GUARDRAILS BEFORE SCALE

An Enterprise Governance Blueprint for Generative AI

30 NOVEMBER 2023 • ENTERPRISE AI SERIES

Paul Forrest

PAPER FIVE · ENTERPRISE AI, 2023 TO 2024

About this paper

This is paper five in a series of nine on the enterprise adoption of generative artificial intelligence.

The fourth paper treated controls as a condition of scaling. This one treats them as the subject, and it arrives at the end of the most eventful quarter for artificial intelligence policy since the technology reached the public.

Most of what follows is a reading of primary instruments, which are short enough that I would rather the reader went to them than trusted my summary. The tiering, the control list and the decision rights are my own and are tagged as such.

HOW TO CITE

> Forrest, P. (2023). *Guardrails Before Scale: An Enterprise Governance Blueprint for Generative AI.* Enterprise AI, 2023 to 2024, Paper Five. Version 1.0, 30 November 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.

PRIMARY An institutional or official document, cited from the issuing body.

VENDOR A vendor announcement or a figure the vendor reports about its own product.

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

LEGAL A statute, regulation, executive order, legislative proposal or court decision.

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

A NOTE ON DATES

Research for this paper closed on 27 November 2023. The European negotiation was still open at that date, and the text describes it as open.

LEGAL AND REGULATORY NOTICE

This paper describes executive action, international statements, voluntary frameworks and legislative proposals, principally in sections one, two and six. Nothing in this document constitutes legal advice. The provision of legal advice sits outside the terms of any engagement with me, and any question about a particular organisation's obligations is a matter for its own qualified advisers.

Statements about what an instrument says are attributed to that instrument. Where I draw a practical implication the instrument does not itself state, the text says so. Legal status is described precisely, because proposed, agreed, adopted, in force and applicable are different states and the distinction carries practical consequences.

Contents

ABSTRACT AND SUMMARY OF THE ARGUMENT

A month in which the policy position changed, and what an enterprise should do about it

Between 30 October and 1 November 2023 three things happened that a governance function must absorb. A United States executive order was issued, the Group of Seven published voluntary instruments for developers of advanced systems, and 28 countries and the European Union signed a political declaration on frontier risk.[1,2,3] LEGAL / PRIMARY None of the three imposes an obligation on an ordinary enterprise deploying an assistant, and all three change the environment in which such an enterprise will be asked to explain itself.

This paper converts that environment into an operating blueprint. It covers risk tiering, lifecycle controls, decision rights, supplier assurance and a 90-day launch. The organising claim is that guardrails are a set of decisions with evidence attached, renewed on a cycle, and that a policy document produces none of them on its own.

The risks themselves are not conventional software risks. Probabilistic output, prompt injection through untrusted content, leakage of material the user never intended to disclose, a supplier changing a model without notice, and human overreliance on fluent text are the five that most commonly defeat controls designed for deterministic systems. AUTHOR

WHAT THIS PAPER ARGUES

Guardrails are an operating system of decisions, evidence and review cycles. An acceptable use policy is one artefact of that system, and organisations that produce only the artefact have documented an intention.

- The executive order should be read for what it is. It directs federal agencies, and its principal reach into the private sector is a reporting duty on developers of the largest dual-use models and operators of large computing clusters.[1] LEGAL An enterprise deploying a copilot acquires no obligation from it.
- The international instruments address developers. The Group of Seven principles and code of conduct published on 30 October are voluntary and are addressed to organisations developing the most advanced systems.[2] PRIMARY A deploying enterprise sits outside their stated scope, which matters before anyone adopts them as a compliance baseline.
- Systems should be tiered by consequence and reversibility. Four tiers run from prohibited through high-consequence and controlled productivity to low-risk experimental, with the tier set by what happens when the output is wrong rather than by which technology was used. AUTHOR
- Govern is a condition rather than a stage. In the framework most organisations are adopting, three functions describe work done on a system and the fourth describes the conditions under which that work happens at all.[4] PRIMARY Treating it as step one of four produces a completed step and an unowned result.
- Supplier assurance is where the bargaining power sits. Change notice, testing access, logging, data terms and an exit plan are negotiable now in a way they will not be once the deployment is embedded.

SECTION 1

The month the policy environment moved

Three instruments in three days. Their scope is narrower than the coverage suggests, and knowing the scope is the first governance task.

