What has to be settled first

Version 0.2 · revised 8 September 2026 · this version as a PDF

Safe for what, fair by whose measure, accurate to whose tolerance? Nobody can answer those for your organisation, which makes the goodness in question yours to find — and there are arrangements in which it cannot be found at all.

The pursuit of 'Goodness' in AI — Sovereign agents, and what it takes to trust one with real work

---

What this series is for

FIG-26 THE BLANK INSIDE EACH WORD

THE WORD AS OFFERED	THE BLANK IT CONTAINS	WHO CAN FILL IT
safe	against which harms?	→ only the answerable body
fair	by whose measure?	→ only the answerable body
accurate	enough for which decision?	→ only the answerable body

WHAT A SUPPLIER ANSWERS INSTEAD

Safe against the harms it chose. Fair by the measure it selected. Accurate to the tolerance it set.

Each answer is true, checkable, and about a different question from the one you had to ask.

A supplier who has answered them for you has told you what they were willing to promise, not what you needed to know.

FIG-26 Safe, fair and accurate look like properties of a system and are not. Each contains a blank — against which harms, by whose measure, enough for which decision — and only the body that will answer for the system can fill it. A supplier answers a na
and the two are easy to mistake for one another.

📖 Ask the Glossary

An organisation asking whether it can trust an AI system is asking a real question, and nobody can answer it on that organisation's behalf. Not a supplier, not a regulator, and not this series.

That is a statement about how the questions are built. Almost everything published on the subject asks whether a system is aligned, safe, fair, accurate, robust or explainable. **Those are the right questions**, and any account treating institutional machinery as an alternative to asking whether a system works is selling paperwork. But every one of them is incomplete as posed, and incomplete in the same way. Safe — for whom, against what, at what cost to what else? Fair — on which of several incompatible measures, each defensible, which cannot be satisfied together? Accurate — to what tolerance, and who carries the residual error? Each contains a blank, and the blank is filled by somebody exercising judgement about a particular situation. **The question is not answerable until a party with the standing to do so has filled it, and the only party positioned to fill it is the one that will answer for the result.**

So the goodness in the title is not a property to be located in a product and verified. It is the reader's own — their purposes, their obligations, their tolerance for being wrong, in their setting. Nobody can hand it over, and an arrangement that appears to be handing it over is doing something else.

This series is therefore a set of instruments for looking rather than an argument to be accepted. Five questions that establish whether an institution governs the thing acting for it. Five levels of permitted autonomy and the criteria for setting them. Four properties that either hold or do not. Twelve requirements in the form they would take in a contract. Each is a way of asking, of a particular arrangement, whether the reader's judgement can actually reach what their software does.

And the same instruments show where the answer is unlikely to be found. Certain arrangements make the questions unanswerable by construction: where the authority to set a limit was never held by anyone in the organisation, where the record shows what a supplier chose to record, where a capability can be withdrawn on notice by a party the institution does not govern. In those arrangements a reader may search for goodness diligently and be looking somewhere it cannot be.

Knowing which arrangement one is in is the first result these instruments produce, and for many organisations it is the only one they will need.

Where the technical questions stop

The difficulty is not that measurement is hard. It is that each of these properties is two-place where it is usually written as one-place: not *safe*, but *safe for this purpose*; not *fair*, but *fair on this construal*; not *accurate*, but *accurate enough for this decision*.

Fairness makes the structure unavoidable, because the incompatibility has been proved rather than argued. Several reasonable statistical definitions of a fair classifier cannot all hold at once except in degenerate cases; a system satisfying one will fail another, and the choice between them is a judgement about which unfairness a particular institution is willing to carry. No amount of testing selects between them. Somebody does, and that somebody either has the standing to make the choice on the institution's behalf or does not.

Safety has the same shape once the system acts rather than advises. A threshold prudent in a small practice is negligent in a hospital and obstructive in a newsroom. The response appropriate to one correspondent is inappropriate to the next. There is no configuration that encodes the right answer, because the right answer depends on circumstances the system does not possess and could not be given.

An institution asking whether a system is good is therefore asking for something no supplier is positioned to provide, and it will generally be provided with something else: a description of testing, of benchmarks, of guardrails, of review processes. None of that material is offered in bad faith and much of it represents real work. It is nonetheless not an answer, because the answer requires a judgement about a situation in which the tester was not present.

