Why Should I Care?

EVIDENCE REVIEW · 2026

# Speed, Sense and Stick

## What AI has really changed in how businesses improve

An evidence review of AI's impact on business improvement and continuous improvement methods

**Claudio Costa**



whyshouldicare.net

Disclaimer. This report is my own independent research, based on information that is available and researchable online. It reflects my personal analysis only, and does not reflect the views or opinions of any company, or of any public or private organisation, including any I work with or am affiliated with, or any named in it.

2026
Copyright © Claudio Costa. All rights reserved.
Published under the Why Should I Care? brand.
whyshouldicare.net

# Contents


START HERE

Introduction 4

Key takeaways 5

THE APPROACH

Three kinds of AI, and where they could help 6

Scope, question and approach 7

TESTING HYPOTHESIS 1: WHAT AI CHANGES

Speed 8

Sense 9

Stick 10

Beyond speed: the rest of Lean 11

TESTING HYPOTHESIS 2: HOW STRONG THE EVIDENCE IS

Evidence for Lean methods 12

Sector evidence 13

Real life examples 14

VERDICTS AND ACTION

Verdicts on the two hypotheses 15

Five recommendations and four practices 16

Limitations and conclusion 17

REFERENCE

Appendix A: evidence tables 18

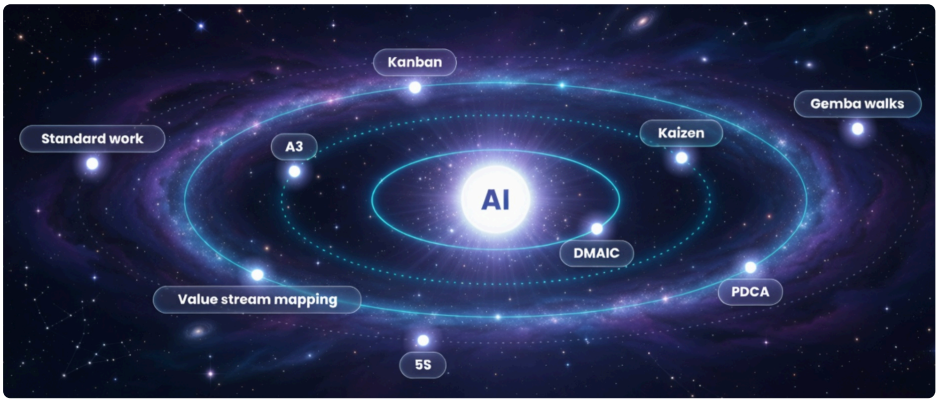
Appendix B: definitions and about the research 20

Sources and acknowledgements 21

INTRODUCTION

# AI has made improvement work faster. Whether it makes it better remains unproven.

Since late 2022, leaders have been told that AI will transform Lean, Lean Six Sigma and Kaizen, and some say it will replace them. Businesses have adopted AI widely, with 88% using it in at least one function (McKinsey, 2025), yet systematic evidence of what it does for these methods is almost non-existent. Plenty of people have a view. Far fewer have examined the evidence.




**The question for this report is how far AI makes these methods faster, more insightful and more effective.**

THREE TESTS

## Speed
Are defined tasks completed faster?

## Sense
Do organisations understand the processes AI acts on, and can the tools be trusted?

## Stick
Do improvements endure, and do people keep the skills and trust to sustain them?

TWO HYPOTHESES

HYPOTHESIS 1

AI has brought new capabilities that accelerate Lean, Lean Six Sigma and Kaizen initiatives and make them more relevant, insightful and effective.

HYPOTHESIS 2, THE FOLLOW ON TEST

If hypothesis 1 holds, is the evidence strong enough to prove it? The evidence that AI makes these methods work better is still thin and mostly indirect.

### The evidence base

Controlled studies, official statistics, practitioner sources and company disclosures from 2022 to 2026, across six sectors: manufacturing, financial services and telecoms, where the evidence is strongest, and healthcare, professional services and logistics, where it is thinner. The sectors illustrate rather than survey. Vendor reported, adjacent and company reported findings are marked. Lean also asks about customer value, flow and waste, so page 11 tests how far the evidence reaches beyond speed. The search was targeted rather than systematic, and its limits are on page 17.

KEY TAKEAWAYS

# AI can speed up defined tasks. Whether it makes improvement better is not yet proven.

HYPOTHESIS 1

ONLY PARTLY SUPPORTED

HYPOTHESIS 2, THE FOLLOW ON TEST

CONFIRMED

- **Speed is real, but only on bounded tasks.**
  According to Cui et al. (2025), 4,867 developers produced 26% more pull requests a week. According to METR (2025), experienced developers took 19% longer.
- **Sense: reliability is the constraint.**
  According to the APEX-Agents benchmark (2026), the best system completed 24% of realistic professional tasks in a single run.
- **Stick: sustainment was weak before AI.**
  According to McKinsey (2024), only 12% of transformations sustained gains beyond three years. No study shows that AI changes this.
- **Lean is more than speed.**
  Evidence on customer value, waste, daily management, go and see and kata is largely absent (page 11).
- **No audited saving is attributed to AI.**
  Across six sectors, the figures found are modelled, company reported or self reported (page 13).

FOUR NUMBERS THAT FRAME THE REPORT

**2% to 88%**
Reported AI use, depending on who is asked (Exhibit 4).

**19%**
Slower, in the best known real world trial, while developers believed they were faster.

**12%**
Of transformations sustained gains beyond three years, before generative AI.

**None**
Audited savings attributed to AI, found in any of six sectors.

THE CONCLUSION

**The evidence helps to answer hypothesis 1 only in part: AI speeds up defined tasks, but no controlled study shows that it improves how businesses improve. The old discipline still decides the outcome. AI makes it faster to apply, but it does not make it optional.**

WHICH AI, AND WHERE IT COULD HELP

# AI is not one technology: three kinds are tested here, and each could help improvement work in a different way

Discussion of AI often treats it as a single thing. This report separates three kinds, because the evidence for each is different, and so are the ways each could support Lean, Lean Six Sigma and Kaizen.

Exhibit 1

## Three kinds of AI appear in this report, and the evidence for each differs

| KIND OF AI | WHAT IT DOES | EXAMPLE USES IN IMPROVEMENT WORK | WHERE IT APPEARS IN THIS REPORT |
|---|---|---|---|
| **Machine learning** | Software that learns patterns from data in order to predict or classify, such as which machine is likely to fail or which part is defective. | Predictive maintenance, inspection and forecasting. | Manufacturing evidence, which predates generative AI (Exhibit 13), and a systematic review of machine learning in Lean (Exhibit 12). |
| **Generative AI** | Models that produce text, code, images or summaries from a prompt, usually as an assistant that a person works with. | Drafting, data analysis, dashboard preparation, suggesting possible root causes, coaching and knowledge capture. | Controlled task studies (Exhibit 5) and most of the early evidence on skills and trust (Exhibit 8). |
| **Agentic AI** | Systems that plan and carry out several steps towards a goal with some autonomy, using tools and data from other systems. | Running defined steps of a process and acting on what they find. | Benchmarks of realistic work (Exhibit 6) and deployment and cost evidence (Exhibit 7). |

Source: author's classification, based on how the sources cited in this report describe the technology they studied.

