THE ENTERPRISE AI SKILLS MAP

Building Workforce Readiness for Generative AI

31 MAY 2023 • ENTERPRISE AI SERIES

Paul Forrest

PAPER THREE · ENTERPRISE AI, 2023 TO 2024

About this paper

This is the third of a planned nine papers on the enterprise adoption of generative artificial intelligence.

The first two papers treated adoption as a learning problem and capability as something an individual acquires. This one treats it as something an organisation designs, which raises a different set of questions about who needs what, and who decides.

Two surveys and two field studies carry the evidence here. The four-layer stack, the population weighting and the assessment measures are mine, and each carries a tag saying so where it appears.

HOW TO CITE

> Forrest, P. (2023). *The Enterprise AI Skills Map: Building Workforce Readiness for Generative AI.* Enterprise AI, 2023 to 2024, Paper Three. Version 1.0, 31 May 2023.

HOW TO READ THE EVIDENCE TAGS

Every substantive claim carries a tag stating what kind of evidence stands behind it. A claim resting on two kinds carries both.

PREPRINT Research posted as a preprint or working paper, not peer-reviewed at this date.

SURVEY Survey or polling data, cited with its sample and fieldwork dates.

AUTHOR My own framework or judgement, proposed and not yet proven.

A NOTE ON DATES

Research for this paper closed on 26 May 2023. The two surveys that carry most of its weight were published on 30 April and 9 May, which is why it is dated at the end of the month.

Contents

ABSTRACT AND SUMMARY OF THE ARGUMENT

What employers say they need, what the surveys actually measured, and what follows

Two substantial pieces of survey evidence arrived in the last month, and between them they look likely to set the terms of the workforce discussion for the rest of the year.[1,2] SURVEY Both are worth taking seriously. Both are also being quoted in ways their own methodology sections do not support, and the gap between the two is where this paper begins.

One headline has travelled further than the rest. It concerns 23 per cent of jobs changing by 2027. That figure describes structural churn, meaning jobs created plus jobs eliminated, against a projected net decline of about 2 per cent.[1] SURVEY Quoted as though it forecast a quarter of roles disappearing, it supports a conclusion the report does not reach.

What the evidence does support is narrower and more useful. Employers expect a large share of existing skills to be disrupted, they rank analytical and creative thinking above technical artificial intelligence skills in their training plans, and the largest organisations rank AI skills first.[1] SURVEY This paper builds a four-layer capability model on that footing, maps it across four workforce populations, and proposes a 90-day approach that starts from tasks.

WHAT THIS PAPER ARGUES

The scarce capability is the combination of machine fluency with domain judgement and responsible execution. Organisations are currently buying the first of the three, because it is the only one a vendor can supply.

- The survey headline needs reading correctly. Structural churn of 23 per cent is growth of 10.2 per cent and decline of 12.3 per cent, with 69 million roles created and 83 million displaced for a net fall of 14 million.[1] SURVEY
- The skills employers name first are cognitive rather than technical. Analytical thinking leads company training priorities at 48 per cent and creative thinking follows at 43 per cent, with AI and big data third at 42 per cent, rising to first among organisations of more than 50,000 people.[1] SURVEY
- Capability is best built in four layers, which are machine fluency, task craft, critical judgement, and governance and data discipline. The first is the cheapest to teach and the least durable. AUTHOR
- Tasks should be mapped before roles are rewritten. A job description rewritten in 2023 to accommodate a technology six months old will be rewritten again, and the exercise consumes political capital that the task work needs.
- Assessment should measure transfer rather than completion. Course completion measures attendance, while verified task quality, escalation behaviour and privacy compliance measure whether anything changed.

SECTION 1

What the last month of survey evidence actually says

Two reports, both large, both sponsored, and both quoted more loosely than they deserve. The numbers are worth stating properly before anything is built on them.

The World Economic Forum published its Future of Jobs Report on 30 April 2023, drawing on a survey of 803 companies across 27 industry clusters and 45 economies, together employing more than 11.3 million workers.[1] SURVEY Nine days later Microsoft published a work trend report based on a survey of 31,000 full-time employed or self-employed people across 31 markets, conducted by an independent research firm between 1 February and 14 March 2023.[2] SURVEY Both are employer-facing or vendor-sponsored, and both should be read with that in view.

The churn figure, stated correctly

The Forum reports that employers anticipate structural labour market churn of 23 per cent of jobs over the next five years. That figure is the sum of two movements, growth of 10.2 per cent and decline of 12.3 per cent, which nets to a projected fall of about 2 per cent, or 69 million roles created against 83 million displaced.[1] SURVEY It is a survey of employer expectation, and it describes movement in both directions.

