

AI Work Redesign: Know When to Assist, Rebuild, Wait, or Quit

A five-part scoreboard for deciding how much of the work - not merely the job title - to redesign around AI.

DecisionGraph Weekly · DGW-2026-10-03 · September 28-October 3, 2026 · Draft for Human Review

The decision

Start with the smallest intervention that improves the complete workflow. Move from AI assistance to targeted redesign or a broader rebuild only when evidence for value, performance, control, resilience, and organizational capacity justifies the added cost and disruption.

The question is: **Which jobs are ready to be rebuilt around AI, which should only get AI assistance, and what scoreboard tells us when to GO BIG, PUT GUARDRAILS AROUND IT, WAIT, or QUIT?**

Here, “jobs” means the work being performed - tasks, handoffs, decisions, roles, and workflows - not merely job titles. Evidence that AI improves one task does not prove that an entire job should be redesigned or removed.

AI everywhere is not a defensible default. Do not rebuild more of the job or workflow than the evidence justifies.

The 60-second view

AI produces different results in different work. In a field study of 5,179 customer-support agents, bounded AI assistance increased issues resolved per hour by about 14% on average, with larger gains for less-experienced workers. In a randomized study of 16 experienced open-source developers completing 246 tasks, access to early-2025 AI tools increased completion time by about 19%. Those findings describe different tasks, users, tools, and operating conditions. Together they show why a general mandate to “use more AI” is weak decision logic.

Executives have four choices. Use **AI Assistance** when a bounded task can improve while the workflow and accountability remain intact. Choose **Targeted Redesign** when a specific handoff, role, process step, or decision right blocks value. **Rebuild Around AI** only when whole-workflow economics, performance, ownership, controls, resilience, and workforce readiness support the larger change. **Wait or Quit** when a decision-critical input is unresolved or when the intervention fails the required test.

Use five lenses - **Value, Performance, Readiness & Control, Resilience, and People & Change**. Each organization must set its own thresholds before testing. The scoreboard does not produce a universal score; it shows which conditions support expansion, require limits, justify a pause, or call for stopping.

Decision confidence: strong for the direction; organization-specific for implementation.

Why now

The case for AI-enabled work redesign is becoming more ambitious while the evidence remains uneven. Commercial surveys report associations between deeper workflow redesign and stronger financial outcomes, but they do not establish that redesign caused those results. Field experiments show that AI can improve bounded work and, in selected settings, replace some coordination that would otherwise require a team. A randomized developer study also shows that experienced people can become slower with AI, even when they expect to become faster.

This is a portfolio decision as much as a technology decision. A U.S. Government Accountability Office review of 126 IRS AI use cases found rapid portfolio growth alongside gaps in benefit information, inventory quality, ownership, skills planning, and strategic alignment. The lesson is practical: expansion without measurement and accountability can make the portfolio harder to manage before it makes the work better.

The business consequence

The wrong intervention wastes value in two directions. A company can underreach by adding a helpful assistant to a workflow whose real constraint is an approval, handoff, role boundary, or decision right. It can also overreach by rebuilding a workflow when bounded assistance would capture most of the benefit with less transition cost, less control exposure, and more reversibility.

The unit of analysis must therefore be the complete workflow outcome. Measure what reaches the customer or operating result after review, correction, exception handling, escalation, integration, security, compliance, training, and change costs are included. A faster task is useful only when it improves the system around it.

Four intervention paths

AI Assistance

Keep most of the work intact and use AI for bounded tasks such as drafting, summarizing, searching, classifying, or supporting a decision.

Choose assistance when the task is well defined, the output can be checked, human ownership remains clear, and the effect can be measured against a baseline. This path usually offers the fastest learning and the greatest reversibility.

Main risk: task-level gains can disappear after review, correction, rework, exceptions, and downstream quality effects are counted.

Targeted Redesign

Change the specific handoffs, roles, process steps, or decision rights that prevent a task-level gain from improving the workflow outcome.

Choose targeted redesign when assistance works but a measured bottleneck remains - for example, duplicated review, slow approval, unclear routing, or an ownership gap. Change the constraint and preserve the parts of the operating model that still work.

Main risk: redesigning the wrong bottleneck adds complexity without improving the outcome.

Rebuild Around AI

Redesign the broader workflow around AI or agent capabilities only when whole-workflow evidence supports the larger intervention.

Choose this path only when complete economics are positive, performance is reliable, data and controls are ready, ownership is explicit, exception handling and fallback work, the workforce can absorb the change, and reduced reversibility is acceptable. Survey evidence alone is not enough.

Main risk: this path carries the greatest transition cost, organizational disruption, control burden, workforce impact, data dependency, and switching cost.

Wait or Quit

Wait when a decision-critical input is unresolved and a bounded test can answer it. Name the missing evidence, its owner, the test, and the reassessment date.