Exhibit 2

## Five ways AI could help improvement work have been suggested, and none has yet been proved

| METHOD | HOW AI COULD HELP | KIND OF AI | EVIDENCE TODAY |
|---|---|---|---|
| **Lean** flow and waste | Build a time stamped picture of how work actually flows, alongside sampling based value stream mapping, and predict equipment failure or spot defects. | Process mining and machine learning | One peer reviewed process mining case and one systematic review. |
| **Lean Six Sigma** DMAIC | Prepare and analyse data, draft dashboards and project documents, and suggest possible root causes for the team to test. | Generative AI | Commentary, frameworks and a few cases; no controlled DMAIC study. |
| **Kaizen** daily improvement | Help novices frame problems and capture know how, so more people can contribute improvement ideas while people keep the lead. | Generative AI | Inference from adjacent studies; no kaizen specific study. |
| **A3 and coaching** management system | Act as a coaching aid inside a management system run by people (Lean Enterprise Institute, 2026). | Generative AI | Practitioner position only. |
| **Redesigned processes** | Carry out defined steps of a process, but only after it has been redesigned and within the reliability limits in Exhibit 6. | Agentic AI | Benchmarks and deployment cases only; no controlled study in improvement work. |

Source: author's synthesis of the sources in Exhibit 12; Lean Enterprise Institute (12 March 2026); Bain (June 2026). These are possible uses, not findings.

### How to read the rest of the report

The uses above are the plausible ones that the sources suggest. The tests that follow ask whether the evidence supports them: Speed asks whether they make work faster, Sense whether they can be relied on, and Stick whether the gains last.

SCOPE, QUESTION AND APPROACH

# Three tests, Speed, Sense and Stick, frame what the evidence can and cannot tell us

**The question.** What has changed since generative AI entered mainstream business use and agents began to be deployed, and what does it mean for Lean, Lean Six Sigma and Kaizen?

**The window.** 2022 to 2026. Earlier evidence is used only as a baseline and is labelled as such.

**The evidence.** Controlled studies and field experiments, official statistics, practitioner surveys, company disclosures, and examples from six sectors, chosen on stated tests (Appendix A, pages 18 and 19). Most task level evidence comes from knowledge work, software engineering and digital performance, not from Lean programmes, so it is read as evidence on business improvement, and on the methods only where stated.

**The rule.** Vendor claims and predictions are reported as claims and predictions, never as findings.

Exhibit 3

## Each test has its own question and its own way of failing

### SPEED

**Measures**
Whether defined tasks get done faster.

**Failure mode**
Tasks become faster without the process improving.

### SENSE

**Measures**
Whether organisations understand the process they act on, and whether the tools are reliable.

**Failure mode**
Automating a process nobody understands.

### STICK

**Measures**
Whether gains last, and whether people keep the skills and capacity needed to sustain improvement.

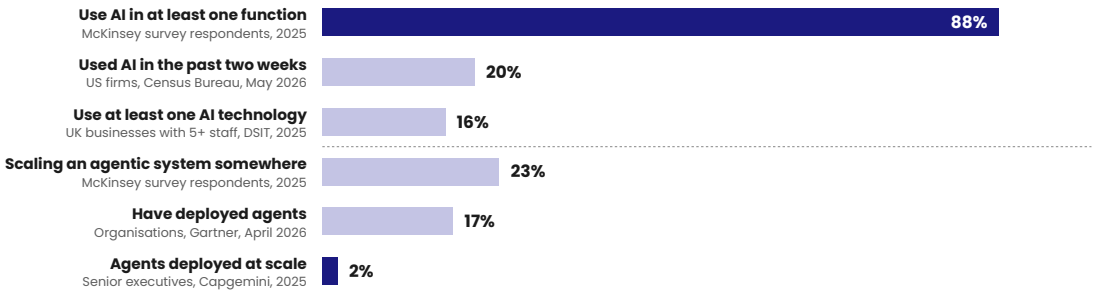
**Failure mode**
Gains fade after go live, and skills, trust and learning erode.

Source: author's framework.

Adoption says nothing about speed, yet it frames everything that follows. Headline figures are not directly comparable, because they cover different populations and define use differently. Consultancy surveys tend to reach organisations already engaged with AI, whereas official statistics cover the wider business population.

Exhibit 4

## Reported AI use ranges from 2% to 88%, depending on who is asked and what counts as use




Bars above the dashed line show any AI use, bars below show AI agents. Sources: McKinsey, The state of AI in 2025; US Census Bureau, AI use in businesses, 26 May 2026; DSIT, AI Adoption Research Report, January 2026; Gartner, Hype Cycle for Agentic AI, 15 April 2026; Capgemini Research Institute, Rise of agentic AI, 16 July 2025. Census question wording changed on 17 November 2025. Populations and definitions differ, so the bars are not directly comparable.

### The UK qualification

Use is widespread, but the financial effect is not yet clear. According to DSIT (2026), 75% of UK businesses using AI report better workforce productivity, yet 77% report no change in revenue so far. According to McKinsey (2026, n=1,719), 37% see some effect on profit and only 6% qualify as AI high performers, the same as in 2025.

TEST 1 · SPEED SUPPORTED ON BOUNDED TASKS

# Speed: gains are real on bounded tasks, but they shrink as work becomes realistic

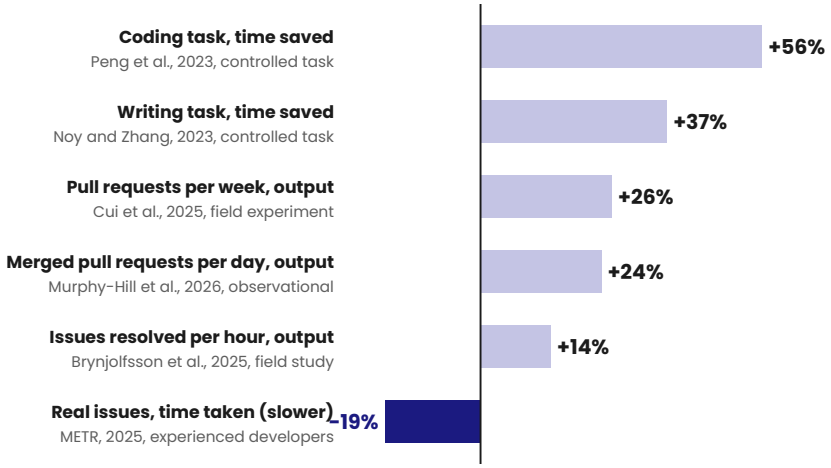
**The question.** Do defined tasks get done faster with AI, and does that hold in real work? **What the evidence shows.** Controlled studies provide the strongest evidence that AI can accelerate clearly defined tasks. Results are more mixed in real working environments, and more activity does not necessarily produce better outcomes.

**19%**

slower, while believing they were 20% faster (METR, 2025)

Exhibit 5

## Controlled studies show substantial task level gains, but the best known real world trial found a slowdown




A positive figure is a gain: either time saved on a task or more output. The navy bar shows experienced developers took longer with AI. Each bar shows a different measure, so the bars are not directly comparable, and a cut in time is not the same as a rise in output. These studies test generative AI assistants. Sources: Peng et al. (2023); Noy and Zhang (2023); Cui et al. (2025); Murphy-Hill et al. (2026); Brynjolfsson et al. (2025); METR (2025). The last two studies below (Dell'Acqua; Chen and Stratton) measure other outcomes and are not plotted.

## Gains are largest for less experienced staff

According to Brynjolfsson, Li and Raymond (2025), support agents resolved 14% more issues per hour on average and 34% more among novices, with minimal effect for experienced staff. That suggests AI could spread standard ways of working faster, although the evidence comes from support and writing, not kaizen practice.

## Perceived and measured gains differ

According to METR (2025), experienced developers expected a 24% speed up, measured a 19% slowdown, and still believed afterwards that they had been 20% faster. A follow up in February 2026 describes its own results as very weak evidence, because of selection effects.

## Gains depend on the task

According to Dell'Acqua et al. (2023), 758 BCG consultants completed 12% more tasks and worked 25% faster on tasks within the AI's ability, but were 19% less likely to be correct on a task outside it. Knowing which is which takes process knowledge as well as technical skill.