The objection from attribution law

There is a serious objection at this point and it deserves stating at full strength.

Delegation is not an undefined term. Agency law has worked on it for centuries. And electronic transactions law addressed automated systems directly, well before this debate: where a person deploys a system that operates automatically, **the acts of that system are attributed to the person who deployed it**. Software is not an agent in law precisely because it cannot be answerable. It is the means through which a principal acts, and the principal carries what it does — as with a defective spreadsheet, or a machine on a factory floor. ⚠️ Described from general knowledge of the model law and its national enactments; the instruments were not retrieved for this draft and should be before this passage is relied on.

That position is correct, and it does not weaken the argument. It is the argument.

If the acts of the system are attributed to the deploying institution, then **the institution's arrangements are not one governance option among several. They are the thing the liability attaches to.** The question of who set the threshold, who may change it, and what

record establishes what was done under it is not administrative housekeeping around a technical artefact — it is the determination of what the institution has actually committed itself to, in advance, in every act the system performs on its behalf.

The attribution rule tells an institution that it will answer. It does not tell it what it will be answering *for*, and that is settled entirely by arrangements the institution makes or fails to make — so a body that has never decided what its system may do has made a commitment anyway, without noticing that it was making one.

Structural governance, and governance by document

There is a distinction running underneath everything that follows, and naming it early saves a great deal of confusion later. Almost all AI governance in circulation is **governance by document**: a policy, a charter, an acceptable-use standard, a schedule of requirements, a clause in a contract. Such a document states what should happen, works by being read and complied with, and fails silently: nobody notices the policy that stopped being followed.

Structural governance puts the constraint where compliance is not the mechanism. The limit is not a rule the system is asked to observe; it is a capability the system does not hold. Not *the agent must not send external email* but *the agent has no credential that any mail server will accept*. Not *the agent must log its actions* but *the agent acts by writing into the record, so there is no gap between the doing and the recording in which a discrepancy can live*. Not *the agent must not exceed its scope* but *the token it holds carries the scope, and the tool refuses anything outside it, and the refusal happens somewhere the agent cannot argue with*.

A documentary control fails when somebody is mistaken, distracted, overruled, or gone. A structural control has to be dismantled, and dismantling leaves evidence.

That is the test. It is also why, in the widely circulated safeguards diagram examined later in this piece, the innermost of its five rings matters more than the four outside it and still does not save it. Prompt filtering, output validation and content detection are the only items in the picture operating on the machine rather than on the people, and each of them inspects text and asks the system to behave — which places them on the documentary side of the line despite being made of code.

Two qualifications, because the distinction is easy to overstate.

The first is that some things cannot be made structural and should not be pretended into it. A grant of authority is a decision, and decisions are recorded in minutes. Whether a limit is drawn in the right place is a judgement, and judgements are written down. What can be

structural is not the *deciding* but the *enforcement*: the decision is documentary and the boundary is not, so that an organisation which changes its mind must change the structure rather than merely reinterpret the paragraph.

The second is that structure has its own failure mode, and it is the mirror image. A documentary control can be ignored; a structural one can be wrong and immovable, enforcing last year's judgement against this year's circumstances with nobody able to override it in the moment it matters. Anything built this way needs a deliberate route to change it, held by somebody entitled to use it — which returns the argument to authority, and is why these questions come before the technical ones rather than after.

A reader can tell the difference by asking, of any control offered to them, what would happen if everyone concerned were mistaken about what the system was doing. **Where governance by document is all that exists, the arrangement is ungoverned and well-documented**, and much of what is sold as AI governance is in that condition.

What fills the gap: "just use it", and the measure that comes with it

Structural governance is difficult and slow. Documentary governance is achievable and partial. That leaves a gap between what an organisation can readily do and what would actually hold — and something occupies it.

What occupies it is an argument that sounds like pragmatism: **get on with it, start using the thing, learn by doing, do not let perfect be the enemy of good.** As advice about software procurement that is often sound. As a response to a question about authority it is not an answer, and it is worth being precise about what it does instead.

