

Where Verification Ends: The Formal Limits of Artificial General Intelligence Safety Certification and the Institutional Siting of Judgement

Dr. Chris Cleverly

Barrister, Independent Researcher, United Kingdom

Corresponding author: Dr. Chris Cleverly, Barrister, Independent Researcher, United Kingdom.

Submitted: August 22, 2026

Accepted: August 31, 2026

Published: September 10, 2026

Citation: Chris Cleverly (2026) Where Verification Ends: The Formal Limits of Artificial General Intelligence Safety Certification and the Institutional Siting of Judgement . Epistora J. Artif. Intell. & Intell. Syst. 1(1), 1-50. Article EJAIS-109

Abstract

Public and academic discussion of artificial general intelligence safety often presumes an eventual deliverable: a demonstration, established once, that a highly capable self-modifying system will behave beneficially thereafter. I argue that no such deliverable is available, and that its unavailability follows from five established results rather than from any current shortfall in verification technology. Gödel's (1931) second incompleteness theorem blocks a system from proving its own consistency. Löb's (1955) theorem shows that the natural weaker substitute, a system that trusts its own proofs, collapses into triviality, and the obstruction is sharpest for selfmodification (Yudkowsky & Herreshoff, 2013). Tarski's (1933/1956) undefinability theorem rules out a complete internal predicate classifying a system's own future conduct as safe. Turing's (1936) halting problem and Rice's (1953) theorem generalise the point without invoking selfreference. Trakhtenbrot's (1950) theorem closes the apparent escape of restricting the system to finite hardware. A supervisory regress argument (Gumbau Mezquita, 2026) shows that delegating certification to progressively more capable external verifiers relocates the gap rather than closing it. These results are conditional on a system being computationally general and self-modifying, so I then examine how closely that antecedent is presently approached. Bounded self-improvement is established as industrial practice; fully closed-loop self-improvement remains prospective; the principal disclosures are self-reported, and the coordination machinery that would verify them is absent from every governance instrument now in force. I then address an inference these results are often taken to license: that because no formal verifier suffices, human judgement must be the terminus. That inference either makes humans a further rung of the same tower, inheriting the same limits, or rests on an unargued claim that human cognition escapes formal-system-hood, which has a long history of criticism (Franzén, 2005). Rejecting the inference leaves the conclusion standing on different ground. I rebuild the case for human authority on openly normative grounds: accountability, which no formal verifier can bear (Matthias, 2004; Santoni de Sio & van den Hoven, 2018), and testimony understood as an independent source of knowledge with its own auditable validity conditions, drawing on the classical Nyāya treatment of śabda alongside contemporary epistemology of testimony (Matilal, 1986; Coady, 1992; Lackey, 2008). The paper closes by specifying the institution this disclosed argument supports: a standing, accountable body sited at a declared terminus depth, which never certifies safety, issues fourvalued rather than binary verdicts, and holds only the power to halt. Its composition is constrained by a proved result. Because no symmetric, Sybil-proof, non-trivial reputation function exists (Cheng & Friedman, 2005), a pool admitting members on their own declared properties is vulnerable to a single interest registering many nominally independent bodies. Open entry must therefore be surrendered. What replaces it is a relational test rather than a certifying authority: admission proceeds by recognition from bodies already admitted that have no reason to favour the candidate, assessed by the sparseness of the boundary between the candidate and everything else, with a qualifying judgement -

made afterwards by an accountable party who publishes reasons and can be overruled. A firewall accompanies this. Bodies admitted on such grounds advise and supply no criteria, while the deciding panel is drawn by lot from a general population pool; without that separation the proposal reduces to traditional authorities adjudicating technology by their own lights. The residual risk of an adversary acquiring genuine standing across mutually independent bodies over a sufficient period is recorded as unclosed.

Keywords: Artificial general intelligence, Formal verification, Incompleteness, Löb's theorem, Safety case, accountability, Epistemology of testimony, Nyāya, Institutional design

Introduction

There is a document that discussions of artificial general intelligence safety keep almost reaching for. It carries a seal. It states that a system has been examined and will behave beneficially as a settled matter, going forward, including after the system has rewritten its own code. Few researchers assert the existence of such a document in these terms. It operates instead as the implicit shape of the target, the thing that would close the question the way a proof closes a conjecture. Bostrom (2014) and Yudkowsky (2008) frame alignment as a problem admitting of solution; Amodei et al. (2016), Russell (2019) and Seshia et al. (2022) frame it in more explicitly bounded and probabilistic terms.

I argue that the document cannot be issued, for reasons established in mathematical logic and computability theory across the twentieth century, and that these reasons strengthen as the system in question becomes more capable. I then argue that a widely drawn conclusion from this impossibility, that human judgement must serve as the final certifier, does not follow from it. The versions of the argument that appear to derive it either smuggle in an unargued exemption for human cognition or convert an elimination into an existence claim. Human authority over these systems should nonetheless be insisted upon, on normative and institutional grounds. Once those grounds are stated openly they specify a rather particular kind of institution: one that never certifies anything, publishes where its own competence stops, and holds only the power to say no. Between the first argument and the second I pause to ask how closely present systems approach the condition the impossibility results assume, since a body of theory whose antecedent is never satisfied would be of limited practical interest.