Churn is not loss, and the difference changes the response. An organisation planning for a quarter of its roles to vanish will cut. An organisation planning for a quarter of its roles to change composition will redesign work and move people, which is harder, slower and considerably more likely to be what the evidence describes.

The companion figure is the one for a board. Employers expect 44 per cent of skills to be disrupted within five years.[1] SURVEY That is a claim about the content of jobs rather than their number, and it points at training and task redesign rather than at headcount.

THE 23 PER CENT, DECOMPOSED

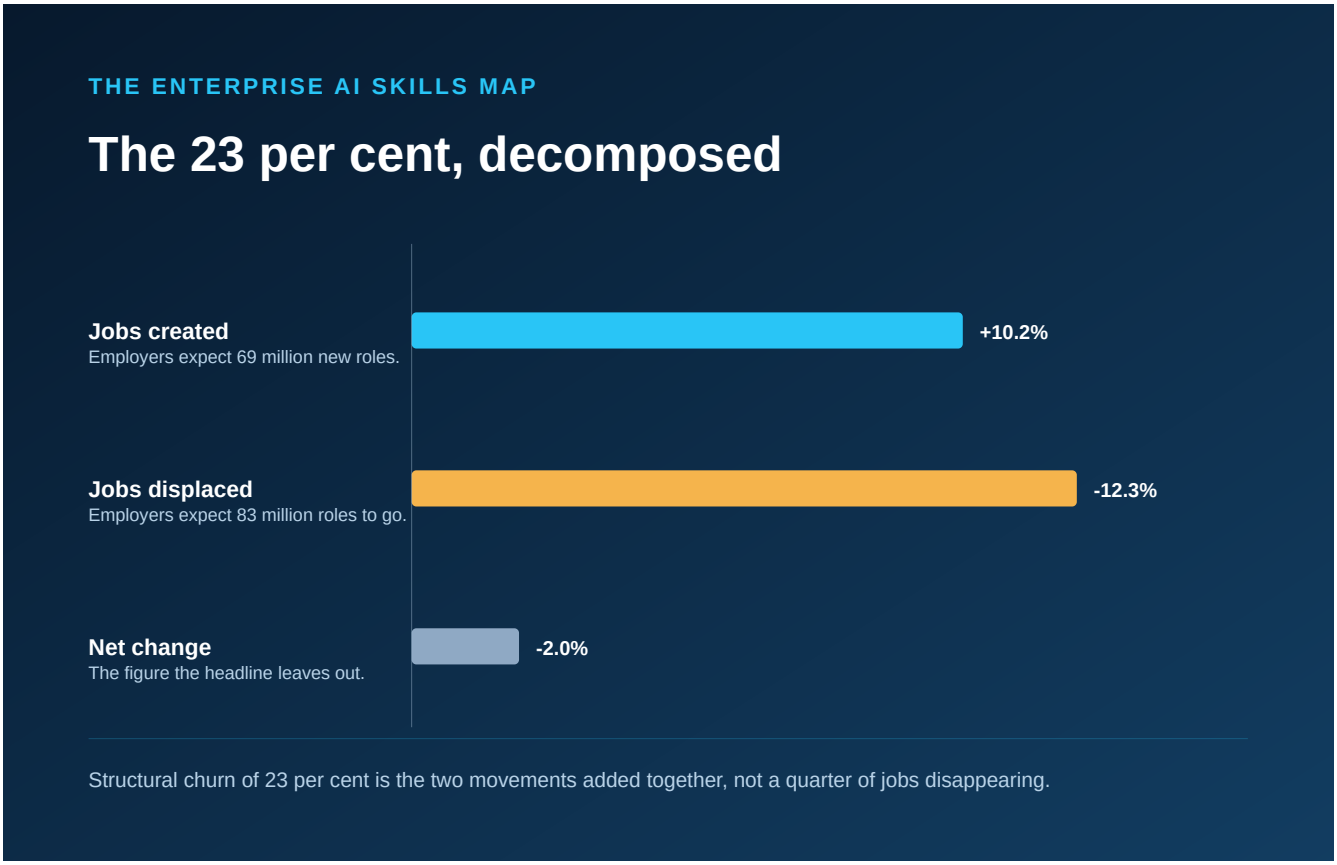

FIGURE 1 The figure that travelled describes structural churn, which is job creation and job destruction added together. The net expectation is a decline of about 2 per cent, and the two movements are what the headline conceals.