It converts a question about who decided into a question about attitude. The person asking under whose authority a system is acting becomes, in the framing, the person who is resistant, blocking, behind the curve, not on board. Nothing has been answered; the asker has been reclassified, and because the reclassification is social rather than argumentative there is no reply available that does not confirm it — insisting is evidence of the trait.

That the mechanism is real can be established from the safeguards material itself. That safeguards diagram examined below places **"Psychological Safety — foster safe space to question and discuss AI use"** in its outermost layer. As a cultural recommendation that is sensible enough. What it concedes is that **asking questions about AI carries a social cost high enough to need a programme** — and the diagram treats that as a cultural amenity to be provided, rather than as evidence that the governance question has been converted into a loyalty test. The same layer carries the observation that a substantial share of employees

conceal their AI use from their employer. An organisation in which staff hide what they are doing, and in which questioning requires designated safety, has arrived at the condition every instrument in this series exists to detect: nobody in it can find out what is happening.

What adoption actually measures

Having converted the question, the framing supplies a measure, and the measure is adoption: seats, active users, queries, tokens, the proportion of staff using the tool weekly. Its defects are worth setting out one at a time, and it is not simply that it is a crude proxy.

Adoption is guaranteed to rise independently of benefit. It rises when a feature is switched on by default, when an alternative is withdrawn, when a workflow routes through it, when use is mandated, and when the interface makes the assisted path the shortest one. None of those has any relationship to whether the work got better. A number that increases under conditions the supplier controls is a measure of exposure, not of value.

It is the one number a supplier can produce, which is why it is the number offered. A supplier cannot measure whether an institution's correspondence improved, whether its decisions were sounder, or whether its obligations were met, because those are facts about the institution's situation. It can measure usage precisely, in real time, and present it as evidence.

The loop closes on itself. High adoption is offered as proof of value; proof of value justifies wider deployment; wider deployment raises adoption. Nothing anywhere in that circuit touches whether anything improved, and it requires no external input to keep turning, so an institution reporting adoption to its board is reporting the supplier's success and calling it its own.

And it inverts the governance question rather than displacing it. Once adoption is the measure, every control becomes a cost. The approval step that would have caught something appears in the reporting as friction; the limit that held appears as a reason the number is below target; the person who declined to route a task through the tool appears as an outlier. **The institution's own reporting now penalises the machinery that protects it**, quietly, in the ordinary course of management review, without anyone deciding that governance should be discouraged.

Measuring adoption is not itself the trap. The trap is that it is the only thing measured, that it cannot fall for reasons that would matter, and that everything capable of making it fall for a good reason has been reclassified as an obstacle.

What an institution can measure instead

The alternative is not a better metric of the same kind, and any offered would be subject to the same capture. What an institution can establish, at any moment and without a supplier's cooperation, is whether it can answer for what was done: which grants are in force, who

made them, what was done under each, what the system declined to do and at which boundary, and whether somebody who was not present can reconstruct it.

That is not a number and it does not trend. It is a capability an institution either has on a given Tuesday or does not, and unlike adoption it cannot be raised by switching something on.

A compiler with two masters, 1988

If the diagnosis holds, the characteristic failure of these systems should be constitutional rather than behavioural: not a system doing something wrong, but a system whose acts cannot be traced to an authority. That failure has a name and it predates the present debate by decades.

Norm Hardy, in ACM SIGOPS Operating Systems Review in 1988, described a compiler running with authority drawn from two sources — the invoker's, and its own licence to write to its home files — unable to distinguish them:

> The compiler serves two masters and carries some authority from each to perform its respective duties. **It has no way to keep them apart.**

The detail that matters is that nothing malfunctioned and no code changed. *"When the code was written to produce the output it was correct! What happened to make it wrong? The precise answer is that it became wrong when we added home files license to (SYSX)FORT."* The defect entered from outside the program, in a grant made elsewhere. **No amount of testing the artefact would have found it, because the artefact was not defective.** The technical questions, however well answered, do not reach a failure of that kind.

It is not history. The Model Context Protocol's authorization specification, version 2025-06-18 — the protocol most current agent tooling uses — carries a section headed **"Confused Deputy Problem"** and a requirement in the imperative: **"The MCP server MUST NOT pass through the token it received from the MCP client."** Thirty-seven years separate the two, and the same defect was waiting at the end of them.