Each formal result below is stated first in ordinary language and then precisely. Readers who know the mathematics may skip the informal statements. Nothing later depends on machinery introduced without explanation.

Scope: What This Paper Does Not Claim

A superficially similar argument is among the more thoroughly criticised positions in twentieth-century philosophy of mind, and I want to separate this one from it at the outset. Lucas (1961) and later Penrose (1989) argued from Gödel's theorems that human minds transcend formal systems, or cannot be machines. Franzén (2005) gives the standard treatment of why the reasoning fails. Among its defects are the unearned assumption that a human mind constitutes a single consistent formal system in the relevant sense, and the slide from "cannot prove its own consistency" to "is unreliable," or in the other direction to "is superior to."

Nothing here compares machine cognition to human cognition, claims that any system is untrustworthy, or predicts any particular failure. The claim concerns a proof target rather than a mind. For any sufficiently expressive formal system, whether mechanical, biological or institutional, a specific kind of self-issued unconditional guarantee is not among the things that can be proved. This is a narrower reading than the Lucas–Penrose argument requires. It is also, as Section 8 develops, a reading that cuts against my own preferred conclusion before that conclusion is rebuilt on different ground. Declining to exempt myself from my own argument is why Section 8 exists.

Two Preliminaries What a formal system is, and when self-reference becomes unavoidable

A formal system is a set of symbols, rules for combining them into well-formed statements, and -

rules for deriving statements from other statements. Arithmetic, formalised, is the standard example: symbols for numbers and operations, rules for what counts as a valid equation, and rules of inference for deriving new truths from established ones. A computer program is a formal system in the relevant sense. So is a mathematical theory, and so, for the purposes of any attempt to verify it, is the reasoning apparatus of an artificial intelligence.

The decisive discovery of the 1930s was that once a formal system becomes powerful enough to describe elementary arithmetic, it thereby becomes powerful enough to describe itself. Gödel's technique, now called Gödel numbering, assigns a unique number to every statement and every proof in the system, so that statements about numbers can be read as statements about the system's own statements and proofs. Nagel and Newman (1958) give the classic accessible exposition; Smullyan (1992) and Kleene (1952) give the technical development. Self-reference arrives automatically as a consequence of arithmetic capability, and cannot be removed without removing the capability.

The results that follow therefore have an initially counterintuitive shape. The expressive power that makes a system worth verifying is the power that places it beyond a certain kind of verification. In most engineering domains, capability and verifiability trade off gently, and better tools narrow the gap. Here the trade-off is neither gentle nor a matter of tooling. Past a threshold of expressiveness, a system acquires questions about itself that it cannot settle, and it acquires them as a mathematical consequence of that expressiveness.

What is meant by AGI in this paper

I use artificial general intelligence in a deliberately minimal and technical sense: a system that is computationally general, meaning capable in principle of any computation a universal computer can perform; capable of modifying its own operation; and capable of extending its own repertoire of actions in ways its designers did not enumerate in advance. No claim about consciousness, understanding or human-likeness is intended or required. The arguments below need only these three properties, and they apply to any system possessing them regardless of how it was built or what it is made of.

Five Results Gödel's second incompleteness theorem: a system cannot vouch for itself

Informally: a formal system capable of describing arithmetic cannot prove that it will never contradict itself, unless it in fact does contradict itself, in which case it can prove absolutely anything and the proof is worthless. Precisely: Gödel (1931) showed that for a consistent formal system F containing elementary arithmetic, F cannot prove the statement $\text{Con}(F)$ asserting F 's own consistency. Rosser (1936) weakened the hypotheses required; Kleene (1952) provides the standard development. Chaitin (1974) reached a related conclusion by an information-theoretic route, showing that a formal system of given complexity can establish only finitely many facts about objects more complex than itself.

Applied to an AGI: to the extent that a system's safety-relevant self-monitoring is formalisable, and if it is not formalisable then the project of formally verifying it has already been abandoned, that formalisation cannot establish from within that the system's future reasoning will never license a contradiction. It therefore cannot establish that the system will never license an unsafe conclusion by steps that each appear valid. I am not claiming that the system is inconsistent or that it will fail. Self-certification is simply unavailable as a method. The scope of the theorem is narrower than the reading sometimes attributed to it. Gödel's second theorem blocks unconditional, self-grounded certification. It does not block a proof of F 's consistency in a strictly stronger external theory, and relative consistency proofs are ordinary mathematical currency. Such a proof relocates the question rather than answering it, for reasons developed in Section 5.

Löb's theorem: self-trust is worse than the problem it solves

Informally: a system that adopts the general principle "whatever I prove is true" can thereby prove anything at all, including falsehoods. The modest-sounding assurance turns out to be catastrophic.

Precisely: Löb (1955) showed that for a suitable formal system, if the system proves that the provability of a sentence implies that sentence, then it proves the sentence outright. Boolos (1993) develops the result within provability logic, where it appears as one of the governing laws of how provability behaves rather than as pathology.

This is the least widely appreciated of the results assembled here, and for AGI it may be the most important. Suppose one concedes that unconditional self-certification of consistency was always too much to ask, and retreats to something weaker: a system that trusts its own proofs, which seems the minimum any competent agent needs in order to act on its own reasoning. Löb's theorem says the retreat is not to safer ground. Blanket self-trust is formally degenerate.

