The four properties underneath

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

Four properties that either hold or do not, a thirty-seven-year-old defect now written into the protocol most agent tooling uses, and one test that sorts every safeguard on the market into two piles.

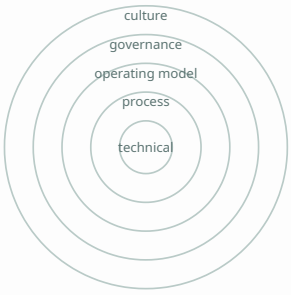
The pursuit of 'Goodness' in AI — Sovereign agents, and what it takes to trust one with real work

---

The concentric-rings picture

FIG-29 THE RINGS, AND THE QUESTION THAT COLLAPSES THEM

THE PICTURE ALMOST EVERYBODY HAS



Every layer is worth having. The picture is an accurate rendering of what the field believes.

If everyone in the organisation were entirely mistaken about what the agent was doing, which of these layers would still hold?

FOUR OF THE FIVE ASSUME THE BELIEF IS RIGHT

- culture · ethical oversight
- governance · defined decision rights
- operating model · staff training
- process · approval paths, monitoring
- technical · filters, output validation

inspects text. Does not constrain what the system may DO.

The ring that survives being wrong is the one the picture omits: an identity that is not borrowed, a capability that is not ambient, authority that narrows, and a record someone who was not there can rely on.

FIG-29 The concentric-rings picture of AI safeguards, drawn as the fi
one test: if everyone in the organisation were entirely mistaken about w
ing, which layer would still hold? Four assume the belief is right. The innermost inspects text and does not constrain what the system may do. The ring that survives being wrong is the one the picture omits.

A diagram circulates widely, in several versions, showing AI safeguards as concentric rings. Culture on the outside, then governance, then an operating model, then process, with technical safeguards at the centre. Each ring carries sensible content: ethical oversight, defined decision rights, staff training, approval paths, monitoring, and — at the core — prompt filtering, output validation, content filters, detection of personal information.

Every layer in that picture is worth having. An organisation with none of them is worse off than one with all of them, the people who draw these diagrams are not selling anything, and dismissing the genre would be both unfair and a mistake, because the diagram is an accurate rendering of what the field currently believes.

It is also missing the only ring that survives being wrong, and one question makes the gap visible.

> **If everyone in the organisation were entirely mistaken about what the agent was doing, which of these layers would still hold?**

Run it inward. Culture is a disposition; it holds if people are right about what is happening. Governance assigns decision rights over conduct people can see. An operating model turns principle into routine, and process embeds approval paths into delivery. Each of those depends on the organisation's understanding being accurate. Then the innermost ring, the technical one, where the constraints are supposed to become real — and every item in it inspects **text**. Prompt filters read the input. Output validation reads the output. Content filters and detection routines read what passed through. Not one of them constrains what the system may **do**.

Nothing in that picture would stop an agent sending a message, moving money, or combining two records it was never authorised to combine. The word usually attached to it is *guardrail*, and a guardrail on a road stops a vehicle whatever the driver intends. These are instructions, evaluated by the system being instructed — a metaphor that imports the physics of a steel barrier and delivers a note pinned to the dashboard.

The grammar of the diagram

The gap matters less than the reason it is invisible, and the reason is grammatical.

Read the verbs. *Embed* AI ethics into organisational mindset. *Set* accountability and oversight. *Assign* ownership for AI decisions. *Build* technical safeguards into AI systems. *Prevent* unsafe or harmful prompts. *Block* harmful content. *Ensure* adherence to laws, policies and standards. *Detect and protect* sensitive personal information.

Every one is a verb of accomplishment. Each reports a completed action on a state of affairs, and each is applied to something that has not been accomplished and in several cases cannot be. A filter does not *prevent*; it reduces, probabilistically, and its failure rate is the entire subject of the injection literature. A compliance review does not *ensure* adherence; it samples. Accountability is not a parameter that can be *set* — an institution either stands behind an act or does not, and no configuration screen moves that. Mindsets are not *embedded*.

The effect of that grammar is to foreclose the reader's question before it can be formed. Somebody reading "prevent unsafe prompts" does not go on to ask whether anything is prevented, because the sentence has already answered it. The verb does the work an argument would have had to do, and does it without ever making a claim that could be tested.