On 30 October 2023 the President of the United States issued an executive order on safe, secure and trustworthy artificial intelligence, published in the Federal Register on 1 November.[1] LEGAL The same day the Group of Seven published a leaders' statement together with international guiding principles and a code of conduct for organisations developing advanced systems.[2] PRIMARY On 1 November, 28 countries and the European Union signed a frontier-risk declaration.[3] PRIMARY

What the executive order actually reaches

The order is an instruction to federal agencies. It directs the standards institute to develop guidelines for red-team testing, requires a report on content authentication and synthetic content labelling, establishes an interagency council, and instructs the budget office to issue guidance to agencies on their own use of these systems.[1] LEGAL Those are duties on government.

Its principal reach into the private sector runs through defence production authority and applies to developers of dual-use foundation models above a stated compute threshold, and to operators of large computing clusters, who acquire reporting obligations.[1] LEGAL A separate provision directs the preparation of a rule that would require providers of infrastructure services to report foreign training runs above a stated threshold, and to ensure that foreign resellers verify the identity of the foreign customers they sign.[1] LEGAL The thresholds are described in the order as interim, pending further definition. An enterprise fine-tuning a commercial model for internal drafting is nowhere near them.

This matters because the order is being read as regulation of enterprise AI use, and it is not that. An organisation building a compliance programme against it will build against duties it does not hold, and will miss the actual exposure, which arrives through data protection law, sector regulation, contract and employment obligations that already applied before any of this.