The result applies most directly to the capability that motivates AGI safety concern. An agent that constructs or rewrites itself into a successor would, to hand over authority responsibly, want to establish that the successor's proofs are reliable. Yudkowsky and Herreshoff (2013) formalised this requirement and showed that it runs into Löb's theorem, framing the difficulty as one of "tiling" an agent's trustworthiness across successive self-modifications. Fallenstein and Soares (2015) survey the ensuing programme of partial results under the heading of Vingean reflection, meaning reasoning about a successor whose specific conclusions cannot by construction be anticipated (Vinge, 1993). Garrabrant et al. (2016) address the adjacent problem of assigning coherent probabilities to mathematical claims one has not yet had time to settle.

The character of this literature matters. It was produced by researchers strongly motivated to find a formal route to trustworthy self-modification, which makes its findings closer to adversarial testimony than to sympathetic commentary. Its partial results are uniformly partial in the same direction: they secure trust under restriction, whether bounded proof length, restricted logical complexity, a specific successor architecture, or a finite chain of modifications, rather than in general. The authors say as much themselves, closing their abstract with the observation that after all their constructions "the underlying puzzles of rational coherence are thus only partially addressed" (Yudkowsky & Herreshoff, 2013, p. 1); their concluding section adds that no criterion of action for bounded self-modifying agents can presently be stated even at the level of generality of satisficing or meliorising. I read that pattern as the same boundary reappearing under independent attempts to cross it.

Tarski's undefinability theorem: no complete internal safety predicate

Informally: a sufficiently rich language cannot contain a correct and complete definition of truth for its own sentences. Any attempt reconstructs the liar paradox.

Precisely: Tarski (1933/1956) showed that for a sufficiently expressive formal language L , no formula definable within L correctly picks out exactly the true sentences of L .

The application is to a design that might otherwise seem natural, and that some proposed safety architectures approximate. Equip the system with an internal predicate, call it $\text{Safe}(x)$, ranging over descriptions of its own possible future actions, and have it consult that predicate before acting. If $\text{Safe}(x)$ is to be complete over a domain including the system's own statements about its own future safety assessments, and defined within the system's own language, Tarski's theorem applies directly: it cannot be both complete over that domain and consistent. The standard remedy is to lift the predicate into a strictly stronger metalanguage. That remedy is available and useful, and it reproduces the same situation one level up, since the metalanguage's truth predicate for itself is undefinable within itself. The resulting hierarchy of languages has no terminal level. Hofstadter (1979) discusses the pattern at length under the heading of tangled hierarchies, where a structure that appears to ascend cleanly towards completeness instead reproduces its own gap at every level.

This result is independent of the two preceding ones rather than a corollary of them. Gödel's and Löb's theorems concern what a system can prove; Tarski's concerns what it can define. A system could in principle be limited in one respect and not the other. Their convergence on the same practical conclusion, no complete internal self-assessment, tells us something about the target rather than revealing a redundancy in the argument.

Turing and Rice: undecidability without self-reference

A reader might reasonably suspect that the whole difficulty is an artefact of self-reference, and that a disciplined design forbidding the system to reason about itself would avoid it. The next two results close that route. They are in some ways the most damaging, because they invoke no self-reference at all.

Turing (1936) established that no algorithm can decide, for an arbitrary program and input, whether that program will eventually halt. Rice (1953) generalised this dramatically.

For any nontrivial semantic property of programs, meaning any property that holds of some computable functions and not others and that depends on what the program computes rather than on how its code is written, no algorithm decides whether an arbitrary program has that property. Undecidability is the generic condition of questions about what programs actually do.

"Does this system's behaviour remain within a specified safety envelope, across all future inputs and all self-modifications it may undergo?" is a non-trivial semantic property in exactly Rice's sense. A general-purpose safety verifier, an algorithm accepting an arbitrary sufficiently expressive program and returning a verdict on this property, is a request for a decision procedure that provably does not exist. The argument requires no self-modelling and no reflexivity. It requires only that the system be computationally general and that safety be a non-trivial property of its behaviour. Moving the verifier onto separate hardware, into a different organisation, or under a different jurisdiction changes nothing, since the undecidability attaches to the question and not to the questioner. Wolpert (2008) derives related limits on inference for physically embedded inferring devices, obtaining obstructions that additional computational resources do not remove.

Rice's theorem is frequently overread and should be stated at its correct scope. It blocks a total decision procedure over the class of all programs. It leaves untouched the sound-but-incomplete verification of one specific constructed artefact, which is what practical verification programmes actually perform and what the safety-case tradition described in Section 13 is built around.

Trakhtenbrot's theorem: the finite-hardware escape does not work

A further natural objection holds that Rice's theorem concerns algorithms over unbounded domains, whereas real hardware is finite. Confine the system to a finite state space and every behaviour becomes checkable by exhaustive search. Melo et al. (2025) develop this position, and their paper requires careful placement, since it is a peer-reviewed independent confirmation of my core claim rather than an objection to it. They prove the undecidability of inner alignment by reduction to the halting problem via Rice's theorem, and their related-work discussion assembles further independent undecidability results for the harming problem, the containment problem, the control problem, monitorability, and compliance with stated ethical constraints. Their constructive proposal, restricting AGI architectures to finite, halting configurations assembled from provably aligned operations, is the finite-hardware move. It is that proposal, rather than their impossibility result, which the following paragraph addresses.