## More activity, no more resolved issues

According to Chen and Stratton (2026), across 718 firms commits and pull requests per worker rose by about 20% to 30%, yet resolved issues showed no detectable increase, review time rose 49%, and the share of pull requests needing changes nearly doubled.

### Evidence grade

●●○ Moderate to strong on tasks, limited on outcomes.

**In plain words.** A bounded task is a clearly defined piece of work, such as drafting a document or writing a function.

### What this means for hypothesis 1 (Speed)

Speed has the strongest evidence of the three tests, but it does not answer the central question alone. Faster tasks do not necessarily produce better processes or better quality outcomes.

### The macro picture has not yet moved

According to Humlum and Vestergaard (2025), Danish data rule out effects on earnings or hours larger than 2% two years after ChatGPT. Acemoglu (2024) estimates a total factor productivity rise of no more than 1% over ten years; Goldman Sachs (2023) estimates 7% on global GDP. Neither is tested over its full period.

TEST 2 · SENSE RELIABILITY IS THE CONSTRAINT

# Sense: AI capability is advancing rapidly, but reliability under real working conditions remains the binding constraint

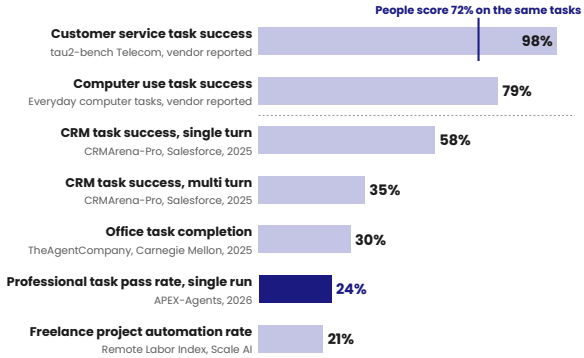
**The question.** Can organisations rely on AI tools, and agents in particular, inside real processes they understand? **What the evidence shows.** Benchmark scores are rising fast, but success on realistic professional tasks is still low and inconsistent, and the return and failure statistics in circulation measure different things.

**24%**

of realistic professional tasks completed in a single run by the best system (APEX-Agents, 2026)

Exhibit 6

## Structured scores beat the human baseline, but completion on realistic professional tasks stays below a quarter to a third




Source: OpenAI, Introducing GPT-5.5, 23 April 2026; Salesforce, CRMArena-Pro, arXiv 2505.18878; Carnegie Mellon via The Register, 29 June 2025; APEX-Agents, arXiv 2601.14242; Scale AI, Remote Labor Index. The benchmarks measure different capabilities and should not be compared directly. The two rows above the dashed line are vendor reported structured tasks; the rows below are realistic work. The vertical line shows how well people score on the same everyday computer tasks (72%).

Exhibit 6 shows scores falling as work becomes realistic. According to APEX-Agents (2026), the best system completed 24% of professional tasks in a single run, and only 1% to 13% in all eight repeat runs: consistency is the gap, not peak performance. Not shown in the exhibit: METR (January 2026) finds the task length models complete at 50% success has doubled about every 4.3 months since 2023.

Real world deployment tells a similar story. In May 2025, Klarna's chief executive said AI cost less but delivered lower quality, and the company resumed hiring. In 2025 the Commonwealth Bank of Australia reversed 45 redundancies attributed to AI. In Moffatt v. Air Canada (2024), a tribunal held the airline responsible for its chatbot's advice.

### Evidence grade

●○○ Limited: vendor reported benchmarks and early deployment stories.

**In plain words.** An agent is software that plans and carries out several steps towards a goal, using tools and data from other systems.

### What this means for hypothesis 1 (Sense)

Rising benchmark scores do not show that AI agents can run reliably inside real processes. Practitioners warn that automating a poorly designed process amplifies its weaknesses, but no controlled study has quantified this. Going to see the work is also part of Sense (page 11).

Exhibit 7

## Return and failure statistics are widely quoted and measure different things

| SOURCE | FIGURE | WHAT IT ACTUALLY MEASURES |
|---|---|---|
| **MIT NANDA, July 2025** | 95% of organisations getting zero return. | Directional estimate for custom enterprise tools reaching production; critics question the sample. |
| **BCG, September 2025** | 5% create value at scale, 60% no material value. | 1,250 executives in 68 countries. The 2026 edition is defined differently and is not like for like. |
| **Deloitte, January 2025** | Nearly three quarters meet or exceed ROI expectations. | 2,773 respondents with initiatives already running. |
| **KPMG, June 2026** | 49% scaled back, narrowed, delayed or paused agents. | A pause or scale back rate, not a cancellation rate; n=2,145 senior leaders. |
| **Gartner, June 2025** | Over 40% of agentic projects cancelled by end 2027. | A prediction, not an observed rate. No measured cancellation rate had been published by October 2026. |

Source: MIT NANDA, The GenAI Divide; BCG, The Widening AI Value Gap; Deloitte, State of Generative AI in the Enterprise; KPMG, Global AI Pulse Q2 2026; Gartner, 25 June 2025.

TEST 3 · STICK UNPROVEN, WITH EARLY RISKS

# Stick: sustainment was already the weak point, and early evidence on people introduces new risks

**The question.** Do the gains last, and do people keep the skills, trust and capacity to sustain improvement? **What the evidence shows.** Sustaining gains was already the weak point before AI, and early evidence suggests AI may weaken skills and trust. None of it comes from improvement teams.

Sustaining gains was already hard before AI. According to McKinsey (2024), most organisations achieved less than a third of the impact they expected from technology transformations. The survey ran in 2022, before generative AI, so it cannot show whether AI helps or hurts. Lean relies on gains lasting through daily management, standard work and coaching, and no study was found that tests whether AI improves that.

**12%**

of transformations sustained gains beyond three years (McKinsey, 2024), before generative AI

### Evidence grade

●○○ Limited: early evidence, mostly adjacent to improvement work.

**In plain words.** Sustainment is whether the gains from an improvement are still in place years later, with people still able to keep improving.

Exhibit 8

## Early evidence raises questions about skills, trust, employment and cost visibility

**17% worse**
**Skill erosion**
Students scored 17% lower once unguarded AI access was removed (Bastani et al., PNAS, 2025).

**319 workers**
**Critical thinking**
In a survey of 319 knowledge workers, higher confidence in AI went with less critical thinking effort (Lee et al., CHI 2025).

**66%**
**Unchecked reliance**
Rely on AI output without checking; 56% report mistakes in their work (KPMG and University of Melbourne, 2025).

**41%**
**Low quality output**
Received low effort, AI generated work, labelled workslop (Harvard Business Review, September 2025).

**43% to 27%**
**Trust in autonomy**
Trust in fully autonomous agents fell in a year (Capgemini Research Institute, July 2025).

**26%**
**Cost visibility**
Of leaders say AI operating costs are fully visible (KPMG Q2 2026, n=204 US leaders).

Source: as cited in each tile. None of these studies examined kaizen or Lean teams, so implications for improvement culture are inferred. On employment, Brynjolfsson, Chandar and Chen (2025) report a 13% relative decline for ages 22 to 25 in exposed occupations; the Yale Budget Lab (2025) finds no aggregate disruption.

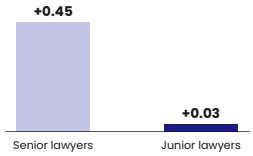
## Do people keep the skill? Better drafts, little skill gain

This is the one randomised trial found that tests whether AI builds the skill that lasting improvement depends on. According to Autor et al. (NBER Working Paper 35720, 2026), after three months, unassisted contract editing skills improved by 0.45 standard deviations among senior patent lawyers, but by only 0.03 among junior lawyers, a result that was not statistically significant. Junior lawyers still produced better drafts with AI, but the narrower gap is little evidence that their own skills improved.