FOUR INSTRUMENTS, AND WHO EACH ONE BINDS

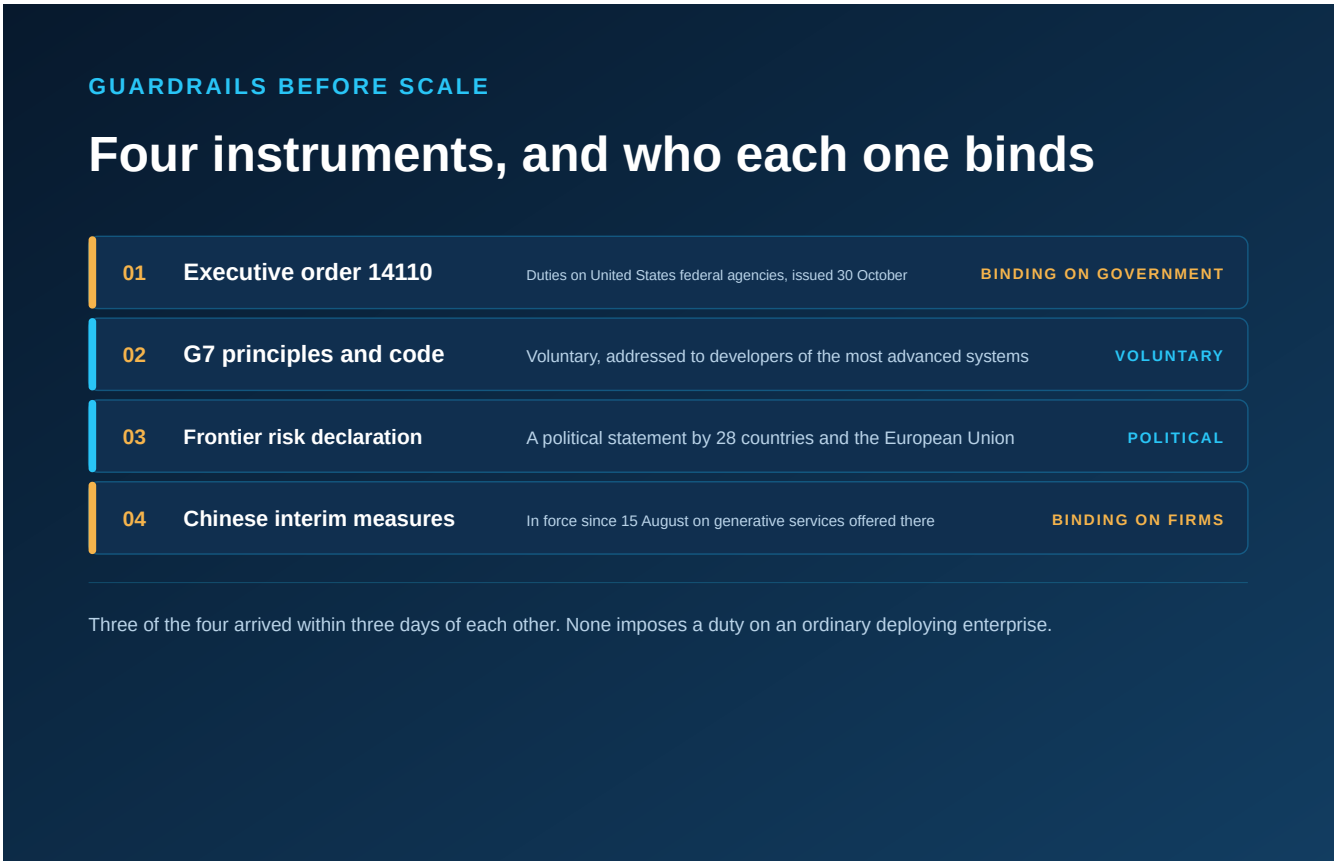

FIGURE 1 Three of the four arrived within three days of each other at the turn of the month. Only the first and the fourth create duties on anybody, and neither set of duties falls on an ordinary enterprise deploying an assistant.

Primary instruments, read from their issuing bodies. PRIMARY / LEGAL

The international instruments, and who they address

The Group of Seven principles and code of conduct are voluntary and addressed to organisations developing the most advanced systems.[2] PRIMARY The declaration signed on 1 November is a political statement committing signatories to cooperation on frontier risk, and it imposes nothing on companies.[3] PRIMARY Both are worth reading. Neither is a compliance baseline for a deploying enterprise. Adopting them as one produces a control framework calibrated to a different party's obligations.

> THE EUROPEAN POSITION AS AT 30 NOVEMBER 2023
>
> The proposed European regulation remains in trilogue between the Parliament, the Council and the Commission, with the Council having adopted a general approach in December 2022 and the Parliament its negotiating mandate in June 2023. LEGAL
>
> No agreed text exists at the date of this paper, no adoption has taken place, and nothing is in force. Any planning assumption about final obligations is an assumption about a negotiation in progress, and organisations should treat published summaries of the likely outcome accordingly.

One binding instrument exists and attracts far less attention. Measures governing generative services took effect in China on 15 August 2023, which makes them the first binding national rules of their kind

anywhere.[5] LEGAL Organisations operating there have obligations now, which is a different position from the one most Western commentary describes.

SECTION 2

From principles to something observable

A principle is a statement of intent. A control is a thing somebody does, on a date, leaving a record.

Every framework in circulation names desirable characteristics. Systems should be valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed.[4] PRIMARY Nobody disagrees, which is the difficulty. A characteristic that nobody disagrees with generates no decisions.

Translation into a control requires three things the principle does not supply. Who does it, on what evidence, and what happens when the answer is unsatisfactory. A requirement that systems be explainable becomes operational when it reads that before release the product owner records which decisions the system influences, what the user is shown about its basis, and who signs that this is adequate for the tier.

FROM A PRINCIPLE TO A CONTROL

FIGURE 2 Four questions that turn a desirable characteristic into something an assurance function can test. A principle surviving all four is a control.

My framework, applied to the trustworthiness characteristics in the NIST framework. AUTHOR / PRIMARY

A structural point about the framework most organisations are adopting

The framework published in January 2023 organises work around four functions, and they are not four sequential stages.[4] PRIMARY Three of them describe activities performed on a system. The fourth describes the culture, accountability structures and policies within which those activities take place, and it is cross-cutting by construction. An organisation that runs it as step one will complete a governance workstream, close it, and find 18 months later that the mapping and measurement work has no owner. AUTHOR / PRIMARY

SPECIFICATION

Turning any principle into a control, in four questions

- What observable thing would be true if this principle were being honoured, stated so that two people would agree on whether it was true?
- Who produces that evidence, as part of what activity, and on what cadence?
- Who reviews it, and what authority do they have when it is inadequate?
- What is the record, where does it live, and how long is it kept? A control leaving no artefact cannot be assured and will not survive the first audit.

SECTION 3

Four tiers, and the argument for keeping them few

Tiering by consequence rather than by technology. Four is enough, and organisations that build nine spend the year classifying.