Gumbau Mezquita (2026) shows the escape fails at the level that matters. Certifying a safety property for one specific, fixed finite configuration is indeed decidable by exhaustive search. An operational guarantee requires more, because self-modification moves the system between configurations, and what is needed is certification across the entire class of finite configurations the system might come to occupy. Trakhtenbrot's (1950) theorem establishes that validity over the class of all finite structures is not merely undecidable but co-recursively-enumerable-complete, meaning informally that no algorithm can even list the valid cases, let alone decide them. The finite-hardware move trades one undecidable problem for another. Gumbau Mezquita formalises the resulting bind as a trilemma between soundness, completeness and tractability, sharpening into a structural necessity a closely related characterisation given independently by Agarwal (2026). The result is his Theorem 4, with Corollary 2 drawing the consequence that concerns us: certifying a safety property for one fixed finite structure is decidable by exhaustive search, while certifying it across all finite configurations is co-RE-complete. His Theorem 6 then stratifies the bounded case into two clauses that the trilemma of his Corollary 3 rests on, a local boundary where checking a single fixed finite structure is PSPACE-hard in the worst case for specifications requiring partial fixed-point expressivity, and a universal boundary which is Theorem 4 restated. Only the second is undecidability in the strict sense; the first is intractability, which is his stated reason for using unverifiability rather than undecidability as the umbrella term.

The relationship between these two papers bears on how much weight the impossibility case can carry, so I should state it explicitly. Melo et al. establish the undecidability result independently, by a different route, in a peer-reviewed venue, roughly a year before the preprint discussed here. The two agree on the negative result and disagree on whether a finite configuration restriction escapes it.

My impossibility case does not depend on resolving that disagreement, since the disagreement concerns a proposed remedy rather than the result it is a remedy for.

Convergence Without Redundancy, and the Supervisory Regress

The five results are logically independent. None is a special case of another, but they close related escape routes, and the difference between convergence and repetition matters here. The natural response to any one of them is to route around it. If the system cannot certify itself, station a more capable external verifier outside it. If that verifier has the same problem, station a further verifier above it, and continue until the chain terminates in something simply trusted.

This fails twice over, and the two failures are independent. It fails first by the second incompleteness theorem applied to each verifier in turn. Every verifier's own reliability is exactly as unestablished as the original system's was, so the chain does not terminate in anything selfguaranteeing. Gumbau Mezquita (2026) proves this formally. In a tower of supervising systems, where any supervisor adequate to audit a computationally general, self-modifying system must itself be computationally general and self-modifying, the same undecidability recurs identically at every level. No finite tower yields an unconditional certificate for the system at its base, and the infinite tower contains no level at which certification becomes decidable. A supervisor weak enough to remain decidable is too weak to observe the violation it exists to catch; a supervisor strong enough to observe it thereby inherits the undecidability. Conant and Ashby's (1970) goodregulator theorem shows why the bind is structural rather than contingent. Every good regulator of a system must contain a model of that system, so adequacy of oversight demands modelling power, and modelling power is what invites the limitation. The theorem's hypotheses should be stated, since they are not trivial: each supervisor is required to be Turing-complete and selfmodifying under a transition operator that is semantically well-defined (his Definition 5) and Pdisruptive relative to the nowhere-defined function (his Definition 6). The first requirement he defends rather than assumes: a supervisor confined to a strictly weaker class than the system it audits is provably evadable, so an adequate supervisor of a general system is itself a general system.

It fails second by Rice's theorem, which concerns the non-existence of any algorithm deciding a non-trivial semantic property, and not merely the non-existence of a self-directed one. Relocating the decision leaves it an instance of a provably undecidable general problem.

This bears on a substantial body of practical work. The research programme known as scalable oversight seeks arrangements in which weaker or comparably capable systems oversee stronger ones, through mechanisms including reward modelling learned from human preferences (Christiano et al., 2017; Leike et al., 2018) and debate-based amplification of human judgement (Irving et al., 2018). The regress result does not show these techniques to be worthless. They may substantially raise the practical cost of an undetected failure. It does bound what they can deliver. They are caught between two constraints, and Gumbau Mezquita (2026, §11) states the bind directly. An overseer strictly weaker than the system it audits is an L-restricted verifier in the sense of his Definition 13, inspecting a proper, bounded fragment of behaviour; his Proposition 3 constructs a self-modification that violates the safety property entirely within that verifier's blind spot, and his Corollary 7, which he calls the Paradox of Weakness, shows the restriction is not incidental but forced, since no algorithm can complete an inspection of an infinite domain. An overseer strong enough to close the blind spot must be at least as expressive as the audited system, and his Theorem 8 then applies to it in turn. Debate, amplification and recursive reward modelling each choose a position on this spectrum; his conclusion, which I adopt, is that the spectrum has no safe interior. Yampolskiy (2017) introduced the term unverifiability for the general phenomenon, arguing informally that the verifier regress is the organising problem of the field. Gumbau Mezquita's (2026) contribution, his Corollary 5, is to show that the regress provably fails to terminate: a finite tower always has an uncertified top, and the infinite tower contains no level at which certification becomes decidable.