Employees might also hesitate to suggest improvements that could automate their own roles. That is a hypothesis, supported only indirectly by the finding that 59% of respondents worry about automation (Edelman, 2025).

Exhibit 9

### Unassisted skill after three months, change in standard deviations




Source: Autor et al., NBER Working Paper 35720, 2026; 133 patent lawyers, randomised trial.

### What this means for hypothesis 1 (Stick)

There is no direct evidence that AI makes gains last. Before AI, only 12% of transformations sustained gains for three years, and early evidence on skills, trust and cost visibility points to added risk, not relief. In manufacturing, plants that dropped KPI reviews lost more (page 13). Smalley (The Lean Post, 2026) argues that Toyota develops AI through its own people, guided by jidoka (see Appendix B). That is a practitioner view, not a finding. Practice and leadership matter too (page 11).

BEYOND SPEED · THE REST OF LEAN

# Lean is more than speed, and the evidence barely reaches customer value, waste, go and see or capability

**The question.** Lean asks how well a business serves its customer, how work flows, which steps add value and how people learn. Does the evidence show AI helping with these, as well as with speed?

Exhibit 10

## Evidence is thickest for speed and thinnest for the human parts of Lean

| LEAN ELEMENT | HOW AI COULD HELP | WHAT WAS FOUND | STATUS |
|---|---|---|---|
| **Customer value** | Summarise customer feedback and complaints, so teams define value more precisely. | No evidence found: no study has tested AI for this in real improvement work. | NONE FOUND |
| **Flow** | Build a time stamped picture of how work moves and where it waits. | According to Menano de Carvalho et al. (2026), customs document handling fell from 15 to 4 minutes (Exhibit 11). According to Vossler et al. (2026), a hospital team used AI to analyse 500 patient records in about 30 minutes, whereas a manual review of 25 records took about 100 hours. The AI found everything the earlier manual Lean review found, which suggests AI could widen how many cases a Lean review covers. One hospital, not yet peer reviewed. | LIMITED |
| **Waste: value adding, non value adding and necessary** | Classify each process step by the value it adds. | According to De Michele et al. (2025), a preprint tested a language model on classifying process steps by Lean value across 50 process models. No field study in real improvement work was found. In the author's view, people must make this call, because a step that looks like waste to a model may be needed for safety, compliance or customer reasons that only the business knows. | LIMITED |
| **Standard work** | Draft and update standard work from observed steps. | No evidence found: no study has tested AI for drafting or updating standard work. | NONE FOUND |
| **Daily management** | Prepare KPI boards and flag deviations before the tier meeting. | No study of AI inside tier meetings or KPI reviews was found. According to McElheran et al. (2025), plants that dropped KPI reviews lost more productivity after adopting AI (Exhibit 13). | ADJACENT |
| **Problem solving: A3 and PDCA** | Suggest possible causes and draft A3s for the team to test. | According to Shen and Tamkin (2026), in a preprint experiment AI assistance weakened conceptual understanding and debugging, although three of six ways of using it preserved learning. | LIMITED |
| **Go and see** | Spot patterns in records and sensor data before a visit. | No study was found that compares AI monitoring with going to see the work. Heaton (Lean Enterprise Institute, 2025) argues there is no substitute for the gemba, and that AI can create an illusion of learning. | PERSPECTIVE |
| **Kata, practice and coaching** | Coach practice routines, or reduce the need for them. | No study of AI with the Improvement Kata or coaching routines was found. The one Kata study located, Lodgaard et al. (2023, construction), involved no AI and reported no outcome data. | NONE FOUND |

Source: as cited in each row. Grades are the author's judgement. Vossler et al., arXiv 2604.20055; De Michele et al., arXiv 2504.06600; Shen and Tamkin, arXiv 2601.20245; Heaton, Lean Enterprise Institute, 27 August 2025; Lodgaard et al., International Journal of Lean Six Sigma, 2023.

### Sense: go and see

**Key message.** Lean treats going to see the work as the way to sense what is really happening. AI can analyse records quickly, but records are not the same as observation.

**Evidence.** AI works from records, which show what was logged, not always what happened. Studies show it can analyse records fast, but none shows that records can replace observation.

### Stick: practice, leadership and capability

**Key message.** Lasting improvement depends on people's skill and on how leaders lead. Relying on AI can erode skill, since completed work can be mistaken for learned capability.

**Evidence.** According to Budzyń et al. (2025), in four Polish centres the rate at which doctors detected adenomas in colonoscopies without AI fell from 28% to 22% after AI was introduced (observational). According to Van Kleeff and van Harten (2025), at a Dutch hospital, line managers' Lean leadership drove performance, not Lean techniques or staff involvement.

EVIDENCE FOR THE IMPROVEMENT METHODS

# Lean, Lean Six Sigma and Kaizen: the evidence is thin, conceptual and mostly adjacent

**The question.** Is there direct evidence that AI improves Lean, Lean Six Sigma or Kaizen practice itself? The closer a study sits to real Lean practice, the weaker the evidence becomes.

Exhibit 11

## Direct evidence is absent, and what exists is adjacent, estimated or conceptual

| EVIDENCE TYPE | WHAT WAS FOUND |
|---|---|
| SYSTEMATIC REVIEW, ADJACENT | According to Serevan and Kuşakçı (2026), 54 articles show improvements in defects, downtime and cycle time from machine learning in Lean manufacturing. Validation was limited and no combined effect size was reported. |
| PEER REVIEWED CASES | According to Menano de Carvalho et al. (2026), customs document handling fell from 15 to 4 minutes, and shop floor reporting from about 48 to about 15.5 hours a month. Baselines were estimated, with no control group. |
| PREPRINT, SINGLE SITE | According to Zhao et al. (2026), an agent's main root cause finding matched engineers' consensus in 169 of 200 reports (85%), and time to identify problems fell from 47 to 19 minutes. Not peer reviewed and not presented as Lean practice. |
| CONCEPTUAL OR VENDOR | DMAIC roadmaps proposing hypotheses without data, book chapters asserting benefits, and vendor content. |
| CONTROLLED STUDY | No evidence found: no controlled study of generative AI within DMAIC, A3, value stream mapping or kaizen. |

Source: Serevan and Kuşakçı (2026); Menano de Carvalho, Peças and Jorge, Sustainability 18(10), May 2026; Zhao, Fan and Li, engrXiv, June 2026; Căsăneanu Dascălu et al. (2025); Laureani (2026). The review also found no controlled study of AI's effect on suggestion schemes or kaizen events.

Exhibit 12

## AI offers potential benefits and distinct risks for each approach, but direct evidence is limited

| APPROACH | POSITIVE IMPACT | NEGATIVE IMPACT OR RISK | EVIDENCE STATUS |
|---|---|---|---|
| **Lean** flow, waste, management system | Process mining adds time stamped flow data beyond sampling based value stream mapping. Machine learning supports predictive maintenance and inspection. | Automation can amplify process weaknesses; Forrester cautions against automating steps that should be removed. Process mining misses unrecorded work. | One peer reviewed process mining case and one systematic review of machine learning. |
| **Lean Six Sigma** DMAIC, data driven | Data analysis, dashboard preparation, drafting and generating possible root causes. | Vectara (May 2026) reports hallucination rates of 2% to 24% when summarising supplied material. Claims that AI will replace belt certified roles are untested. | Commentary, conceptual frameworks and a few cases; no controlled DMAIC study. |
| **Kaizen** employee led, daily | Better support for novices and knowledge capture could let more people solve problems. Practitioners frame AI through jidoka. | Adjacent studies raise concerns about skill development. Employees might withhold ideas that automate their roles, which is untested. | Inference from adjacent studies; no kaizen specific study. |

Source: Serevan and Kuşakçı (2026); Khakpour et al. (2025); Forrester; Vectara hallucination leaderboard, 11 May 2026; KPMG and University of Melbourne (2025); Bastani et al. (2025); Smalley (2026).