The instinct is to tier by technology, which produces categories like generative, predictive and automation, each with its own control set. It fails because the same model in two workflows carries entirely different exposure, and because the technology labels will be obsolete before the policy is approved.

FOUR TIERS, SET BY CONSEQUENCE RATHER THAN BY TECHNOLOGY

TIER	WHAT SITS HERE	WHAT THE TIER REQUIRES
Prohibited	Uses the organisation will not permit at any maturity, stated positively and kept short.	A written list with a named owner and a review date. Anything absent from it is permitted by default, so the list must be read as a whole rather than extended informally.
High consequence	Output that reaches a customer, affects a person materially, enters a regulated filing, or cannot be withdrawn.	Independent evaluation before release, human decision authority retained, full logging, named accountable owner, and re-evaluation after any model change.
Controlled productivity	Internal work on approved data, where the recipient can verify and the error stays inside.	Approved tool, data boundary rule, stated verification rung, incident route, and periodic sampling of output quality.
Low-risk experimental	Bounded exploration on non-sensitive material with no path to a decision.	Registration so the organisation knows it is happening, a stop condition, and nothing else. Over-controlling this tier drives the activity underground.

Why the lowest tier should be left nearly alone

The temptation is to apply the high-consequence control set everywhere. The reasoning is that consistency is safer. The practical effect is that exploratory use moves to personal accounts and personal devices, where the organisation has no visibility at all. Registration with a stop condition and almost nothing else is the trade that keeps the activity where it can be seen. AUTHOR

The summer survey evidence is consistent with the gap being large. Among respondents reporting adoption of artificial intelligence of any kind, 21 per cent said their organisation had established policies governing employee use of generative tools.[6] SURVEY The panel is self-selected and the figure describes the panel. The direction is nonetheless the one visible inside most organisations this autumn.

The prohibited list, and how to write it

Write it positively and keep it under ten items. A long prohibition list reads as comprehensive and functions as permissive, because a reader encountering a use not on the list infers that it is allowed. Where the organisation means that a category needs authorisation and is not forbidden, say so, since conflating the two produces either paralysis or quiet non-compliance.

SECTION 4

Controls across the lifecycle

Something has to happen at eight points in a system's life. Approval is the point everybody covers and the least informative of the eight.

Governance effort concentrates almost entirely at approval, which is the single moment in the lifecycle when the system is not operating. The exposures that produce incidents arrive earlier, at intake and integration, and later, when a supplier changes something.

1. Intake is a registration record of two pages, with purpose, owner, data classes, tier and intended users. Without it the organisation cannot answer how many systems it has, which is the first question any regulator or insurer asks.
2. Mapping records what the system touches, meaning which data, which decisions it influences, which people are affected, and what happens downstream of its output.
3. Evaluation is a representative task set with a quality standard and a tolerance, run before release. It is the control most organisations do not have and cannot buy.
4. Approval is a decision at the right level for the tier, recorded, with residual risk explicitly accepted by somebody entitled to accept it.
5. Deployment means access enforced at retrieval, prompts and outputs logged to the standard the tier requires, and users told what the system is and is not for.
6. Monitoring covers usage, refusals, escalations, user corrections and incidents, reviewed by a named person. A dashboard nobody opens is not monitoring.
7. Incident response needs a route that works at four on a Friday, with a defined severity scale and an owner. Near misses are recorded, because they arrive first and cost nothing.
8. Retirement settles what happens to the data, the configuration and the workflow when the system is withdrawn or replaced.

The control that has no established method

Evaluation is the genuine gap. Conventional software is tested against expected outputs. A probabilistic system requires a task set, a quality standard and a tolerance, and there is no published method an enterprise can adopt without building it. The executive order directs the standards institute to develop guidelines for red-team testing.[1] LEGAL That work has not been published at the date of this paper, and organisations needing an evaluation capability this quarter will have to construct one and refine it when guidance arrives.

> **PROMPT INJECTION, BRIEFLY**
>
> Where a system processes content it did not originate, a document, a web page or an email, that content can contain instructions the system may follow. Input validation does not help here, because instruction and data are identical. Treat any untrusted content entering a prompt as hostile, and keep systems that read external content away from systems that can act.