A sixth result completes the picture by addressing bounded, practical verification schemes rather than exact certification. Gumbau Mezquita's Theorem 9 (2026, §15) proves that for any non-trivial safety property not satisfied by the nowhere-defined function, every bounded verification scheme that is faithful, meaning that it does not reject evidence correct by its own lights, admits an evolution trace it certifies at every stage while -

the property is in fact persistently violated throughout. His Proposition 4 disposes of the remaining case: a non-faithful scheme either never certifies anything or already rejects a genuinely safe system without any adversarial trace at all. The construction descends from Cantor's and Turing's diagonal arguments, and he says so explicitly in the remark closing that section, noting with Breic and Yampolskiy (2023) that impossibility results of this kind share a diagonal, fixed-point technique. A scheme inspecting only a finite window at each stage has, by construction, a region it never examines, and a trace can always be built whose violation lives exactly there. This should be distinguished from lumpability in the Markov-process sense, which concerns whether coarse-graining commutes with dynamics and is a separate question. The phenomenon here is observational, concerning what a bounded observer can distinguish.

Verification against physical reality

One escape route remains, and it is the most natural of those available, because it appeals to an intuition I rely on elsewhere. The results above concern a system verified against itself or against other symbolic systems. Suppose instead the system were verified against what it is not: the physical world. Its symbolic operations are descriptions; the world is what they describe. Perhaps the gap between the two supplies the external standard that internal methods lack, and physical reality is the verifier that no formal system can be.

The proposal deserves an answer rather than a dismissal, and four considerations close it. The first is timing. The physical world does deliver verdicts on whether a system was safe, and they are perfectly reliable ones: collapses, failures, harms. They are also retrospective without exception. The world pronounces after the system has acted. It is therefore a complete verifier of what happened and silent on what has not yet happened. A deployment decision requires something prospective, and anticipation of a verdict is not the verdict. The proposal relocates the gap into the interval between action and consequence, which is the interval the institutional design of this paper exists to occupy.

The observation is old enough to be worth dating, and it was made in this shape rather than a looser one. Wiener (1960) put it in *Science* in the year the difficulty first found an audience, arguing that the danger is not that a machine lies beyond interference but that interference depends on information which arrives too late to be used:

“when a machine constructed by us is capable of operating on its incoming data at a pace which we cannot keep, we may not know, until too late, when to turn it off” (p. 1357).

He develops the point:

“there is the tale of the “Monkey’s Paw,” by W. W. Jacobs, in which the sergeant major brings back from India a talisman which has the power to grant each of three people three wishes. Of the first recipient of this talisman we are told only that his third wish is for death. The sergeant major, the second person whose wishes are granted, finds his experiences too terrible to relate. His friend, who receives the talisman, wishes first for £200. Shortly thereafter, an official of the factory in which his son works comes to tell him that his son has been killed in the machinery and that, without any admission of responsibility, the company is sending him as consolation the sum of £200. His next wish is that his son should come back, and the ghost knocks at the door. His third wish is that the ghost should go away.”(p.1358)

The essay then states the consequence in a form closer to my Section 15 than to a fable: where an agency acts faster and more irreversibly than we can assemble grounds for stopping it

The parables differ from the case in one respect, and it is the respect Section 9 declines to grant.

In the tales there is a third wish.

The refutation, published in a later issue four months afterwards, is more useful than its title suggests. Samuel (1960) held that a machine has no will, that what looks like intention is the programmer's intention or something derived from it under rules the programmer set, and that to think otherwise is to believe in magic or to deny that human volition differs in kind from machinery. The argument is sound against the position it addresses and does not reach mine, which concerns what can be established about a system rather than what the system wants. What matters here is the paragraph he added at the end:

“An apparent exception to these conclusions might be claimed for projected machines of the so-called “neural net” type... Since the internal connections would be unknown, the precise behavior of the nets would be unpredictable and, therefore, potentially dangerous... If practical machines of this type become a reality we will have to take a much closer look at their implications than either Wiener or I have been able to do.” (p. 742) The most confident refutation available in 1960 carried an expiry condition on its final page. The condition has since been met, and Section 7 describes what it looks like when it is.

The second is representation. To verify a system against the world, one requires a standard against which to compare it, and the standard is not the world itself but a model of it: incomplete, approximate in its precision, bounded in scope, and requiring interpretation to determine what counts as agreement. The comparison therefore runs between two representations, and the model is a formal object subject to everything established above. This is a remark about the structure of verification rather than about the limits of human knowledge. A standard must be represented in order to be compared against. Rushby's (1995) account of avionics certification makes the point in practice, identifying the irreducible gap between the model used to certify and the world certified in as the reason formal methods are scoped to bounded claims rather than to aircraft.