### Practitioner positions

The Lean Enterprise Institute (12 March 2026) positions AI as a coaching aid inside a management system run by people and warns against mistaking activity for progress. Bain (June 2026) calls automating a broken process AI's most expensive mistake. None of the 2025 to 2026 consultancy pieces reviewed quantifies the result of combining AI with Lean or Six Sigma or argues that AI replaces these methods.

EVIDENCE FROM SIX SECTORS

# Six sectors show one pattern: real task level gains, and no audited saving attributed to AI

Exhibit 13

## Six sectors at a glance: the grade shows the strongest source found, not the sector as a whole

### Manufacturing

**BEST SOURCE: STRONG**

Productivity falls after AI adoption, then recovers. At older plants, much of the early loss came from stopping regular KPI reviews (McElheran et al., 2025). Data mostly predate generative AI.

### Financial services

**BEST SOURCE: MODERATE**

US banks with more AI related hiring earn about 0.38 percentage points more on assets (FRBSF, 2026). Correlational, with job postings as a proxy.

### Telecoms

**BEST SOURCE: LIMITED**

Vodafone and BT report operational gains and large savings programmes in FY26. AI's share is not separated and nothing is audited.

### Healthcare

**BEST SOURCE: MODERATE**

A hospital AI pipeline analysed 500 patient records in about 30 minutes, whereas a manual review of 25 took about 100 hours (Vossler et al., 2026, preprint). In the MASAI trial, AI support cut radiologists' reading workload by 44%.

### Professional services

**BEST SOURCE: MODERATE**

Productivity rose by about 50% to 130% in five of six law tasks (Schwarcz et al., 2025). Accountants closed statements 7.5 days faster (Choi and Xie, 2025).

### Logistics

**BEST SOURCE: MODERATE**

Across about 9,300 automation projects at one firm, operating costs fell while labour cost rose (Yilmaz and Peukert, 2024). One warehouse case cut reception time about 40%.

Source: author's synthesis. Sample, design and limits for each source are in Appendix A (pages 18 and 19). Exhibit 17 grades each sector as a whole. Grades are the author's judgement.

Exhibit 14

## Financial benefits: no audited saving attributed to AI was found, only modelled, company reported or self reported figures

| TYPE OF CLAIM | EXAMPLE | FIGURE | STATUS |
|---|---|---|---|
| **Audited and attributed to AI** | None found in any sector. | Not applicable. | NONE FOUND |
| **Modelled** | GOV.UK Digital and Data Benefits framework, April 2026. | Potential annual savings of about €7.6bn (£6.3bn), of which about €1.3bn (£1.1bn) are cost reductions. | LIMITED Potential, not realised |
| **Company reported** | C.H. Robinson (2026); BT and Vodafone (page 18); practitioner example (page 14). | Productivity gains of more than 40%; savings programmes; about €6m (£5m) a year. | LIMITED |
| **Self reported time saved** | Bick, Blandin and Deming, St. Louis Fed, February 2025. | Users save 5% of their hours, 1% across all workers, an implied 1% aggregate productivity gain. | LIMITED |

Source: GOV.UK, Digital and Data Benefits framework, 7 April 2026; Bick, Blandin and Deming, Federal Reserve Bank of St. Louis, 27 February 2025; other sources as in Exhibit A2.

### Why the financial evidence is thin

Hours saved are not savings until they show up as lower spend, fewer hires or more output. Costs such as licences, data and maintenance are often left out, and only 26% of leaders say they see the full cost of AI (KPMG, Q2 2026). So reported benefits are usually gross, not net.

### Read across the sectors

The strongest evidence sits furthest from generative AI, and the newest evidence is often the least measured. The constant is the management system: firms with stronger management practices adopt technology more readily, while plants that let it slide tend to underperform. No example isolates an audited saving from AI. Page 17 shows why task level gains should not be mistaken for programme level savings.

REAL LIFE EXAMPLES

# Two real life examples show the same tool delivering very different results, depending on how it is deployed

**The question.** What do two practitioner accounts add to the evidence, and how far can they be relied on? They are a few snippets of real life, not a sample.

Exhibit 15

## An AI programme run as a large IT project overran and was scaled back, while AI used as one more improvement tool returned about ten times its cost

| | ELECTRICITY NETWORK AND GRID BUILDER | GOVERNMENT OUTSOURCING PROVIDER, SINGLE CONTRACT |
|---|---|---|
| **Scale** | Business turnover of about €14bn ($16bn). | Contract turnover of about €600m (£500m) or more, with profit of about €42m (£35m). |
| **How AI was used** | Major AI projects costing many millions, with the hope that AI would act as a silver bullet. | AI as another tool in the improvement armoury: a very powerful tool, but still one tool among others. |
| **How it was delivered** | As a large IT programme. | Process automation with AI, supported by the usual business change elements: behaviour change, process change and short interval controls. |
| **Result** | Cost overruns and late delivery. Not delivered as intended, so scaled back and implemented in phases, with the first phases bringing little benefit. | Annualised bottom line benefit of about €6m (£5m), a return of 10 to 1 on the investment. A side benefit was an AI capability built across the business on a yellow, green and black belt basis, so that the tool becomes part of all improvement work. |
| **Evidence grade** | LIMITED One practitioner account. | LIMITED One practitioner account. |

Source: two accounts shared with the author by a practitioner with direct experience of both, who was the IT project manager on the first. Organisations are anonymised and details are generalised. Figures are as reported by the practitioner and have not been independently verified. Euro figures are approximate conversions, with the original currency in brackets.

### Limited conclusions from two accounts

From these two direct experiences, AI is no different from any other tool. Deployed as a big IT project, it carries the risks of any change of IT system, such as a new ERP system. The risks are higher because no one yet knows what AI can and cannot do at its boundaries, and those boundaries are moving quickly. Used bottom up within a Lean or Six Sigma programme, focused on what AI is known to do well and badly and delivering within those limits, there is a very large amount of low hanging fruit.

## What these examples can support

That the way AI is deployed seems to matter as much as the technology. Both accounts are consistent with a pattern seen elsewhere in this report: results depend on the management system around AI, and on redesigning work before applying it.

## What they cannot support

Any claim about cause or about how common these outcomes are. There are two cases, from one source, with unaudited figures and no comparison group. The two organisations also differ in size, sector and scope, so the difference in result cannot be put down to the approach alone.

HOW THE HYPOTHESES STAND

# Hypothesis 1 is only partly supported and hypothesis 2 is confirmed

Exhibit 16

## Evaluation of the first hypothesis against the evidence on pages 7 to 13

| PART OF THE HYPOTHESIS | VERDICT | WHAT THE EVIDENCE SHOWS | WHAT PREVENTS A FIRMER VERDICT |
|---|---|---|---|
| **Faster** | SUPPORTED | Gains on bounded tasks in several studies, with the largest improvements among less experienced staff (Cui et al., 2025; Noy and Zhang, 2023). | The evidence relates to specific tasks rather than entire programmes. Experienced developers were 19% slower in one trial (METR, 2025). |
| **More relevant** | PARTLY | Adoption is widespread in some surveys, and practitioners consistently argue for its relevance. | Reported adoption rates vary considerably, ranging from 2% to 88%, depending on who is surveyed. |
| **More insightful** | LIMITED | A systematic review of 54 adjacent articles reports gains from machine learning in Lean manufacturing. | Validation was limited. No combined effect size was calculated, and none of the studies examined Lean methods. |
| **More effective** | NOT PROVEN | Manufacturing data show an initial decline in productivity followed by recovery where strong management routines are in place. | No controlled studies of Lean methods exist, no audited savings in telecoms have been reported, and sustainment was 12% before AI adoption. |

Source: author's synthesis of pages 5 to 13, Exhibits 13 and 14 and Appendix A.