SECTION 5

Decision rights, and the one nobody has assigned

Five parties hold decisions, and the gap between them is almost always the same gap.

Governance documents describe roles. What they rarely do is state who may accept a risk that cannot be eliminated, which is the decision on which every deployment actually turns. In its absence the decision is taken by whoever is least uncomfortable, usually the person under most delivery pressure.

WHO DECIDES WHAT

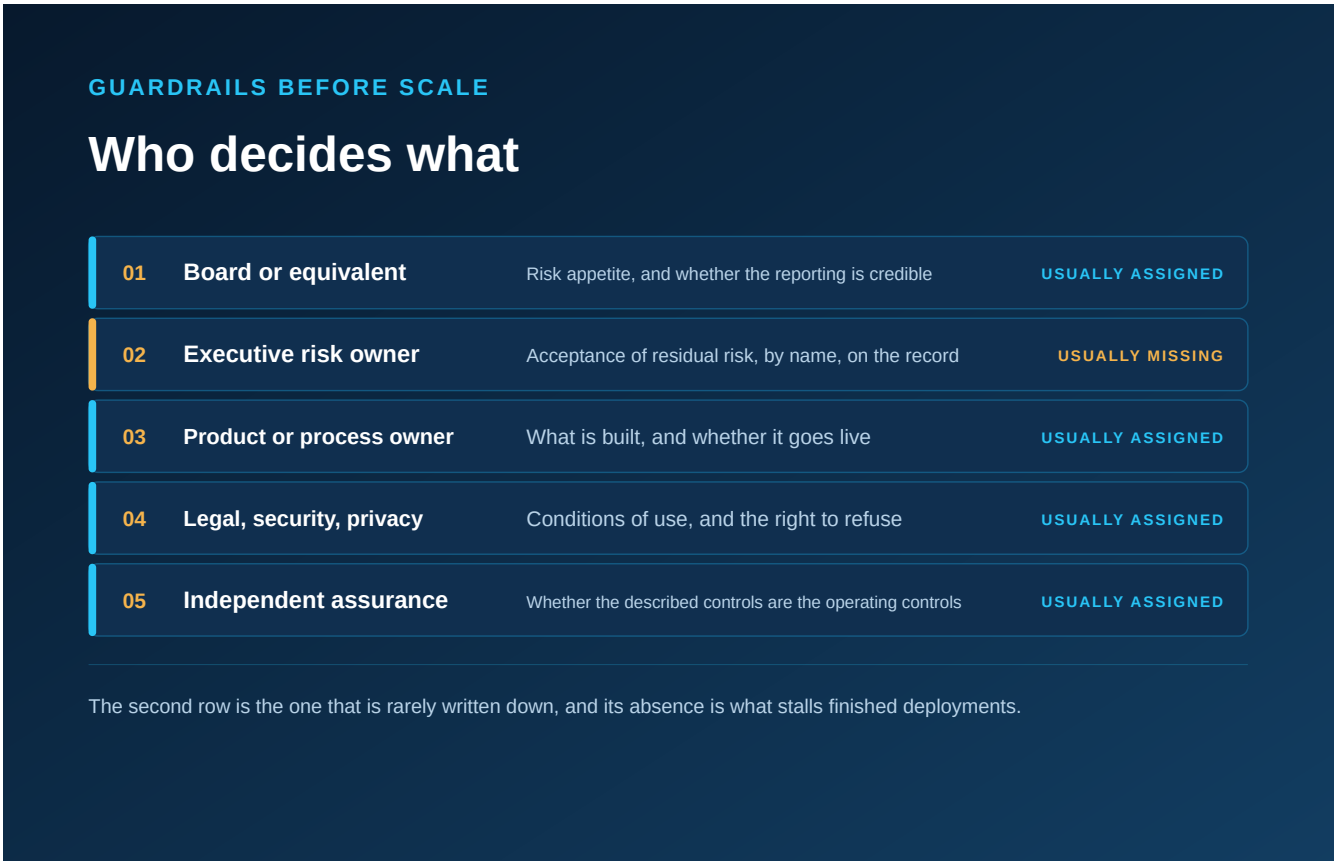

FIGURE 3 Five parties and the decisions each holds. The column that is usually missing names who may accept residual risk, and that decision determines whether anything ships.