World Economic Forum, Future of Jobs Report 2023. SURVEY

Which skills employers say they will train

The ranking here is frequently reported backwards. In company training priorities from now to 2027, analytical thinking leads at 48 per cent of surveyed organisations, creative thinking follows at 43 per cent, and artificial intelligence and big data comes third at 42 per cent.[1] SURVEY The Forum's own framing of the contrast is worth quoting, because it is more precise than most of the commentary built on it.

> *While AI and big data ranks only 15th as a core skill for mass employment today, it is the number three priority in company training strategies from now until 2027, and number one priority for companies with more than 50,000 employees.*
>
> WORLD ECONOMIC FORUM, FUTURE OF JOBS REPORT 2023

Two things follow. Large organisations are moving first and hardest on technical capability, which fits the pattern of every prior technology cycle and says nothing about whether they are right. And the skills employers rank above it are cognitive, which is the finding this paper's capability stack is built around.

The second survey, and what it does not contain

The Work Trend Index report identifies analytical judgement, flexibility and emotional intelligence at the top of the skills leaders believe employees will need, with intellectual curiosity, bias detection and the ability to delegate to and prompt a system following.[2] SURVEY It publishes this as a ranked list. It does not attach a percentage to each skill, and figures that circulate in that form have been supplied by somebody other than the authors.

The report's one hard number on this point is that 82 per cent of leaders say employees will need new skills to prepare for the growth of artificial intelligence.[2] SURVEY It is a statement about expectation from a company that sells the relevant products, and it should be attributed on both counts.

SECTION 2

A capability stack in four layers

Machine fluency, task craft, critical judgement, governance. Organisations buy the first because it is the only one with a supplier.

Most of what is currently sold as artificial intelligence training addresses the bottom layer, and that layer will be worth least in 18 months. The reason is structural. Machine fluency can be taught by somebody who does not know your business, in a room, to 100 people at once. Nothing above it can.

WHAT EACH LAYER COSTS TO BUILD, AND HOW LONG IT LASTS

LAYER	WHAT IT CONSISTS OF	TIME TO COMPETENCE	DURABILITY
Machine fluency	Prompt construction, iteration, knowing what the class of system does and does not do.	Hours to days.	Low. Product changes obsolete specific technique within months.
Task craft	Knowing which parts of your own work the system handles and where it fails in your domain.	Weeks of supervised practice on real work.	Moderate. The domain knowledge persists, the specific failure list does not.
Critical judgement	Verification discipline, source checking, recognising a confident wrong answer.	Months, and it builds on existing professional expertise.	High. It transfers across systems and across roles.
Governance and data discipline	Data boundaries, retention, disclosure, escalation, and knowing when to refuse.	Days to establish, continuous to maintain.	High, provided the organisation keeps it current.

Why the bottom layer decays

Specific prompting technique is tied to the behaviour of a specific model generation, and the models are changing on a scale of months. A course built in March around one system's quirks will be partly wrong by the autumn. That is no argument against teaching it, since people must start somewhere. It is an argument against building the programme's centre of gravity there, and against measuring the programme by how many people completed it in the first year.

Why the third layer matters most

Critical judgement is the capacity to notice that a confident output is wrong, and it is largely a function of domain expertise that the organisation already has. What the organisation usually lacks is the habit of applying it to machine output, because machine output arrives in a register that professionals have been trained to read as authoritative. Formatted, fluent, free of hedging, and wrong in a way surface reading never catches. AUTHOR

The evidence on where the gains concentrate is suggestive. A study of customer support agents circulated in April reports an average increase of around 14 per cent in issues resolved per hour, with the greatest effect among novice and lower-skilled workers.[3] PREPRINT A study of professional writing tasks from March reports a similar pattern, with the largest gains among initially weaker writers.[4] PREPRINT If that pattern holds, the technology compresses the distance between novice and expert on well-defined tasks, and the value of expertise migrates towards the judgement layer.

A CAUTION ABOUT THAT INFERENCE

Both studies examine tasks with clear quality criteria and fast feedback, and neither is about the professions. Neither examines work where the error surfaces months later. Extending the novice-compression finding to that kind of work is an argument, and this paper makes it as one.

SECTION 3

Four populations with different obligations

One programme for everybody costs more and teaches less. The weighting across populations is what makes the thing affordable.