### Verdict on hypothesis 1

Only partly supported. Speed is supported on bounded tasks, and relevance in part. Insight has limited support and effectiveness is not proven, because no controlled study tests AI inside the improvement methods.

Exhibit 17

## Hypothesis 2 tested: evidence across selected sectors is too limited to settle the first hypothesis

| EVIDENCE AREA | SECTOR GRADE | WHAT THE EVIDENCE SHOWS | WHAT IS MISSING |
|---|---|---|---|
| **Lean, Lean Six Sigma and Kaizen methods** | INSUFFICIENT | A systematic review of adjacent studies, a few case reports, a preprint and conceptual or vendor material. | No controlled peer reviewed studies measuring the baseline impact of generative AI within DMAIC, A3 problem solving, value stream mapping or kaizen activities. |
| **Manufacturing** | MODERATE | Census data on about 28,500 US plants show an initial productivity decline influenced by management routines. UK firm data suggest a positive association between management practices and technology adoption. | No recent data available to analyse: the evidence predates generative AI. Neither source examines continuous improvement methods directly or establishes causality. |
| **Financial services** | LIMITED | US banks with higher levels of AI related hiring reported returns on assets about 0.38 percentage points higher. UK firms also report widespread AI adoption. | No specific data available to analyse AI's impact on improvement work: the evidence is correlational or self reported. |
| **Telecoms** | LIMITED | Vodafone and BT report operational improvements and substantial cost reduction programmes, but it is not clear how far these are due to AI initiatives. | No independently audited data available to analyse: savings are company reported and not attributed specifically to AI. |
| **Healthcare, professional services and logistics** | LIMITED | Task level gains in trials and field studies. One hospital preprint ran a Lean analysis far faster. | No specific data available to analyse from inside an improvement programme, and no audited saving. |

Source: author's synthesis of pages 12 and 13, Exhibits 13 and 14 and Appendix A.

### Verdict on hypothesis 2

Confirmed. The strongest evidence, in manufacturing, predates generative AI. The newer evidence, in financial services and telecoms, is correlational or company reported, and in healthcare, professional services and logistics it is mostly task level.

RECOMMENDATIONS

# Five actions to prove both hypotheses, and four practices for using AI within a management system

Exhibit 18

## Five recommendations to strengthen the evidence supporting both hypotheses

| | RECOMMENDATION | ACCOUNTABLE ROLE | EXIT GATE | HYPOTHESIS IT SERVES |
|---|---|---|---|---|
| 1 | SENSE<br>**Redesign before applying AI**<br>Observe the work, remove steps and name one owner before any agent is built. | Process owner | Baseline signed off and owner named before funding. | H1: gains come from better processes, not faster bad ones. |
| 2 | SPEED STICK<br>**Test AI inside the methods, with a control**<br>Pilot AI within DMAIC, A3, value stream mapping and kaizen, against a baseline and a comparison group. | Improvement function lead | Results published regardless of outcome, including null findings. | H2: closes the biggest gap, no controlled evidence. |
| 3 | SPEED<br>**Start where the evidence is strongest**<br>Begin with bounded analysis, drafting and agent assisted tasks within a single process. | Process owner | Pilot demonstrates a sustained improvement against the baseline after rework. | H1: speed is proven only on bounded tasks. |
| 4 | SPEED STICK<br>**Measure the full cost and fund in stages**<br>Establish a baseline first, then release funding against verified net benefits at 12 and 36 months, measured as rework, capacity, service levels and customer outcomes, not hours alone. | Finance as benefits owner | Baseline, stopping rule and audit route agreed before funding. | H2: moves company reported gains towards audited ones. |
| 5 | STICK<br>**Develop capabilities and share the evidence**<br>Train staff to critically evaluate AI generated outputs, keep KPI reviews running, and publish results in a common format. | Improvement function lead | Staff competence assessed, governance routines established, and results shared externally. | H1 and H2: sustains gains and builds a body of evidence. |

Source: author's synthesis. Pairings and roles are a proposed model, not a tested finding.

Exhibit 19

## Four operating practices help AI work inside the management system

| | PRACTICE | WHAT TO DO | EVIDENCE BEHIND IT | ACCOUNTABLE ROLE |
|---|---|---|---|---|
| A | **Build AI into daily management** | Use AI to prepare KPI boards and flag deviations before tier meetings, while the discussion and the actions stay face to face. | Adjacent. Plants that dropped KPI reviews lost more productivity after adopting AI (McElheran et al., 2025). No study of AI inside tier meetings was found. | Site or function leader |
| B | **Apply AI to waste elimination first** | Use AI to find waiting, rework and unneeded steps and remove them, before automating what remains. | Practitioner position (Bain, 2026; Lean Enterprise Institute, 2026). No controlled study. | Process owner |
| C | **Upskill Lean and Six Sigma practitioners** | Add data engineering, process mining and the evaluation of AI output to belt training, alongside coaching and deliberate practice. | None found on training content. Technology alone had no effect in a garment factory trial, but did with people practices (Adhvaryu et al., 2026). | Improvement function lead |
| D | **Combine AI with the management system, problem solving and culture** | Treat AI as one tool inside the routines, problem solving and improvement culture that already sustain gains, not as a replacement for them. | Adjacent. Line managers' Lean leadership drove hospital performance, while Lean techniques did not (Van Kleeff and van Harten, 2025). | Executive sponsor |

Source: author's synthesis. The practices are a proposed model. The evidence behind each is adjacent or practitioner based, and none has been tested with AI.

LIMITATIONS OF THE EVIDENCE, AND CONCLUSION

# What the evidence supports, what it does not, and what comes next

### The evidence supports

AI can accelerate bounded analysis, drafting and agent assisted tasks, with the largest gains typically observed among less experienced staff.

That processes should be redesigned before AI is applied, a position argued consistently by practitioners and consultancies.

Governance, trust and human capability require explicit attention, but the evidence does not demonstrate that governance alone constrains implementation success.

### The evidence does not support

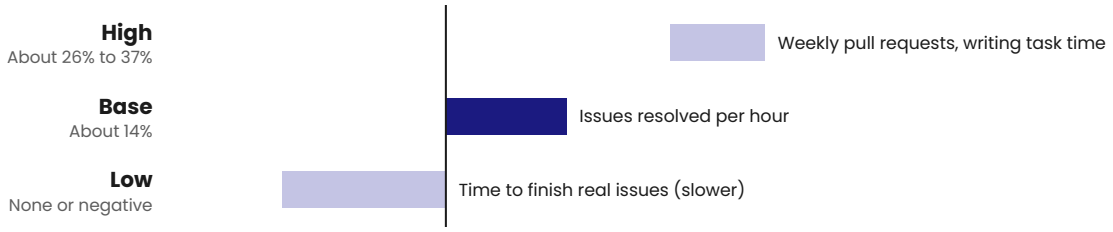
That AI has made continuous improvement the norm.

That AI has rendered Lean obsolete or revitalised it remains a proposition to be tested rather than a conclusion.

That AI improves or weakens the sustainability of gains remains uncertain, as the 12% baseline predates generative AI.

Exhibit 20

## Task level results show gross gains, not the net value a programme keeps




Illustrative only. Each bar reflects a different measure of task level gain. Sources: METR (2025); Brynjolfsson et al. (2025); Cui et al. (2025); Noy and Zhang (2023).