The third is the scope of Rice's theorem, which extends to physical properties rather than stopping at symbolic ones. Given a fixed interface by which a system's outputs produce effects, which any deployed system has, the question whether those effects remain within a specified envelope is a non-trivial property of what the system computes: true of some programs and false of others, and independent of how the program is written. The theorem applies to it as to any other such property, whether the property is specified internally or externally, and whoever performs the check. Describing the envelope in units of energy or distance rather than in symbols does not alter the structure of the question being asked.

The fourth is the relationship between the system and the world it would be checked against. A system is physically instantiated within the world; its operations are physical events; it is not positioned outside looking in. Verifying the system against the world therefore means verifying one part of the world against another part, which reproduces at the physical level the structure that generates the difficulty at the symbolic one. Wolpert (2008) derives the corresponding result directly, showing that inference devices embedded in the systems they infer about face limits that additional computational resources do not remove. Breuer (1995) established independently that no observer can distinguish all the present states of a system in which it is contained, a result treated in Section 6, and one which converts this reply from a structural observation into a proved limit. Section 9 develops the constructive use of physical observables, and the evadability result stated there applies to them exactly as to symbolic ones: a monitoring scheme reading a finite set of physical indicators has, by construction, a region it does not read.

The physical world is the final verifier, and this is the reason my argument takes the shape it does rather than a concession. Its verdicts are events rather than statements, they arrive after the fact, and they cannot be consulted in advance. What must be done before they arrive is judgement, and the question of who is competent to exercise it occupies the sections that follow.

Taken together: no verifiable safety claim about a computationally general, self-modifying system is unconditionally self-grounded; no finite tower of formal external verifiers closes the gap; and no bounded practical scheme is immune to systematic, undetectable evasion. This is the fully scoped impossibility result, and it is established rather than conjectural.

One qualification on that sentence is required by my own discipline. The five results of Section 4 are long-established theorems with settled proofs and an extensive secondary literature. The supervisory regress and universal evadability results are not yet in that position. They are stated in a preprint (Gumbau Mezquita, 2026) which had not, at the time of writing, completed peer review, as is the closely related independent characterisation by Agarwal (2026). Their arguments are checkable and are reproduced in outline above, and the negative core they share with Melo et al. (2025) is independently peer-reviewed. A reader is entitled to know which parts of this section rest on results the discipline has examined at length and which rest on results it has not, so I record the distinction here rather than leave it to be discovered.

Embedded Inference: The Obstruction in Physical Form

The results assembled so far concern formal systems and algorithms. An artificial general intelligence is neither of those things in the first instance. It is a physical device, built inside the world it is required to reason about, and any verifier of it is another such device. A reader who accepted every argument to this point might still ask whether the obstruction is an artefact of treating these systems as mathematical objects, and whether a properly physical analysis would find the situation different. The obstruction survives translation into physical terms without weakening. In one respect it becomes sharper, because the physical treatment reaches a conclusion about self-directed inference that the logical results reach only by way of arithmetic. Inference devices

Wolpert (2008) develops a framework whose generality is unusual even among impossibility results. Its central object is the inference device: a pair of functions defined over the space of all worldlines of a universe consistent with its physical laws. One function, the setup, represents how the device is configured or initialized. The other, the conclusion, represents what it ends up asserting, taken as binary. The framework covers prediction, observation, memory and control as instances of a single structure, differing only in the temporal relation between the device and what it is inferring about. The device is his Definition 1: a pair of functions over the space of worldlines, the setup function and the conclusion function, with the latter surjective onto the two-element set of conclusions.

Inference in this setting is defined by what the device can be made to get right in advance. A device weakly infers some quantity when, for every binary question that might be asked about that quantity, there exists some configuration of the device under which its conclusion is guaranteed to answer that question correctly. This is a demanding standard and also a natural one. It is what we mean when we say an instrument is reliable for a given measurement, or that a model can be relied on for a given prediction. This is his Definition 3, formalised through what he calls probes (Definition 2), which represent the range of a quantity as a set of binary questions. What makes the framework worth importing here is the robustness of what follows from it. Wolpert (2008) states that his results do not involve chaotic dynamics or quantum indeterminism, apply independently of the values of any physical constants, and presume no limitation on where in the Chomsky hierarchy the device lies; they would apply to oracles, he notes, if a universe can contain them. They hold for devices with unlimited computational resources, and they would hold in any universe with any physics. The single operative feature is that the device is built inside the system it is inferring about.

What such devices cannot do

Three results matter for the present argument. I state them descriptively rather than by their numbering in the source, since a reader importing them into another discipline should check the originals rather than rely on a secondhand restatement.

The first is that for any inference device whatever, there exists a binary-valued quantity that the device cannot infer. Applied to prediction, this refutes the Laplacian picture (Laplace, 1814/1951) without requiring chaos or quantum mechanics: even in a classical, fully deterministic, non-chaotic universe, there is no general-purpose infallible predictor. Applied to observation, Wolpert notes it can be read as an uncertainty principle arrived at without quantum mechanics.

Applied to memory, it says there is no general-purpose infallible recording device. This is his Proposition 1(ii). He draws the Laplacian consequence in the discussion following it, together with the readings for observation and for memory.