The instinct in most organisations is to define a single curriculum and run everyone through it, which has the advantage of fairness and the disadvantage of being wrong for all four groups simultaneously. The obligations differ, so the time should differ.

FOUR POPULATIONS, FOUR OBLIGATIONS

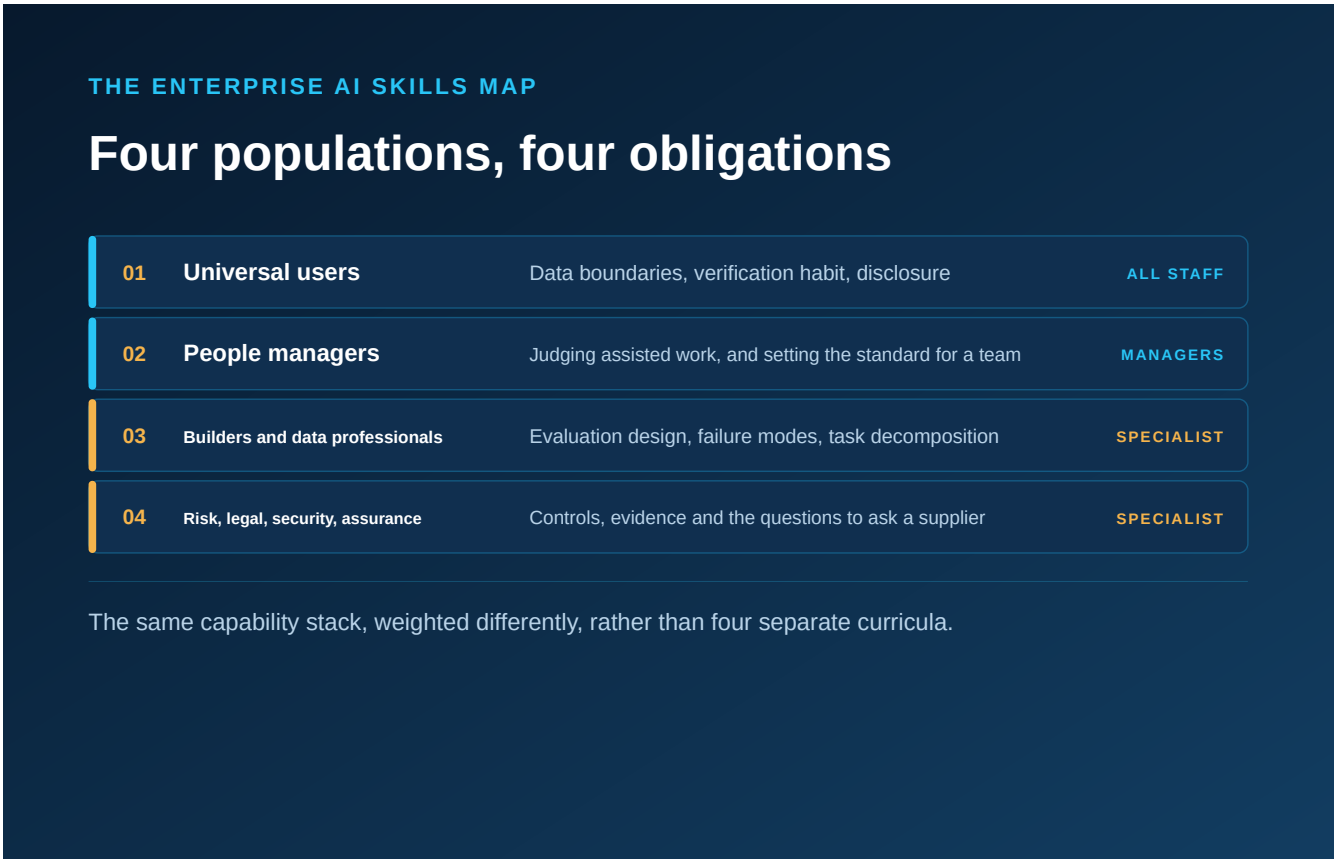

FIGURE 2 The same capability stack, weighted differently for universal users, people managers, builders and data professionals, and the risk and assurance functions. The weighting is what makes a programme affordable.