Then there is the form of the artefact itself, which carries the load-bearing assertion and never states it. Nowhere does the diagram say *these five layers are sufficient*. If it did, the claim would be a proposition, and a proposition can be examined, contradicted and found wanting. Instead the picture *shows* sufficiency: five closed concentric rings with nothing outside the outermost, each nested inside the next, the whole bounded. A reader does not argue with a taxonomy. They fill it in. The subtitle — all five layers matter — completes the move by making the only remaining question *whether you have done all five*, which concedes at the outset that five is the number.

🔑 **This is why the vocabulary matters more than the omission.** The missing ring is a fact about the diagram and could be pointed out in a sentence. The grammar and the form are what make the omission unnoticeable, because they have already told the reader that the picture is complete and that the constraints are real. A reader who has absorbed *guardrail*, *prevent*, *ensure* and *all five layers* has been given a world in which enforcement exists, and will assess any alternative against that world rather than against the one they are in.

It is the same mechanism this series identifies in the evidence base, arriving through language rather than through citation. There, a source states its own limits and the qualification dies on the way to the summary. Here, no qualification is ever made, so none can be lost — the accomplishment is asserted in the verb and the completeness in the shape.

And the diagram is not the work of a vendor. That is the significant part. It reads as a faithful rendering of what the field now believes, drawn by somebody with nothing to sell, which means the vocabulary has been absorbed rather than imposed. A supplier's claim can be contested at the point of sale. A shared understanding cannot, because by the time it reaches a board it has no author to question.

The four properties below are an attempt at the missing ring. Each either holds or does not, and each can be checked by watching a system work rather than by reading a description of it.

Delegation, in the standards

Start with “delegation”, because the state of it is worse than one would expect.

The IETF’s own security glossary does not define it. RFC 4949, *Internet Security Glossary, Version 2* — a document whose purpose is to define terms — has no entry for delegation. The string appears three times in the whole glossary, inside other entries. RFC 9700, the current best-practice document for OAuth 2.0 security, does not contain the word. OAuth 2.0 itself, the protocol most of the industry uses *for* delegation, uses it once, adjectivally.

The one normative definition retrievable from a standards body comes from a provenance model rather than an access-control one. W3C’s PROV-DM, a Recommendation of 30 April 2013:

> Delegation is the assignment of authority and responsibility to an agent (by itself or by another agent) to carry out a specific activity as a delegate or representative, **while the agent it acts on behalf of retains some responsibility for the outcome of the delegated work.**

The clause that survives is the one saying responsibility does **not** transfer. PROV also states its own limit: it does not say who bears responsibility, or in what degree.

A word without a definition cannot be checked. When a supplier says an agent has been “delegated authority”, there is no standard to hold the claim against, and the phrase means whatever that supplier’s implementation happens to do.

The first property: an identity that is not borrowed

Two standards define “principal”, and the gap between them is the subject of this series.

Saltzer and Schroeder, 1975, in the paper that founded the field: *“A principal is, by definition, the entity accountable for the activities of a virtual processor.”* Their glossary is blunter — the entity to which authorisations are granted, **“thus the unit of accountability”**. Authorisation is granted to a principal *because* it is accountable.

RFC 4949 defines the same word without accountability in it at all: *"A specific identity claimed by a user when accessing a system… equivalent to the notion of login account identifier."* And it adds: *"Each principal can spawn one or more subjects, but each subject is associated with only one principal."*

A system can satisfy RFC 4949 completely and contain no accountable party.

The consequence for agents is exact. **An agent holding an operator's API key is not a principal.** It is a *subject* operating that operator's principal — a second body using one person's standing. Everything it does is, in the only sense the record can support, that person's act. A supplier describing this as "the agent has its own identity" is describing a subject and charging for a principal.

One body did once require the accountable version. FIPA's Agent Management Specification, SC00023K, made it a condition of being an agent at all: *"An agent must have at least one owner… and an agent must support at least one notion of identity."* Owner and identity, both checkable, written into a standard. ⚠️ `fipa.org` today serves a gambling affiliate site; the Internet Archive's capture timeline places the change between 21 July and 5 September 2026, and the specifications survive in the archive.

The property required: the acting software is distinguishable from the person on whose behalf it acts, at the point where the act is recorded, without the person having to remember which it was.