My framework. AUTHOR

DECISION RIGHTS, AND THE ONE THAT IS USUALLY UNASSIGNED

PARTY	WHAT THEY DECIDE	WHAT THEY SHOULD NOT BE ASKED TO DECIDE
Board or equivalent	Whether the organisation has an adequate system of control, and whether the risk appetite it has set is being observed.	Individual use cases. A board approving deployments has become an operating committee and has stopped providing oversight.
Executive risk owner	The risk appetite itself, the tier definitions, and acceptance of residual risk above a stated threshold.	Technical control design. The appetite is a business judgement and the controls are an engineering one.
Product or process owner	The use case, the workflow design, the verification rung, and whether to proceed at a given tier.	Whether the data terms are acceptable. That sits with legal and security, and a product owner under delivery pressure is the wrong party to ask.
Legal, security, privacy	Data handling, contractual terms, retention, disclosure obligations and supplier assurance.	Whether the business benefit justifies the residual risk. Review functions advise on risk and do not hold appetite.
Independent assurance	Whether the controls described are the controls operating, tested rather than asserted.	Anything they would later have to audit. Assurance that helped design a control cannot then attest to it.

The board's job, stated narrowly

A board cannot assess use cases and should not try. Its question is whether a system of control exists, whether it is operating, and whether the organisation is inside the risk appetite the board has set. Three pieces of evidence answer it. A count of registered systems by tier, a list of accepted residual risks with who accepted them, and an incident and near-miss record. All three fit on one page and none of them exists in most organisations this month. AUTHOR

Independent assurance and the conflict to avoid

Internal audit, or its equivalent, should test whether described controls are operating controls. The failure mode is predictable and common, which is that the assurance function helps design the framework because it has the relevant expertise, and then cannot credibly attest to it. Where that expertise sits nowhere else, buy the design work in and keep the attestation independent.

SECTION 6

Supplier assurance while the terms are still open

Six contractual terms are negotiable before deployment and considerably less so afterwards.

Most organisations will consume these capabilities through suppliers, which makes the contract the principal control. The window in which terms can be improved is narrow and it is open now, because the vendors are competing for enterprise commitments and because the deployment is not yet embedded in anybody's daily work.

SUPPLIER ASSURANCE, AND WHAT TO ASK FOR IN WRITING

TERM	WHAT TO REQUIRE	WHY IT MATTERS MORE HERE THAN IN ORDINARY SOFTWARE
Training and retention	Whether inputs are used for training, how long they are retained, by whom they may be read, and on what legal basis.	The input is the organisation's confidential material, and the default terms of consumer products differ sharply from enterprise ones.
Change notice	Advance notice of model or behaviour changes, with a stated period and a description of what changed.	A silent model update can change output quality overnight, and evidence gathered before it no longer describes the system in use.
Testing access	The right to run your own evaluation set against the production configuration, at intervals you choose.	Vendor benchmarks measure general capability. Only your own set measures the thing you deployed.
Logging	Access to logs sufficient to reconstruct what was asked and what was returned, with a stated retention period.	Without it an incident cannot be investigated and a disputed output cannot be evidenced.
Sub-processors and region	Who else touches the data, where it is processed, and notice of change.	The supply chain under a hosted model is longer than under conventional software and less visible.
Exit	What is returned, in what format, over what period, and what survives termination.	Configuration, evaluation sets and prompt libraries represent most of the accumulated investment and are the easiest to lose.

Change notice, which is the term to fight for

Of the six, advance notice of model change is most often absent, and its absence does the most damage. An evaluation performed in October describes a system that may behave differently in December, without any announcement, and the organisation will discover the difference through a user complaint. A notice period describing what changed makes the evaluation capability worth building. AUTHOR

A recent product announcement illustrates how quickly the ground moves. On 6 November 2023 OpenAI described a facility allowing custom versions of its assistant to be created without writing code, built from instructions, additional knowledge and selected capabilities, with enterprises able to deploy internal-only versions from later that week.[7] VENDOR Whatever its merits, it lowers the effort required to connect a capable system to organisational knowledge, which raises the importance of the intake register in section

four. A control that depends on projects being large enough to notice stops working when building one takes an afternoon.

SECTION 7

A governance capability in 90 days

Inventory, interim policy, a forum that meets, three assessments and a one-page dashboard. Nothing here requires a new function.