My framework. AUTHOR

FOUR POPULATIONS, AND WHAT EACH ACTUALLY NEEDS

POPULATION	PRIMARY OBLIGATION	WHERE THE TRAINING TIME GOES	WHAT FAILURE LOOKS LIKE
Universal users	Know what may not be entered, and know that fluent output can be wrong.	Two hours, once, plus a domain error catalogue kept current by somebody else.	Confidential material in a prompt, or an unverified output used as though checked.
People managers	Redesign the task, not the person. Set the verification rung and defend the time it costs.	A day, concentrated on workload design and on how to review assisted work.	A team told to use the tools and judged on the same deadlines as before.
Builders and data professionals	Know the data boundary, the retention terms and what the system can be made to reveal.	Several days, technical, including adversarial testing of their own build.	A working prototype connected to data nobody approved, demonstrated to an executive.
Risk, legal, security, assurance	Assess a probabilistic system without pretending it is deterministic software.	Several days, on evaluation method rather than on tooling.	A control framework that reads well and specifies nothing observable.

The population that is usually underserved

People managers receive the least attention and carry the most consequence. They set the workload, decide what review is affordable, and answer for the output. A manager who has been told that the tools save time and has not been told that verification costs time will absorb the saving into the deadline, which removes the check that made the saving safe. That is a design error, committed at the level above the person doing the work. AUTHOR

The population that is usually overserved

Universal users need less than they are given. Two hours covering the data boundary, the failure modes, the escalation route and a handful of domain examples will do more than a day of prompting technique, provided somebody maintains the domain examples. The maintenance is the expensive part and it is where the budget should go.

SECTION 4

Map tasks before touching roles

Job architecture is the slowest thing in any organisation to change and the most expensive to change twice.

There is considerable pressure this quarter to rewrite job descriptions, create new titles and publish a revised competency framework. It should be resisted for at least two more quarters, on the practical ground that a role definition built around a technology six months old will be rebuilt, and the second rebuild will meet more resistance than the first.

Work at the task level instead. A task is small enough to observe, specific enough to baseline, and politically cheap to change. The output of the exercise is a list of tasks that changed, by how much, with what verification, which is exactly the input a role redesign will need when the time comes to do it properly.

SPECIFICATION

How to run a task inventory in one function

- Take one function and list the 20 or so recurring tasks that consume most of its time. A sample of 20 is enough to work with and few enough to finish.
- For each, record what it currently costs in time and rework, taken from completed work.
- Mark which of the five filter tests from the first paper each task passes, and in particular whether the person receiving the output can tell that it is wrong.
- Record who currently does it, at what grade, and what they would do with the time if the task shortened. An unanswered version of that question is how a productivity gain turns into an unfilled vacancy nobody decided on.

The question the inventory forces

Somewhere in the second week a team will find a task that the technology handles well and that exists mainly to train junior staff. Document review, first-draft analysis, routine correspondence. Automating it produces a measurable saving and removes the route by which people acquire the judgement the third layer depends on. I do not have a general answer to this and I am suspicious of anyone who offers one quickly. It should at least be a decision somebody takes on purpose. AUTHOR

SECTION 5

A learning architecture that produces evidence

The architecture has four components, and one of them is not training at all.

The purpose of the architecture is to produce two things at once. Capability in people, which is what it is funded for, and evidence about the organisation's own failure modes, so that the next cycle is cheaper.