The second property: a capability that is not ambient

Ambient authority is authority that is exercised but not chosen. Miller, Yee and Shapiro, in *Capability Myths Demolished*:

> We will use the term ambient authority to describe authority that is exercised, but not selected, by its user… the caller of a function such as open() does not choose any credentials to present with the request; **the request merely succeeds or fails.**

Their image is a world of doors without keys. You approach, and the door opens if it deems you worthy. You presented nothing and selected nothing, so you cannot afterwards say which authority you were using — you were not using one. You were permitted.

⚠️ Citation care: the term is often attributed to Mark Miller's 2006 dissertation. It does not appear there. The document runs to 229 pages and contains the phrase zero times; the 2003 paper is the source.

Almost all current tool-calling is ambient by this test. An agent holding a broadly scoped token, a shell, or a filesystem call does not select which authority it exercises. It acts, and the act succeeds or fails. On the technical meaning of the word, that is not delegation.

The property required: the acting software must *name* the authority it is using, so that a record of what it did is also a record of what it claimed to be entitled to.

The third property: authority that narrows

Two principles are routinely treated as one.

Least privilege is Saltzer and Schroeder's design principle (f): *"Every program and every user of the system should operate using the least set of privileges necessary to complete the job."*

Least authority looks identical and is not, and Miller records the difference in his own footnote — it is not clear precisely what Saltzer and Schroeder meant by "privilege". His distinction, from §8.1 of the dissertation, carries this piece. *Permission* is what the access-control topology allows you to reach. *Authority* is what you can cause, including through anything that will act on your request:

> When Alice and Bob arrange this relying only on the "legal" overt rules of the system, we say Alice is providing Bob with an **indirect access right**… that she is **acting as his proxy**, and that **Bob thereby has authority to write it.**

An agent granted least privilege may hold unbounded authority, if it can ask something more privileged to act for it. Bounding permission is a claim about a diagram; bounding authority is a claim about consequences. Most least-privilege assurances about AI agents are the first kind offered as the second.

The property required: authority attenuates along a chain. A delegate cannot confer more than it holds, and cannot recover what it has given away by asking a third party.

The confused deputy, 1988 and 2025

In 1988 Norm Hardy described a compiler running with authority from two sources — the invoker's, and its own licence to write to its home files — and unable to tell them apart. *The Confused Deputy (or why capabilities might have been invented)*, ACM SIGOPS Operating Systems Review 22(4), October 1988:

> 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 part worth sitting with is that 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 arrived from outside the program, in a grant made elsewhere.

It is in the specification of the protocol most agent tooling now uses. The Model Context Protocol's authorization specification, version **2025-06-18**, 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.**

Its companion document, *Security Best Practices*, describes the same shape. ⚠️ A related passage frequently quoted alongside this one — that a single omnibus scope *"masks user intent per operation"* — is **not** in the 2025-06-18 revision. It appears under "Scope Minimization" in the **2025-11-25** revision. The live documentation URL for 2025-06-18 serves the later text, which is how the misdating occurs. The industry rediscovered the defect, wrote it into a specification, and the fix it names is attenuation — the third property above.

The fourth property: a record legible to somebody who was not there

A delegation nobody can reconstruct afterwards cannot be distinguished from no delegation. RFC 4949 sets the bar: a security audit trail is a record *"sufficient to enable the reconstruction and examination of the sequence of environments and activities surrounding or leading to an operation… from inception to final results."*

Two things this property is **not**, both of which are sold as though it were.

It is not proof the claims are true. W3C's verifiable credentials work is explicit that verification does not imply evaluation of the truth of the claims encoded. A verified record is one whose authorship can be checked. What it says may still be wrong.

It is not non-repudiation in the sense the word implies. RFC 4949: *"Non-repudiation service does not prevent an entity from repudiating a communication."* It distinguishes technical non-repudiation — a signature was made by a particular key — from legal non-repudiation, which turns on how well control of the private key can be established. For an agent holding a key on someone's behalf, that second question is the unanswered one. ⚠️ The IETF depre-

cates a competing definition of this term, instructing its own authors not to use it. That definition is published by the Committee on National Security Systems in CNSS Instruction 4009, **not** by NIST.