The temptation at this point is to commission a framework, which takes six months and produces a document. What follows is the smallest thing that produces decisions, and it can be run by an existing risk or technology function with a named owner.

Days 1 to 30, find out what exists

Inventory. Ask every function what they are using, including personal accounts, with an explicit amnesty for past use. The amnesty is not a courtesy and it is the mechanism, because an inventory built without one records what people are willing to admit. Publish an interim data boundary rule in one page while the inventory runs. The full policy can follow.

Days 31 to 60, tier and assess

Apply the four tiers to what the inventory found. Most entries will fall into controlled productivity and low-risk experimental, and the small number that do not are where the effort goes. Take the three highest-consequence uses and assess them properly, end to end, including the evaluation step. Three real assessments teach the organisation more than a complete framework, and they produce the template for the rest.

Days 61 to 90, stand up the forum and the record

Convene a review forum with the five parties from section five, meeting monthly, with authority to approve at tier and to stop. Publish a one-page dashboard covering registered systems by tier, accepted residual risks and their owners, incidents and near misses, and evaluation coverage. Send it to the board quarterly, asked for or not.

> **WHAT THIS DOES NOT PRODUCE**
>
> The 90 days produce a governance capability that can make decisions and evidence them. It does not produce a mature evaluation function, because that requires task sets built for each significant use and refined over several cycles.
>
> It also does not produce compliance with any specific instrument, because at the date of this paper no instrument imposes obligations on a deploying enterprise that would not have applied anyway through existing data protection and sector regulation. What it produces is the record an organisation will need when that changes, which on the current trajectory is a matter of quarters.

Whether the European negotiation concludes this winter and what shape it takes are not things this paper can know. The position worth being in is one where the answer changes the paperwork rather than the practice.

END MATTER

Method, evidence and limits

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

Selection

Primary legal and institutional instruments cited from their issuing bodies, one industry survey cited with its method, and one vendor announcement cited as a vendor announcement. Commentary on the instruments is not cited, on the ground that the instruments are short enough to read.

Legal status

Legal status is stated precisely throughout. The executive order was issued on 30 October and published in the Federal Register on 1 November. The European proposal is in trilogue with no agreed text. The international instruments are voluntary and addressed to developers. Conflating any of these states is the most common error in current governance writing and it changes what an organisation believes it must do.

Timing

The European trilogue is scheduled to resume in early December, and this paper describes the position before it. A reader consulting the paper afterwards should treat sections one and six as a snapshot.

Gaps in the evidence

No published evaluation method that an enterprise could adopt without building it. No certifiable management system standard specific to artificial intelligence has been published yet, though the international standards bodies have one at the final stage of drafting. And no evidence establishes whether the control set proposed here is proportionate, because no organisation has operated one long enough for that question to be answerable.

The parts that are my own

The four tiers, the lifecycle control list, the decision rights table and the supplier terms are my own. The reading of the January framework's fourth function as cross-cutting is my reading of a published document, and readers who need to rely on it should go to the framework itself.

My interest

I advise on governance design and could be a supplier of the 90-day work in section seven. The paper argues against commissioning a comprehensive framework, which is the larger and more lucrative engagement.

END MATTER

Sources considered and set aside

Instruments and reports I considered for this paper and did not rely on. Several are worth reading for other purposes.

SOURCE	PUBLISHED	WHY IT WAS SET ASIDE
Frontier Model Forum, founding announcement	26 July 2023	An industry body formed by four developers. It publishes intentions, and an industry body's intentions are not a control.
Stanford Center for Research on Foundation Models, Foundation Model Transparency Index	18 October 2023	Scores ten developers on how much they disclose. Useful for choosing a supplier and silent on how to govern the system once bought.
OpenAI, Preparedness team announcement	26 October 2023	A developer describing how it will evaluate its own frontier models. It is the supplier's control, and it does not become the customer's by being read.
Infocomm Media Development Authority of Singapore, AI Verify	June 2023	A testing toolkit and framework from one jurisdiction, open-sourced in June. Worth watching, and outside the scope of a paper written for organisations that mostly are not there.
United States Senate Judiciary subcommittee, hearing on oversight of artificial intelligence	16 May 2023	Testimony. Interesting on the politics and inadmissible on the obligations, because nothing said at a hearing binds anyone.
ISO/IEC 23894, guidance on artificial intelligence risk management	6 February 2023	Guidance, with nothing to certify against, and largely the general risk management standard restated with examples. Read it once.