### From task gain to net value: measure outcomes, not hours

Hours saved are not savings until they appear in an outcome that can be measured: less rework, more capacity, faster service against agreed levels, or a better customer experience. Programme value needs four inputs: the gain on affected tasks, their share of programme work, the share of the gain kept after oversight and rework, and the full cost of running AI. According to Chen and Stratton (2026), across 718 firms commits per worker rose about 20% to 30% while resolved issues showed no detectable increase and review time rose 49%: more output, but also more rework. Few of the studies reviewed here measure these outcomes directly, and only 26% of leaders report full visibility of AI costs (KPMG, Q2 2026), so task level findings should not be presented as programme savings.

CONCLUSION

## What to do with the evidence

Speed on bounded tasks is no longer in doubt. What matters now is whether faster work becomes better, lasting improvement, and only controlled tests inside the methods can show that.

#### FOR LEADERS

Fix the process before automating it. Understand the work, assign clear accountability, and establish the full cost baseline before introducing AI agents. Combine AI with the management system, problem solving capability and improvement culture.

#### FOR IMPROVEMENT PROFESSIONALS

You are best placed to produce the missing evidence. Run controlled pilots within your own improvement methods and publish the findings, including null results.

#### FOR THE WIDER FIELD

Share results using a common reporting format, including those that show no improvement. This is how the second hypothesis can be tested, and either confirmed or rejected.

**The old discipline still decides the outcome. AI makes it faster to apply, but it does not make it optional.**

**About the research.** How the evidence was gathered, and its limits, are set out in Appendix B (page 20).

APPENDIX A1 · EVIDENCE TABLES

# Sector evidence is strongest in manufacturing, thinner in financial services and unaudited in telecoms

### Why these sectors and examples were selected

Manufacturing, financial services and telecoms were selected because they currently offer some of the strongest and most relevant evidence. They are illustrative, not representative of all sectors. We prioritised independent, large-scale evidence that links change to measurable improvement. We excluded vendor case studies, self-reported surveys, projected savings and research with unresolved data limitations.

Exhibit A1

## Six examples across the selected sectors, graded on what each can and cannot support

| SECTOR AND SOURCE | SAMPLE AND DESIGN | WHAT IT SHOWS | GRADE AND LIMITS |
|---|---|---|---|
| **Manufacturing** McElheran, Yang, Kroff and Brynjolfsson, 2025 | US Census data: about 28,500 plants, representing about 300,000, plus about 55,000 firms. Working paper. | Productivity falls after AI adoption, then recovers for firms that survive. At older plants, dropping structured management practices (KPI reviews, target awareness) explains a large share of the early loss, roughly a third to a half depending on the version. | **STRONG** on scale and relevance. Data run to 2021 and mostly cover non generative AI. Not peer reviewed, and the estimates differ between versions |
| **Manufacturing** ONS, 2025 | Official statistics: about 55,000 UK firms sampled, 2023 data. | Firms in the top tenth for management practices (continuous improvement, KPIs, targets) adopt technology at 88%, against 51% in the bottom tenth. AI use rises from about 2% to 10% across the same range. | **MODERATE** Association only. Excludes financial firms. AI showed no significant productivity link |
| **Financial services** FRBSF, September 2026 | 1,006 US banks, over 87% of banking assets. Central bank analysis. | Banks with more AI related hiring earn about 0.38 percentage points more on assets, and 0.61 for large banks. Small banks are largely absent: 84% never posted an AI job. | **MODERATE** Representative but correlational. Job postings are a proxy, and the letter gives no efficiency figures |
| **Financial services** Bank of England and FCA, 2024 | 118 UK firms. Regulator survey. | 75% use AI, up from 58% in 2022. 84% of AI users have an accountable person for their AI framework. Efficiency, productivity and cost gains are expected to grow most, but are expectations. | **LIMITED** Self-reported, small sample, 2024. No later edition found |
| **Telecoms** Vodafone, FY26, May 2026 | Company results presentation. | Virtual assistant resolves more than 70% of queries end to end. Network operations productivity up more than 10% a year. | **LIMITED** Operational measures, not verified savings. No separate AI figure for cost savings |
| **Telecoms** BT, FY26, 21 May 2026 | Company results release. | Savings target raised from about €3.6bn (£3bn) to about €4.4bn (£3.7bn), with about €1.8bn (£1.5bn) a year delivered. Workforce down 7% to about 108,000, linked to automation and AI. | **LIMITED** AI's share is not separated from other drivers, and the programme predates widespread agent deployment |

Source: McElheran et al., Industrial AI working paper, April and December 2025 versions; ONS, Management practices and the adoption of technology and AI in UK firms, 2025; FRBSF Economic Letter 2026-25; Bank of England and FCA, AI in UK financial services, 2024; Vodafone FY26 results; BT Group FY26 results release; City AM, 11 June 2026. Grades are the author's judgement.

APPENDIX A2 · EVIDENCE TABLES

# Healthcare, professional services and logistics show real task level gains, but no audited saving

Exhibit A2

## Three more sectors: the strongest sources found, and what each can support

| SECTOR AND SOURCE | SAMPLE AND DESIGN | WHAT IT SHOWS | GRADE AND LIMITS |
|---|---|---|---|
| **Healthcare** Vossler et al., 2026 | One US hospital. Looked back at past records; not yet peer reviewed. | A hospital team used AI to analyse 500 patient records in about 30 minutes. A manual review of 25 records had taken about 100 hours. The AI found every issue the manual Lean review had found, and some new ones. | LIMITED One site, labels not independently validated |
| **Healthcare** MASAI trial, The Lancet, 2026 | Randomised screening trial, about 106,000 women. | Radiologists spent 44% less time reading screening images with AI support, and 12% fewer cancers appeared between screening rounds (82 against 93). | MODERATE Strong design, but a clinical screening measure, not a Lean programme |
| **Professional services** Schwarcz et al., 2025 | Randomised trial, law students. University of Michigan working paper. | Productivity rose by about 50% to 130% in five of six tasks, and quality improved. | MODERATE Students, not practising lawyers. Working paper |
| **Professional services** Choi and Xie, 2025 | 277 accountants at 79 firms. Stanford working paper. | Accountants using AI closed their financial statements 7.5 days faster and spent 9% less time on routine work. A real change in how people work, but a working paper and not a Lean study. | MODERATE Working paper, publication not confirmed |
| **Logistics** Yilmaz and Peukert, 2024 | About 9,300 automation projects at one logistics firm, 2019 to 2021. Quasi experimental. | Revenue and profit rose and operating costs fell, while total labour cost rose. | MODERATE One firm. Robots and machine learning, not generative AI |
| **Logistics** Moica et al., 2025 | One Bosch warehouse process, ten reception cycles. | Total reception time fell about 40% (5,020 to 3,040 seconds), and savings were about €13,000 a year before infrastructure and maintenance. The paper headlines 79%, but its own value stream map shows about 40%. | LIMITED Five paired samples |
| **Logistics** C.H. Robinson, McKinsey, March 2026 | Interview with the chief executive. | Reports a productivity gain of more than 40% since the end of 2022 and more than 3 million shipping tasks automated, combining Lean practices with AI. | LIMITED Company reported |

Source: full references on page 21. Grades are the author's judgement. No audited distribution study was found.