What keeps attention on the technical half

Two features of the present arrangement keep attention on the technical half, and neither requires anybody to be acting in bad faith.

The first is that the technical questions are the ones a supplier can answer. A supplier can test a model, publish benchmarks, describe its filtering and document its review processes. It cannot supply an institution with the standing to fill the blanks, because standing is not a property the supplier holds and could transfer. Given a question it can answer and a question it cannot, any organisation answers the first — truthfully, and about the wrong half.

The second is vocabulary. Terms drawn from the physical world arrive carrying their physics and deliver none of it. A guardrail on a road stops a vehicle whatever the driver intends; the things called guardrails in this field are instructions, evaluated by the system being instructed. Verbs of accomplishment do the same work: filtering *prevents* harmful content, review *ensures* compliance, training *embeds* values. A filter reduces, probabilistically, and its failure rate is the subject of an active research literature. A review samples. Values are not embedded.

The effect is to foreclose the reader's question before it forms. Somebody reading that a system prevents unsafe outputs does not go on to ask whether anything is prevented, because the sentence has already answered it.

What that produces, at its best, can be examined. A widely circulated diagram sets out AI safeguards as five concentric rings: culture on the outside, then governance, an operating model, process, and technical safeguards at the centre. Its content is sensible — ethical oversight, defined decision rights, staff training, approval paths, monitoring, and at the core prompt filtering, output validation, content filters, detection of personal information. An organisation possessing all five is better placed than one possessing none.

It is worth dwelling on because it is not a caricature. It is a faithful rendering of what the technical framing produces when applied conscientiously, by somebody with nothing to sell — which makes it a better exhibit than any vendor's material, because it shows the framing operating as shared understanding rather than as a pitch.

Apply one test. **If everyone in the organisation were entirely mistaken about what the agent was doing, which of those layers would still hold?** Culture is a disposition and holds where people are right about what is happening. Governance assigns decision rights over conduct that can be seen. An operating model turns principle into routine; process embeds approval into delivery — each depending on the organisation's understanding being accurate. Then the innermost ring, where constraint is supposed to become real, and every item in it inspects **text**: the input, the output, what passed through. None constrains what the system may **do**.

Two absences are structural. **There is no ring for the derivation of authority** — nowhere in five layers does the question arise of who was entitled to permit any of it. And **there is no ring for the supplier**: every layer is internal to the deploying organisation, so all five can be implemented immaculately while the vendor alters the model, changes what it will refuse, or withdraws it.

One box states the whole difficulty in four words. Under Culture, beside psychological safety and AI literacy, sits **“Ethical Judgment — Humans remain accountable for decisions.”** Accountability appears as a *cultural disposition*: something people hold, alongside literacy and a willingness to ask questions. Accountability is not an attitude that people maintain. It is a structure that holds them — a named party, a grant they made, a limit they set, and a record by which somebody else can call them to account. Placed in the culture ring, it depends on people continuing to feel accountable at exactly the moment that is hardest.

The diagram also carries its central claim in a form that cannot be contradicted. Nowhere does it assert that these five layers are sufficient. Had it done so, that would be a proposition, and propositions can be examined. Instead it *shows* sufficiency: five closed rings, nothing outside the outermost. Readers do not argue with a taxonomy; they fill it in. The subtitle — that all five layers matter — completes the move by making the only live question whether one has done all five, which concedes at the outset that five is the number.

That is the generic form: a picture of a world in which enforcement exists, offered in good faith, absorbed as common sense, and missing the ring that survives being wrong. No sentence in it is a lie.

What follows

If the prior questions are prior, specific things follow, and each is developed in a document of its own.

Authority must have a derivation. Who actually holds the controls sets out five questions establishing whether one exists, and examines what three major vendors’ definitions of sovereignty contain — finding no term in any of them for inspection, alteration or stopping.

Powers must be limited, and the limits declared in advance. How much it may do unsupervised sets out five levels of permitted autonomy and meets the strongest published objection to ordering autonomy that way at all.