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. He writes about regulation as a board adviser rather than as a lawyer, which means describing what an instrument says, how firmly it says it, and what follows for an organisation, and stopping short of an opinion on the law.

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

Seven works cited in the text, each with its original date of publication. None is later than 27 November 2023.

1. **The White House.** *Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.* Issued 30 October 2023, published at 88 Federal Register 75191 on 1 November 2023.
   PUBLISHED 30 OCTOBER 2023

2. **Group of Seven.** *Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems*, and *International Code of Conduct for Organizations Developing Advanced AI Systems*, published with the Leaders' Statement. Voluntary, and addressed to developers.
   PUBLISHED 30 OCTOBER 2023

3. **Countries attending the AI Safety Summit.** *The Bletchley Declaration.* Signed by 28 countries and the European Union. A political declaration without binding effect.
   PUBLISHED 1 NOVEMBER 2023

4. **National Institute of Standards and Technology.** *Artificial Intelligence Risk Management Framework 1.0.* NIST AI 100-1. Voluntary, rights-preserving, non-sector-specific and use-case agnostic, organised around Govern, Map, Measure and Manage.
   PUBLISHED 26 JANUARY 2023

5. **Cyberspace Administration of China and others.** *Interim Measures for the Management of Generative Artificial Intelligence Services.* Order No. 15, dated 10 July 2023 and published 13 July 2023, effective 15 August 2023.
   PUBLISHED 13 JULY 2023

6. **McKinsey and Company.** *The state of AI in 2023: Generative AI's breakout year.* Self-selected panel of 1,684 respondents, fielded April 2023.
   PUBLISHED 1 AUGUST 2023

7. **OpenAI.** *Introducing GPTs.* Cited for the described capability and for the internal-only enterprise deployment option, both attributed to the vendor.
   PUBLISHED 6 NOVEMBER 2023

END MATTER

Further sources consulted

Eight 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 27 November 2023.

- **Department for Science, Innovation and Technology.** *Emerging processes for frontier AI safety.*
  PUBLISHED 27 OCTOBER 2023
- **Department for Science, Innovation and Technology.** *Frontier AI: capabilities and risks.* Discussion paper.
  PUBLISHED 25 OCTOBER 2023
- **The White House.** *Voluntary commitments from leading artificial intelligence companies.* First tranche of seven companies, with a further eight on 12 September 2023.
  PUBLISHED 21 JULY 2023
- **European Commission.** Proposal for a Regulation laying down harmonised rules on artificial intelligence, COM(2021) 206 final. Cited at its status on the date of this paper, which is a proposal in trilogue.
  PUBLISHED 21 APRIL 2021
- **Council of the European Union.** General approach on the proposed artificial intelligence regulation.
  PUBLISHED 6 DECEMBER 2022
- **European Parliament.** Negotiating mandate on the proposed artificial intelligence regulation.
  PUBLISHED 14 JUNE 2023
- **Organisation for Economic Cooperation and Development.** *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449. Adopted 22 May 2019 and amended on 8 November 2023, which is the operative text at this date.
  PUBLISHED 22 MAY 2019
- **Department for Science, Innovation and Technology.** *A pro-innovation approach to AI regulation.* Five principles, non-statutory in the first instance.
  PUBLISHED 29 MARCH 2023

WHERE THIS GOES NEXT

The next paper asks who owns all of this

Five papers have now described a learning system, an individual capability, a workforce design, a portfolio and a control framework. Each assumed somebody was accountable for the whole, and none has said who.

The sixth paper will take that question directly, and it is due at the end of March. The executive order has instructed the United States budget office to specify who holds the role inside federal agencies, and that guidance is due within the same window, so the paper should be able to read a government's answer against the private sector's.

THREE THINGS THE NEXT PAPER WILL TAKE FOR GRANTED

- An inventory produced under amnesty, so that it records what is happening rather than what people will admit.
- Four tiers applied to that inventory, with the high-consequence entries assessed properly.
- A named person who may accept residual risk, and a record of what they have accepted.

Paul Forrest · 30 November 2023 · Version 1.0 · Research closed 27 November 2023