. A record that captures what was decided, which alternatives were rejected, what would have to change to justify revisiting it, and who owns the decision simplifies any future rework. Whoever reopens the question now has to say what changed. That is the highest-leverage governance artifact available to a carrier, and it costs nothing but discipline.
- Set an explicit verification standard for AI-assisted work. "Check the model's output" is not a standard. It is a mood. "Verify the three data elements that drive the pricing decision; accept the narrative summary without line-by-line review" is a standard. Without such a standard, professionals verify according to personal risk tolerance, which, under regulatory scrutiny, trends toward verifying everything. Trust should be earned per task and reviewed on evidence, not granted per vendor at contract signing.

The remedy is not to automate less.

It is to be explicit about what you are delegating and what you are reserving.

1 Name a decision-maker.

Not a committee, a person.

2 Write decision records.

Capture reasoning, not outcomes.

3 Set a verification standard.

Explicit, for AI-assisted work.

Originally published in Insights. Shared with permission, all rights reserved.

An Uncomfortable By-product

There is a harder implication here, and the industry conversation tends to walk around it.

If your recovered capacity is being allocated to coordination efforts, technological efficiencies aren't the headline. How your organization allocates authority is.

Coordination expands to fill available time when nobody is empowered to end a discussion. The AI deployment did not create that condition. It revealed it. Then it funded it.

Carriers that get the most out of AI tools over the next three years will not be the ones with the best models. They will be the ones that treat productivity gains as openings to redesign when decisions get made and who owns them and that accept the discomfort of assigning individual accountability in an industry that has spent decades diffusing it.

The technology question is largely settled. The governance question is wide open, and it is what separates carriers whose AI investment shows up in business results from carriers whose AI investment shows up only on a dashboard. ■

> The technology question is largely settled.
> **The governance question is wide open.**

[1] Earnix, *The Earnix 2026 Industry Trends Report: The Race to Reinvent*, 2026.

[2] Kevin Zheyuan Cui, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz, "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers," *Management Science*, 2026.

[3] Shakked Noy and Whitney Zhang, (2023). "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, July 13, 2023, vol. 381, issue 6654, pp.187–192.

[4] Earnix, *The Earnix 2026 Industry Trends Report*.

[5] National Association of Insurance Commissioners, *Model Bulletin on the Use of Artificial Intelligence Systems by Insurers*, December 4, 2023.

[6] Earnix, *The Earnix 2026 Industry Trends Report*.



Originally published in Insights. Shared with permission, all rights reserved.