APPENDIX B · DEFINITIONS AND ABOUT THE RESEARCH

# Definitions of terms used in this report

| TERM | WHAT IT MEANS |
|---|---|
| **A3** | A one page problem solving report, from Toyota, that sets out the problem, causes, countermeasures and follow up on a single sheet. |
| **Agent, agentic AI** | Software that plans and carries out several steps towards a goal with some autonomy, using tools and data from other systems. |
| **Benchmark** | A standard test used to compare AI systems. Scores from different benchmarks are not directly comparable. |
| **Bounded task** | A clearly defined piece of work, such as drafting a document or writing a function. |
| **DMAIC** | Define, Measure, Analyse, Improve, Control: the five step problem solving cycle used in Six Sigma. |
| **Generative AI** | AI that produces text, code, images or summaries from a prompt, usually as an assistant that a person works with. |
| **Gemba, go and see** | Gemba is the place where the work happens. Go and see means leaders and teams observe the real work there before drawing conclusions. |
| **Jidoka** | A Toyota principle: build quality into the process by stopping when a problem occurs. Often described as automation with a human touch. |
| **Kaizen** | Continuous, small scale improvement led by the people who do the work. |
| **Kata (Improvement Kata)** | A structured routine of practice and coaching in which teams move towards a target in small steps. |
| **Lean** | A management approach that delivers customer value by removing waste and improving the flow of work. |
| **Lean Six Sigma** | Combines Lean (waste and flow) with Six Sigma (reducing variation and defects using data). |
| **Machine learning** | Software that learns patterns from data to predict or classify, for example which machine is likely to fail. |
| **PDCA** | Plan, Do, Check, Act: a repeating cycle for testing a change and learning from the result. |
| **Preprint** | A research paper shared before independent peer review, so its findings are less certain. |
| **Process mining** | Using the event records in business systems to reconstruct how work actually flows. |
| **Randomised controlled trial** | A study that assigns people to groups at random, so that differences between groups can be attributed to the change being tested. |
| **Standard work** | The agreed best current way of doing a task, and the baseline for improving it. |
| **Sustainment** | Whether the gains from an improvement are still in place years later, with people still able to keep improving. |
| **Tier meeting** | A short, regular daily management meeting at each level of an organisation that reviews performance measures and problems. |
| **Value stream mapping** | A diagram of every step from request to delivery, showing where time and waste occur. |
| **Waste (muda)** | Any step that uses time or resources without adding value for the customer. |

## About the research

The search was targeted, not systematic, and six sectors were examined, not a full survey. Sources supplied by other research tools were traced to the original publication, and those that could not be traced were excluded. Many 2026 scores are vendor reported, and several studies are observational or preprints. One 2025 systematic review of AI for Lean systems (Emerging Science Journal) was identified but not assessed and carries no weight in the findings. Cost figures are in euros, with original pounds in brackets, and percentages are rounded. Two anonymised practitioner accounts (page 14) are illustrations and have not been independently verified. How the sectors were chosen is set out in Appendix A (pages 18 and 19).

SOURCES

# Where to check the evidence

Selected sources

## Documents cited in this report, grouped by topic


**Adoption and productivity**
McKinsey, The state of AI in 2023, 2025 and 2026
US Census Bureau, AI use in businesses, 26 May 2026
UK DSIT, AI Adoption Research Report, January 2026
Gartner, Hype Cycle for Agentic AI, 15 April 2026
Noy and Zhang, Science, 2023
Peng et al., arXiv 2302.06590, 2023
Brynjolfsson, Li and Raymond, NBER 31161
Dell'Acqua et al., SSRN 4573321
Cui et al., MIT working paper, 2025
METR, early 2025 developer study, 10 July 2025
METR, uplift update, 24 February 2026
METR, Time Horizon 1.1, 29 January 2026
Humlum and Vestergaard, NBER 33777, 2025
Acemoglu, NBER 32487, May 2024
Goldman Sachs, April 2023
Chen and Stratton, 2026
Murphy-Hill et al., arXiv 2607.01418, 2026
Autor et al., NBER 35720, 2026

**Sector evidence**
McElheran, Yang, Kroff and Brynjolfsson, Industrial AI, 2025
ONS, Management practices and technology adoption, 2025
FRBSF Economic Letter 2026-25, 21 September 2026
Bank of England and FCA, AI in UK financial services, 2024

**Agents, failure and return**
The Register on TheAgentCompany, 29 June 2025
Salesforce, CRMArena-Pro, arXiv 2505.18878
APEX-Agents, arXiv 2601.14242, 2026
Scale AI, Remote Labor Index leaderboard
OpenAI, Introducing GPT-5.5, 23 April 2026
Entrepreneur citing Bloomberg on Klarna, 9 May 2025
Information Age on CBA, 21 August 2025
Moffatt v. Air Canada, 2024 BCCRT 149
Vectara hallucination leaderboard, 11 May 2026
MIT NANDA, The GenAI Divide, July 2025
CIO Dive on S&P Global, 14 March 2025
RAND, RR-A2680-1, 13 August 2024
BCG, The Widening AI Value Gap, September 2025
BCG press release, 30 September 2026
Deloitte, State of Generative AI, 21 January 2025
KPMG, Global AI Pulse Q2 2026
Gartner, 25 June 2025
Forrester blog on automating before eliminating

**People and sustainment**
McKinsey, Today's good to great, 19 January 2024
Bastani et al., PNAS, 2025
Lee et al., CHI 2025
KPMG and University of Melbourne, 2025
Niederhoffer et al., Harvard Business Review, 22 September 2025
Capgemini Research Institute, 16 July 2025
Edelman Trust Barometer 2025
Brynjolfsson, Chandar and Chen, Stanford, 2025
Yale Budget Lab, 1 October 2025
Smalley, The Lean Post, 5 August 2026

**Lean, Lean Six Sigma and Kaizen**
Serevan and Kuşakçı, 2026
Khakpour et al., IJLSS, 2025
Căsăneanu Dascălu et al., 2025
Laureani, IGI Global, 2026
Menano de Carvalho et al., Sustainability, May 2026
Zhao, Fan and Li, engrXiv, June 2026
Lean Enterprise Institute, 12 March 2026
Bain, 15 June 2026
Two anonymised practitioner accounts shared with the author, 2026
Toyota newsroom, 22 May 2025; Toyota Times, 3 April 2026

**Further sectors and Lean elements**
Vossler et al., arXiv 2604.20055, 2026
MASAI trial, The Lancet, 30 January 2026
Schwarcz et al., Michigan Law WP 772, 2025
Choi and Xie, Stanford GSB, June 2025
Yilmaz and Peukert, CESifo WP 11321, 2024
Moica et al., Sustainability 17(20), 9081, 2025
McKinsey on C.H. Robinson, 20 March 2026
GOV.UK benefits framework, 7 April 2026
Bick, Blandin and Deming, St. Louis Fed, 2025
De Michele et al., arXiv 2504.06600, 2025
Shen and Tamkin, arXiv 2601.20245, 2026
Budzyń et al., Lancet Gastro Hepatol, 2025
Van Kleeff and van Harten, PPC, 2025
Adhvaryu et al., NBER 35445, July 2026
Lodgaard et al., IJLSS, 2023
Heaton, Lean Enterprise Institute, 27 Aug 2025

**Telecoms and cost transformation**
Vodafone FY26 results and transcript, May 2026
BT Group FY26 release, 21 May 2026
City AM, 11 June 2026



### Acknowledgements

Thanks to my critical friends, who read earlier drafts, questioned the evidence and sharpened the argument. Their comments improved this report. Any errors, and all views expressed, remain my own, and their help does not imply endorsement.

Moore Alison
Simon Veer
Tom Candy
Asim Khan
Kyaw Win Ko
Mike Keen
Magda Cola
Simon Buckley
Parikshit Naik
Duncan Martin

Why Should I Care?

2026
Copyright © Claudio Costa
All rights reserved.
whyshouldicare.net

# Speed, Sense and Stick

AI can speed up the work of improvement. The proof that it makes improvement better is still to come.

**Disclaimer.** This report is my own independent research, based on information that is available and researchable online. It reflects my personal analysis only, and does not reflect the views or opinions of any company, or of any public or private organisation, including any I work with or am affiliated with, or any named in it.

**Claudio Costa**