Quit when measured value is negative or insufficient, quality or reliability deteriorates, exceptions cannot be controlled, compliance or accountability fails, or the intervention no longer justifies its full cost and risk.

Main risk: waiting without an owner and a decision trigger becomes an indefinite pilot.

The five-part executive scoreboard

The scoreboard is a set of decision gates, not a composite score. Do not weight, average, rank, or convert the lenses into a maturity level.

1. VALUE

Question: Is the measurable value sufficient for the full cost?

Measure business value against the current baseline; comparative cost; time to value; and transition and integration cost. Count model and software costs, human review, correction, exception handling, security, compliance, training, process change, integration, and switching or exit costs.

Decision effect: missing or negative economics supports WAIT or QUIT. Positive task economics with uncertain workflow economics supports PUT GUARDRAILS AROUND IT.

Enterprise-specific threshold required.

2. PERFORMANCE

Question: Does the complete workflow perform reliably enough?

Measure quality, reliability, exception rate, and human-review burden at the final workflow outcome. Segment results by task, user group, data condition, and exception type so that averages do not hide failure pockets.

Decision effect: reliable performance inside a defined boundary can support guardrails. Material deterioration, uncontrolled exceptions, or growing review burden supports WAIT or QUIT.

Enterprise-specific threshold required.

3. READINESS & CONTROL

Question: Can the organization own and govern the intervention?

Test data readiness, accountable ownership, and compliance and control performance. Identify who accepts the output, handles exceptions, approves changes, monitors drift, and can stop the system.

Decision effect: missing ownership, unusable data, or ineffective controls supports WAIT or a narrower intervention. Expansion requires evidence that controls work in operation, not merely that policies exist.

Enterprise-specific threshold required.

4. RESILIENCE

Question: Can the organization recover, reverse, or switch?

Test fallback and recovery, plus reversibility and switching cost. Exercise the fallback before scale. Know how the work continues when the model, provider, integration, data, or control fails.

Decision effect: weak recovery or high lock-in raises the evidence required for expansion. An unworkable fallback can justify WAIT or QUIT even when normal-case performance looks attractive.

Enterprise-specific threshold required.

5. PEOPLE & CHANGE

Question: Can the workforce and operating model absorb the change?

Assess workforce readiness across skills, roles, training, adoption, workload, incentives, and change capacity. Measure whether responsibilities become clearer and whether the people doing the work can operate, challenge, and recover the system.

Decision effect: unresolved change capacity supports WAIT or a smaller intervention. A broader rebuild requires evidence that the new roles and decision rights work in practice.

Enterprise-specific threshold required.

How the scoreboard produces a decision

GO BIG

Expand only when the complete workflow shows repeatable value, reliable performance, effective ownership and controls, workable recovery, acceptable reversibility, and sufficient workforce capacity. The decision follows from passed gates across all five lenses, not from enthusiasm about the technology.

Governed mapping: SCALE.

PUT GUARDRAILS AROUND IT

Keep the proven boundary when value is real but limited to a task, role, user segment, data set, environment, or control condition. State what is allowed, what remains human-owned, which exceptions require escalation, and what evidence would permit expansion.

Governed mapping: LIMIT.

WAIT

Pause expansion when a material question can be answered through a bounded test. A valid wait decision names the missing input, owner, test method, and date or event that reopens the decision.

Governed mapping: DEFER.

QUIT

Stop when full-workflow value remains insufficient, quality or reliability fails, exceptions cannot be controlled, accountability or compliance breaks, recovery is unacceptable, or the operating model cannot support the intervention.

Governed mapping: STOP.

What can go wrong

- **The task improves while the workflow does not.** Faster drafting can be offset by checking, rework, escalation, or downstream defects.
- **The wrong work is redesigned.** A visible AI task can distract from the actual bottleneck in a handoff or decision right.
- **Average results conceal weak segments.** Gains for one user group or task type may coexist with losses elsewhere.
- **Ownership becomes ambiguous.** People may rely on an output without knowing who is accountable for accepting, correcting, or stopping it.
- **The organization scales before it can measure.** A growing AI portfolio can outpace benefit tracking, inventory quality, skills planning, and control capacity.
- **The change becomes hard to reverse.** Integration, data, provider, and role decisions can create switching costs before durable value is proven.

What to do

1. **Name one workflow and its owner.** Avoid starting with a job title or an enterprise-wide mandate.
2. **Record the baseline.** Capture throughput, quality, cost, rework, exceptions, review burden, customer or operating outcome, and current control performance.
3. **Choose the least extensive plausible intervention.** Begin with assistance unless evidence already identifies a workflow constraint that requires redesign.
4. **Set the five thresholds before the test.** Define the evidence that leads to GO BIG, PUT GUARDRAILS AROUND IT, WAIT, or QUIT.
5. **Run a bounded comparison.** Compare the intervention with current work using the same outcome measures and relevant user and task segments.
6. **Review the complete economics and failure modes.** Include transition, review, exception, control, recovery, training, and switching costs.
7. **Expand only after the evidence clears the next scope of change.** A successful assistant is evidence for that assistant; it is not automatic proof for a full rebuild.