A delegation must be real rather than nominal. The four properties underneath specifies four properties, each checkable by observation rather than by description.

And the whole must be enforceable. What you can actually require states twelve requirements in the form they would take in a procurement schedule, and reports that eight are mandatory nowhere.

None of this makes a system good. It is what allows an institution’s judgement about what is good to reach what its software does — and, given the attribution rule, to reach it before rather than after the institution is answerable for the result.

What this argument would look like if it were wrong

First, if the blanks turn out to fill themselves — if capable systems reliably infer the purpose, the tolerance and the construal a particular institution would have chosen, across settings, without being told — then the prior questions are answered by capability and the apparatus is redundant. That is a claim about the future which nobody can currently settle. It is worth noting that it would not make the constitutional questions *wrong*, only unnecessary: a pre-condition that is reliably met without effort is still a precondition.

Second, if institutions with no such apparatus turn out to govern their AI effectively — because commercial pressure, professional norms or existing law supply what the arrangement lacks — then the machinery described here is redundant in a different way, and this series describes a problem that solves itself.

Third, and most likely: if the prior questions are answerable in principle but unobtainable in practice, because no supplier will answer them and no institution can compel it, then the argument is correct and useless. What nobody has measured yet treats that as open rather than assuming its answer.

Related

What nobody has measured yet gathers the questions this series cannot answer, and asks readers for four observations that would help settle them.

The six, in reading order. Each stands on its own; read together they build one argument.

- Who actually holds the controls — five questions establishing whether an authority to act actually exists
- How much it may do unsupervised — five levels of permitted autonomy, declared rather than left implicit
- The four properties underneath — four properties a delegation must carry, each checkable by observation
- What you can actually require — twelve requirements, and what currently requires each of them
- What nobody has measured yet — four open questions, and the observations that would settle them
- An invitation to become a Distributor — not part of the argument above: a proposal to work with its author, for anyone who wants to take this further

ACKNOWLEDGEMENT

Leslie Stroh — my big brother, a product philosopher, and a magnificent role model. He helped me find my way to a vision of goodness in AI.

▸DISCLAIMER

> **How this was made.** The version number counts drafts of the text. It does not measure the inquiry behind it, which has run over days and across several AI systems, with argument between those systems and within them, directed, refused and repeatedly redirected by the author. The source material was AI-generated, and then adversarially and iteratively refined across a range of tools — systems built by different companies in different jurisdictions, set against each other and against the author. No one of them produced this text, and no one of them reviewed it alone. **The plurality is deliberate rather than incidental.** A single model carries a single set of priors about which sources are authoritative, and this series argues that an evidence base narrowed in exactly that way is how a contested question comes to look settled. Using one model to investigate that claim would have been the claim refuting itself. To name a single model on it would credit that model with work that was neither its own nor done in a single pass. The diagram on each page was drawn afterwards, with the same assistance, from the argument the essay had already made. **The plurality was also necessary, and the record should say why.** In drafting, the assisting model repeatedly led with United States institutional sources — a national laboratory, an industry association, a market study nineteen years old — and presented conclusions drawn from them as the state of knowledge. On one occasion European measured data contradicting those conclusions was present in the same research return and was placed below them. Framings were proposed that would have argued against this series' own position using that evidence base, and offered as rigour. Each was refused by the author and the material rebuilt. That is the mechanism these documents describe, occurring in their own making, and it is recorded because a series arguing that evidence bases narrow without anyone deciding to narrow them cannot credibly claim its own production was exempt. The framing, the corrections and the judgements are the author's, and so are the errors.

Cite this version

```
What has to be settled first, version 0.2 (September 2026).
agenticgovernance.digital/downloads/the-question-before-the-question-v0-2.pdf
```


Series contents

NEXT
Who actually holds the controls

Alongside: questions and answers · sources and provenance · slides

Share this Mastodon Email Copy link

Found something wrong?

This piece names what would show it is wrong. If you have that — a figure, a source, an argument — it is worth more to us than agreement. Tell us which claim and why.

Which section, if you know it

for example: 2. Four properties

The claim you are responding to

quote it, or describe it

What is wrong with it

Your email, only if you want an answer

Used to reply to you and nothing else. No list, no newsletter, never passed on.

Send this