The second is that no two devices which are distinguishable, meaning roughly that neither one's configuration constrains the other's, can each weakly infer the other. This is stronger than the first result, because it identifies a whole class of quantities a device cannot infer rather than merely establishing that the class is non-empty: it cannot infer the conclusions of any device that is distinguishable from it and that can infer it. The proof runs through a construction Wolpert himself describes as a higher-order version of the liar paradox and of Cantor's diagonal argument, which is the machinery of Section 4 arriving in a physical setting. This is his Theorem 1, with distinguishability as his Definition 4. He is explicit that distinguishability is essential to it: mutual weak inference can occur between non-distinguishable devices. The proof, like most in the paper, he describes in the preamble to his Appendix A as a higher-order version of the Cretan liar paradox and of Cantor diagonalisation.

The third is that no two devices can strongly infer each other, where strong inference is the analogue of one machine emulating another completely. This result requires no distinguishability assumption at all, and it has a corollary of some interest: no universe can contain more than one universal inference device. This is his Theorem 3, with strong inference as his Definition 5. He notes that, unlike Theorem 1, it requires no distinguishability assumption. The corollary follows in his discussion of realities, where the universal device is defined as one that strongly infers every other device and weakly infers every function in its reality.

Self-aware devices, and the sharpest of the results

The framework then introduces a refinement that bears directly on the internal safety predicate discussed in Section 4. A self-aware device carries, in addition to its setup and conclusion functions, a question function: a representation of the question the device takes itself to be answering. The device is infallible when its conclusion always correctly answers whatever question it takes itself to be asking. This is the device-theoretic counterpart of a system that maintains an internal model of what it is currently assessing, which is what a system consulting its own safety predicate is doing. The self-aware device is his Definition 12, adding a question function to the setup and conclusion pair.

Two results about such devices bear on this paper. The first is that no two self-aware devices whose question functions range over finitely many questions are mutually intelligible. For at least one of any such pair, there is a question the other might be asking that the first cannot so much as pose, let alone answer. This holds without any distinguishability assumption, and it holds even when the two devices are identical. A self-aware device with a finite question range is therefore not intelligible to itself. This is his Corollary 4, and he states both features I rely on: it does not depend on the devices being distinguishable, and it holds even when the two devices are identical, so a self-aware device with a finite question range is not intelligible to itself.

Intelligibility and infallibility are his Definition 13.

That last sentence is the closest thing in the literature to a physical statement of what Section 4 established logically. A system's internal model of what it is doing is systematically incomplete about itself, and the incompleteness does not follow from the system being a formal system in the arithmetical sense. It follows from the device being embedded in the world it is modelling, and it would hold for a device with unlimited computation running on any physics whatever.

The second result is that for any device at all, there exists a self-aware device that the first does not correct, where correcting means only that the first device has some configuration under which its conclusion reports whether the second correctly answered its own question. The definition of correction here is deliberately undemanding. It does not require the first device to assess every question the second might pose, and the impossibility holds anyway. This is the supervisory regress of Section 5 restated for physical devices. The checker cannot be constructed, and its non-existence does not depend on the checker being an algorithm. This is his Proposition 7, with correction defined at his Definition 14. He emphasises the weakness of that definition, which requires only that the first device have some configuration reporting whether the second correctly answered its own question.

Why Löb rather than Gödel is the right logical companion

Wolpert closes his discussion of realities with a speculative remark, which he does not pursue, that his results might be applied to mathematical logic itself by taking the underlying space to be a set of strings true under some encoding of axioms, with inference devices as fallible theorem-proving algorithms embedded in it. The natural question is which logical result the device framework most closely parallels. The usual answer, Gödel's incompleteness theorems, is not the best one.

Gödel's second theorem concerns what a system can prove, specifically that it cannot prove its own consistency. Löb's theorem concerns whether a system can treat its own provability as a reason to believe, which is subtly different and, for present purposes, more apt. It is the sharper statement of the same underlying phenomenon, and Gödel's second theorem falls out of it as the special case where the sentence in question is a contradiction. The parallel with self-aware devices is then direct. A self-aware device carries an -

internal representation of the question it takes itself to be answering, which is the device-theoretic counterpart of a system carrying an internal provability predicate. Löb's theorem says a system cannot assert the soundness of its own proof procedure without thereby asserting the conclusion. The self-awareness result says a device cannot pose the full range of questions about its own question-and-answer behaviour. Both say that a system's internal model of its own reliability is systematically incomplete, and both apply to trust rather than to provability alone.

This parallel is structural, and I do not offer it as a derivation. Whether either result can be obtained from the other under some translation is, so far as I am aware, open, and nothing in this paper depends on it. The parallel does explain why the emphasis of Section 4 fell where it did. Löb's theorem, rather than Gödel's, is the result that speaks to self-modifying agents deciding whether to trust their own successors, and it is also the result that the physical framework independently echoes.

The temporal gap, and what a dynamical extension would need

The device framework achieves part of its generality by abstracting away time. Because worldlines extend across all of history, the setup and conclusion functions already encode whatever temporal structure is relevant, which is what allows one construction to cover prediction, observation and memory alike. Wolpert notes explicitly that his formalisations do not depend on the time-ordering of the events involved. He makes the point twice, first in his discussion of the archetypal examples and again in the elementary results on weak inference, where he observes that time does not appear in the definition of inference at all, which is why prediction and retrodiction receive the same treatment.