What to watch

- Whether workflow outcomes improve after review, correction, and exception handling.
- Whether performance differs by expertise, task type, data condition, or operating environment.
- Whether human-review burden rises as use expands.
- Whether ownership, control performance, and stop authority remain clear.
- Whether fallback and recovery work under realistic failure conditions.
- Whether integration or provider choices reduce reversibility faster than value becomes durable.

- Whether roles, skills, workload, and adoption support the intended operating model.
- Whether new evidence changes the selected intervention rather than merely changing confidence at the margin.

Decision checkpoints

Checkpoint	Question	Continue when	Change course when
Before the test	What outcome and baseline are we changing?	The workflow, owner, measures, thresholds, and stop conditions are explicit.	The team cannot define the decision unit or current performance.
After bounded use	Did the complete workflow improve?	Value and performance persist after review, rework, exceptions, and quality effects.	Gains are task-only, segment-specific, or offset elsewhere.
Before redesign	Is a measured constraint blocking value?	The constraint is evidenced and the proposed change is proportionate.	The redesign is based on a general AI mandate or survey association.
Before scale	Do all five lenses clear the required threshold?	Economics, performance, controls, resilience, and change capacity are sufficient.	A material gate is missing, failing, or unowned.
During operation	Are the conditions still true?	Outcomes stay within bounds and monitoring, escalation, fallback, and stop authority work.	Drift, exceptions, cost, control failures, or workforce effects cross a stop condition.

The economic test

Do not compare software price with labor time in isolation. Compare the current workflow with the AI-enabled workflow across the full decision horizon.

Count value created, avoided loss, throughput, quality, and time to value. Subtract software and model cost, integration, review, correction, exception handling, security, compliance, training, process transition, monitoring, recovery, and switching or exit costs. Test how the result changes when adoption, accuracy, volume, review burden, or provider cost moves away from the plan.

This issue does not set a universal dollar range, payback period, or scale threshold. Those inputs are enterprise-specific and must be established before expansion. The immediate decision horizon is to begin with baseline measurement and a bounded evaluation; broader redesign follows only after local evidence clears the relevant gates.

Keep the counterevidence in the decision

Bounded assistance can create material gains. Selected AI use can also replace some coordination work. Neither finding proves that a whole job or workflow should be rebuilt. The METR study shows the opposite can occur: experienced developers using early-2025 tools took longer in its governed setting. Commercial surveys may identify useful patterns, but

association is not causation. Portfolio expansion without measurement, ownership, data quality, skills, and controls can make decision quality worse.

The answer depends on the work, the evidence, and the operating conditions.

Evidence and source links

- McKinsey & Company, [*The State of AI: How Organizations Are Rewiring to Capture Value*](#) (2025). Commercial survey evidence; association does not establish causation.
- Accenture, [*The Agentic Imperative: Generating Impact*](#) (2026). Commercial survey and practitioner guidance; source-reported findings remain context-bound.
- U.S. Government Accountability Office, [*Artificial Intelligence: IRS Actions Needed to Address Skills Gaps, Information Quality, and Strategic Management*](#), GAO-26-107522 (2026). Independent audit in a U.S. federal-agency context.
- Erik Brynjolfsson, Danielle Li, and Lindsey Raymond, [*Generative AI at Work*](#), NBER Working Paper 31161. Field evidence from customer support; transfer requires local validation.
- METR, [*Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*](#). Randomized evidence from selected repositories, experienced participants, and early-2025 tools.
- Dell'Acqua et al., [*The Cybernetic Teammate*](#) (2025). Field experiment with 776 professionals; results remain task- and setting-specific.
- OECD, [*Agentic AI in Organisations: Early Insights from Practitioner Interviews*](#) (2026). Qualitative practitioner evidence on heterogeneous, governance-dependent adoption.

Methodology note

ZAPHAN compares evidence across methods and operating contexts instead of treating every source as equivalent. The public analysis uses original summaries, concise attribution, and links to public source pages. It does not reproduce protected charts, tables, screenshots, PDFs, or substantial source expression. Commercial survey findings are treated as associations; experimental and audit findings retain their populations, conditions, and generalization limits.

Apply the scoreboard to one named workflow

Choose one workflow, define its owner and baseline, and write the five enterprise-specific thresholds that will lead to GO BIG, PUT GUARDRAILS AROUND IT, WAIT, or QUIT. This is a decision exercise, not a functioning assessment form.