There is also a boundary that residency does not reach. What decides what a system does is not where a record is kept but **what is retrieved into the model's active reasoning environment**, because once a record is in context it shapes planning, output, tool selection and external action. A record that never leaves the country, pulled into context by a model running elsewhere, has crossed the only boundary that governed the outcome.

Two consequences follow that are not obvious. **Read is a separate grant from write**: a system that may read anything can place anything into context, and a read grant is therefore an influence grant. And **the model's own output is untrusted input**, as are tool descriptions supplied from outside.

The property required: the record is sufficient for somebody who was not present to establish who authorised what, within what limits, and what was done under it — and is not described as establishing more than that.

The live case: instructions hidden in what the agent reads

The failure the four properties are built against has a current, measured form. Indirect prompt injection places hostile instructions in material the agent will later retrieve rather than in anything a user typed. Greshake and colleagues named it in 2023, at the ACM Workshop on Artificial Intelligence and Security.

The prevalence work is recent. A 2026 study crawled **1.2 billion URLs across 24.8 million hosts** and found **15,300 validated injection instances across 11,700 pages**, roughly **70% of them in non-rendered HTML** — invisible to a person looking at the page. ⚠️ Post-cutoff. Earlier benchmark work found a ReAct-prompted GPT-4 acting on injected instructions in about **24%** of cases.

The counter-evidence belongs here too, and it cuts at the measurements rather than at the threat: a 2025 critique argues existing injection benchmarks suffer **"flawed success metrics, implementation bugs, and most importantly, weak attacks"**, and are easily saturated. Take the prevalence finding as established and the success rates as contested.

The reason this sits in a piece about architecture rather than about model safety is that none of the four properties depends on the model resisting the instruction. An agent that must name the authority it is exercising, that cannot widen it by asking something else, and that

writes what it did into a record it cannot alter, is an agent whose compromise is bounded by what it was entitled to do. Filtering the input is worth doing and cannot be relied upon, and bounding the authority does not require the filter to work.

What the standards ask of an agent

Nothing found requires an AI agent to hold an identity of its own. SPIFFE specifies a workload identifier with no notion of ownership, responsibility or accountability anywhere in it. RFC 8693's token exchange carries a delegation chain in its `act` claim but treats prior actors as informational. NIST SP 800-63-4, finalised in July 2025, states its scope as the identity of *users*.

The nearest thing to a requirement is a control that exists and was then made optional: NIST SP 800-53 **IA-9** requires that system services and applications be uniquely identified and authenticated before communicating — and it is in **no baseline**. Not Low, Moderate, High, Privacy, or the operational-technology overlay.

Work is under way. ⚠️ Two Internet-Drafts dated 2026 propose agent identity frameworks, one stating that *"agent identity does not replace principal identity; it supplements it"* and requiring that each link in a delegation chain carry its own cryptographic binding — which is the third property, written as a protocol requirement. Both are individual submissions, and a 2026 survey names **"recursive delegation accountability"** among gaps that **"no current technology or regulatory instrument resolves"**. Post-cutoff, and unverifiable against this author's knowledge.

FIPA in 2004 remains the only standard found that made an owner a requirement of being an agent at all.

How you would know this is wrong

First, if a system can be shown to satisfy the accountability the four properties aim at by some other route — if filtering, monitoring and review reliably produce an account of who authorised what — then the properties are one implementation among several and the claim that they are necessary is too strong.

Second, this argument rests on an assumption about history that it does not defend: that access control lists prevailed on convenience, in an era when the thing acting was a program a person had started, and that what is acting has since changed. If capability-style architec-

tures were in fact tried at scale and lost on correctness rather than convenience, that assumption fails and the argument should be withdrawn rather than restated.

Third, on the record: there is no floor stated here for how much a record must contain. Too little and nothing can be reconstructed; too much and the record becomes a surveillance artefact that this same argument would object to. That tension is real and unresolved, and naming it is not the same as answering it.

Related

How much it may do unsupervised sets the level of autonomy an institution permits; this piece sets out what must be true for that permission to mean anything.

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

- What has to be settled first — why the prior questions are prior, and what each of the others answers
- 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
- 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

```
The four properties underneath, version 0.2 (September 2026).
agenticgovernance.digital/downloads/what-has-to-be-true-underneath-v0-2.pdf
```


PREVIOUS
How much it may do unsupervised

Series contents

NEXT
What you can actually require

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