FROM LEARNING TO VERIFIED VALUE

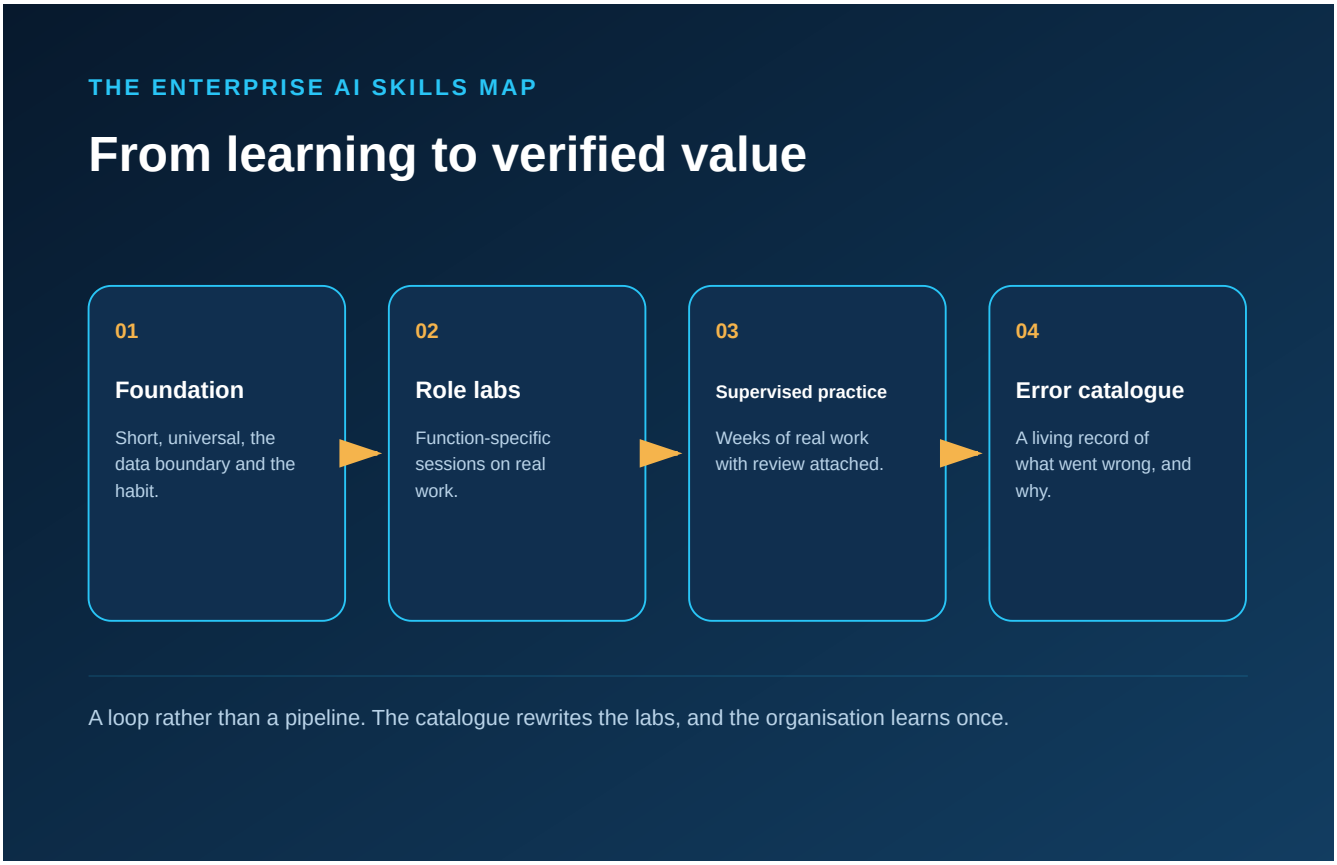

FIGURE 3 A loop rather than a pipeline. Evidence from supervised practice feeds back into the curriculum, and the organisation learns what its own failure modes are.

My framework. AUTHOR

1. A foundation course, short and universal, covering the data boundary and the failure modes. It is delivered once and refreshed when the product or the policy changes.
2. Role labs, meaning function-specific sessions on real work, with the verification rung named and the baseline kept. This is where the budget belongs.
3. Supervised practice, meaning several weeks of real work with review. No other component reliably produces the second and third layers of the stack.

4. A maintained error catalogue, which is a living document and the one component here that is not training. It records how these systems fail on this organisation's work, it is owned by a named person, and it is fed by the labs and the practice period.

The fourth component gets cut first, because it has no completion metric and no vendor. It is also the only artefact from the whole programme that will still be useful in two years, since it accumulates knowledge specific to the organisation. A good catalogue beats good training. AUTHOR

On communities of practice

They work where somebody is paid to run them and decay where they depend on goodwill. Budget a named facilitator with allocated time, or do not start one, because a dormant channel with three messages from March is worse than no channel at all.

SECTION 6

Assessment that survives contact with the organisation

Course completion tells you who attended. Four other measures tell you whether the work changed.

Learning functions measure completion because completion is easy to collect and because it is what the system was built to report. It measures attendance. Everything worth knowing requires looking at work.

1. Verified task quality, meaning a sample of assisted outputs reviewed against the standard that applied before, by somebody other than the person who produced it.
2. Escalation behaviour, meaning how often people escalated instead of guessing, and whether the escalations were appropriate. A rate of zero is a finding and rarely a good one.
3. Data discipline, meaning observed compliance with the boundary rule, tested by asking people what they would do with a realistic borderline case. Asking whether they know the policy tests memory.
4. Transfer, meaning whether the practice appears in work three months later without a facilitator present. This is the measure that matters and the one almost nobody collects.

Of the four, escalation behaviour is the most informative early and the least intuitive. An organisation whose people never escalate has either designed a workflow with no route for it, or created a culture in which admitting uncertainty about a machine output is read as incompetence. Both are fixable, and neither is visible in a completion report. AUTHOR

What not to measure yet

Productivity gain attributable to training. The chain of inference from a course to a business outcome runs through too many uncontrolled steps, and any figure produced this year will be an attribution exercise. Where a sponsor requires one, offer the task-level baseline from the inventory instead. It is a real number about a real task and it makes no claim about causation.

SECTION 7

The next 90 days

Inventory, prioritise, pilot, credential, measure. The credential step is optional and the inventory is not.

This assumes one function, a named owner, and no new budget beyond the time of the people involved. An organisation that cannot assemble those has a governance problem rather than a skills problem, and should wait for the next paper in this series.

Days 1 to 30, inventory and boundary

Run the task inventory in the chosen function. Publish the data boundary rule if it does not already exist, in one page, with a named owner. Identify the four populations inside that function and size them, which usually reveals that the manager population is larger than anyone assumed and has been treated as universal users.

Days 31 to 60, labs and supervised practice

Run role labs on the three highest-value task classes that qualify. Begin supervised practice immediately afterwards, because the gap between a lab and its application is where the learning is lost. Start the error catalogue in the first week of practice, and give it to somebody whose job it becomes. A volunteer will stop.

Days 61 to 90, assess and decide

Assess against verified task quality and escalation behaviour. Decide whether to extend to a second function, hold, or stop. Write down what the error catalogue contains after 60 days, because that document is the deliverable that transfers, and a programme that reaches day 90 with an empty catalogue has been running training.

> **ON CREDENTIALS AND BADGES**
>
> Internal credentials are a reasonable motivational device and a poor measure. They certify that somebody passed an assessment at a point in time against a system that has since changed.
>
> Where they are used, tie them to observed work and give them an expiry. A badge earned in May 2023 against the current generation of tools should not still be current in May 2024 without reassessment, and saying so at the outset avoids an argument later.

Whether any of this holds up depends on a question the available evidence cannot yet answer, which is whether the capability that matters turns out to be durable or whether successive product generations absorb it. The honest position at the end of May is that nobody knows, and that building the judgement layer is the bet that loses least if the answer goes either way.

END MATTER

Method, evidence and limits

How the sources in this paper were chosen and dated, and what the reader should distrust.

Selection

Two large surveys carry most of the empirical weight and both are cited with population, geography, fieldwork window and sponsor attached. Where a survey publishes a ranked list without percentages, this paper reproduces the ranking and does not invent the percentages.

Dating

Every reference carries its original publication date, and the two central surveys carry their fieldwork windows as well.

On sponsored research

One of the two central sources is published by a company that sells products in this market, and the other by an organisation whose members are large employers. Neither fact disqualifies them. Both facts belong beside the numbers, and the text attributes accordingly.

On preprints and working papers

The two field studies of productivity effects were, at the date of this paper, a National Bureau of Economic Research working paper and an unpublished working paper. They carry the preprint tag, which records status and says nothing about quality. Neither had been through peer review, and the argument in section four does not rest on either alone.

What the evidence does not yet cover

No evidence exists on whether capability built now persists across model generations, and that is the central practical question for anyone budgeting a programme. No published assessment instrument for these skills could be located. And no study addresses the junior-task displacement problem raised in section four, which is the issue most likely to matter in five years.

What is mine, and therefore untested

The four-layer stack, the four-population weighting, the durability estimates in the layer table and the assessment measures are my own. The durability column in particular is a judgement, and it is offered so that a reader can disagree with something specific.

Interest declared

I design programmes of the sort this paper describes and would benefit if organisations bought them. The paper argues for smaller programmes than are currently being sold and against the enterprise-wide curriculum that is easiest to deliver, so the bias, if it operates, runs against my own revenue.

END MATTER

Sources considered and set aside

Sources that circulated this spring and did not make it into the argument, and why.

SOURCE	PUBLISHED	WHY IT WAS SET ASIDE
Resume Builder, survey of business leaders on ChatGPT use	February 2023	Widely quoted for the claim that half of surveyed companies were using the tool. The sample is an online panel of 1,000 leaders recruited by a jobs website, and the write-up gives the questions only in summary.
Pew Research Center, A majority of Americans have heard of ChatGPT, but few have tried it themselves	24 May 2023	A consumer awareness survey, United States only, fielded in March. It says nothing about work.
Deloitte, State of AI in the Enterprise, fifth edition	18 October 2022	Fieldwork before the November release, on a definition of artificial intelligence that does not include the systems this paper is about.
KPMG, Generative AI survey of United States executives	April 2023	A survey of 225 executives about intentions. It measures what leaders expect to do, and their people are doing something else.
Microsoft, Work Trend Index 2022	16 March 2022	The previous edition of the survey used here. It predates the tools and is useful only as a baseline for the hybrid work questions, which this paper does not use.

END MATTER

About the author

Paul Forrest advises boards and executive teams on the adoption of new technology in operating businesses, with a practice grounded in delivery rather than in advisory work alone.

His background is in business process reengineering, which he has practised since the 1990s, and in applied natural language processing, which he has worked on since 2014. The task-level bias in this paper comes from the first of those. Reengineering work taught that roles are the last thing to change and the most expensive to change badly, and that the organisations which start with the org chart usually end up rebuilding it twice.

He writes on the condition that a claim carries its evidence, and this series is an attempt to hold that line while a subject moves quickly.

END MATTER

References

Four works cited in the text, each with its original date of publication. None is later than 26 May 2023.

1. **World Economic Forum.** *The Future of Jobs Report 2023.* Employer survey. Cited for structural churn of 23 per cent, growth of 10.2 per cent against decline of 12.3 per cent, 69 million roles created and 83 million displaced, 44 per cent of skills disrupted, and the training priority ranking.
   PUBLISHED 30 APRIL 2023

2. **Microsoft.** *Will AI Fix Work?* Work Trend Index annual report. Survey of 31,000 full-time employed or self-employed people across 31 markets, conducted by Edelman Data x Intelligence, fieldwork 1 February to 14 March 2023.
   PUBLISHED 9 MAY 2023

3. **Brynjolfsson, E., Li, D., and Raymond, L.** *Generative AI at Work.* National Bureau of Economic Research working paper 31161. Customer support agents, average increase of around 14 per cent in issues resolved per hour, greatest effect among novice and lower-skilled workers.
   PUBLISHED APRIL 2023

4. **Noy, S., and Zhang, W.** *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence.* Working paper, cited in that form.
   PUBLISHED 2 MARCH 2023

END MATTER

Further sources consulted

Nine sources I read for this paper and did not cite. They are listed so that the reading behind it can be seen in full, and none is later than 26 May 2023.

- **Eloundou, T., Manning, S., Mishkin, P., and Rock, D.** *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.* arXiv:2303.10130. Around 80 per cent of the United States workforce with at least 10 per cent of tasks affected, 19 per cent with at least half.
  PUBLISHED 17 MARCH 2023
- **Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M.** *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot.* arXiv:2302.06590.
  PUBLISHED 13 FEBRUARY 2023
- **Stanford Institute for Human-Centered Artificial Intelligence.** *Artificial Intelligence Index Report 2023.*
  PUBLISHED 3 APRIL 2023
- **Goldman Sachs.** *Generative AI could raise global GDP by 7%.* Cited as a bank's published projection rather than as a finding.
  PUBLISHED 5 APRIL 2023
- **OpenAI.** *GPT-4.* Release announcement.
  PUBLISHED 14 MARCH 2023
- **National Institute of Standards and Technology.** *Artificial Intelligence Risk Management Framework 1.0.* NIST AI 100-1.
  PUBLISHED 26 JANUARY 2023
- **International Organization for Standardization and International Electrotechnical Commission.** *ISO/IEC 23894, Guidance on risk management.*
  PUBLISHED 6 FEBRUARY 2023
- **Department for Science, Innovation and Technology.** *A pro-innovation approach to AI regulation.* Command paper.
  PUBLISHED 29 MARCH 2023
- **World Economic Forum.** *The Future of Jobs Report 2020.* Cited for comparison of employer expectations across editions.
  PUBLISHED 20 OCTOBER 2020

WHERE THIS GOES NEXT

The next paper takes the portfolio question

This paper assumed that an organisation had decided what to attempt and was asking who needed to be able to do it. The fourth will take the prior question, which is how scattered experiments become a governed portfolio with stage gates and a measurable value case, and it is due at the end of August.

The international labour bodies publish their annual employment outlooks over the summer, and I expect the next paper to lean on that evidence more than this one could.

BEFORE THE NEXT PAPER, THREE THINGS

- A task inventory for one function, with baselines taken from completed work.
- An error catalogue with a named owner and something actually in it.
- A sized view of the four populations, and a manager cohort that has been trained as managers rather than as users.

Paul Forrest · 31 May 2023 · Version 1.0 · Research closed 26 May 2023