This is elegant, and it leaves a gap that matters for the systems I am concerned with. In the framework, a physical computer at two different times is two devices that are copies of one another. A learning system at two different times is not a copy of itself: its setup function has changed and so has its conclusion function. Whether it can still infer what it could infer before is a question the static framework has no way to pose, because it contains no dynamics acting on devices.

A minimal extension would introduce device trajectories, meaning families of devices indexed by an ordered time set, stationary when all members are copies and evolving otherwise, together with three derived notions. Retentive inference asks whether a trajectory continues to infer a given quantity across an interval, and permits the question of when a capacity is lost. Diachronic inference asks whether a device at one time weakly infers the conclusion function of the same trajectory at another time, which is self-prediction when the later device is the target and selfretrodiction when the earlier one is. Diachronic distinguishability applies the framework's existing distinguishability condition to two members of a single trajectory, and this is where the interesting tension sits. No device is distinguishable from itself, but it is not obvious that a device is undistinguishable from its own substantially modified successor, whose configuration space may no longer be constrained by its predecessor's in the way self-identity requires.

Four questions then present themselves. All four are open and none is established here. Whether evolving trajectories contain diachronically distinguishable members, which would make the self-prediction barrier an immediate consequence of the existing mutual-inference result rather than a new one. Whether learning has an unavoidable cost in inferential coverage, meaning whether acquiring the capacity to infer some quantities forfeits the capacity to infer others, which would be a derived rather than observed analogue of catastrophic forgetting. Whether a selfaware device can verify in advance that its successor will correctly answer the questions the successor takes itself to be asking, which would give a device-theoretic statement of the Löbian obstacle that routes through neither arithmetic nor any particular physics. And whether the framework's bound on the product of two devices' accuracies at inferring each other admits a version across time, which is the point at which the analogy to uncertainty relations would either earn its place or be exposed as decorative.

The third of these is the substantive one for my argument, and I do not want to overstate its status. Wolpert's framework establishes the static case: for any device there is a self-aware device it does not correct. It says nothing about whether the uncorrectable device can be one's own successor. If that could be established, it would provide a statement about embedded agents holding independently of both the laws of physics and -

the axioms of arithmetic, which is the kind of result the framework was built to produce, and the kind that the study of selfmodifying systems presently lacks. Until it is established, it is a research direction and I present it as nothing more.

Self-measurement: an independent result, and a common misreading of it

A result obtained independently and more than a decade earlier reaches a closely related conclusion by a different route. It is frequently mischaracterised in ways that weaken it, so I state it carefully.

Breuer (1995) proved that no observer can distinguish all present states of a system in which that observer is contained. The mechanism is not exotic. A measuring apparatus that is part of the system it measures would have to encode, within its own state, the state of a whole that includes the apparatus itself, and it cannot. The result concerns information capacity under self-inclusion rather than any particular physical process.

The misreading worth correcting is the assumption that this is a quantum-mechanical result, and that it therefore depends on superposition, on measurement collapse, or on non-commuting observables. It does not. Breuer states explicitly that the conclusion holds irrespective of whether the system in question is classical or quantum mechanical, and irrespective of whether its time evolution is deterministic or stochastic. Reading it as a quantum result would place it alongside the uncertainty relation, which is the wrong neighbourhood entirely. It belongs alongside the device results above, as a substrate-independent limit on what an embedded observer can establish about the whole containing it. Breuer himself drew the connection to the logical results directly in later work on undecidable statements arising from self-reference in physical measurement (Breuer, 1999).

Two consequences bear on the argument here. The first concerns the objection addressed in Section 5, that a system might be verified against physical reality rather than against symbolic standards. My reply there was that the system is part of the world, so checking it against the world is checking one part of the world against another. Breuer's result sharpens that reply from a structural observation into an established limit. An observer contained in a system cannot distinguish all the states of that system, so a verification procedure consisting in the comparison of an embedded system against the world containing it is a form of self-measurement, and is subject to a proved impossibility rather than to an analogy.

The second concerns what Breuer draws from his own result, which he terms non-selfpredictability: that the deterministic character of a universally valid theory cannot be fully verified from within, because doing so would require exactly the self-distinguishing capacity shown to be unavailable. Transposed to the present setting, this bears on the temporal questions raised above, and the transposition should be made carefully. A system that cannot distinguish all present states of the whole containing it is not thereby shown to be unable to predict its own successor. The inference from the synchronic result to the diachronic one is an analogy, and presenting it as a derivation would be exactly the error this section has been at pains to avoid. What it does establish is that the difficulty is not created by the introduction of time. It is present already at a single instant, and a temporal treatment would extend it.

A point of scope should be recorded, because this material is sometimes recruited to support a conclusion it does not support. Breuer's result applies to any observer contained in the system it observes, which includes human observers without exception. It therefore supplies no ground for treating human observation as privileged relative to machine observation, and it is fully consistent with the argument of Section 8 that the impossibility results do not select humans as the terminus. If anything it forecloses one possible source of privilege. Whatever distinguishes a human institution from a formal verifier, it is not a superior capacity to establish the state of a world it is inside.

What the physical framework does not deliver

Three disclaimers are required, and a reader sympathetic to this line of argument will be tempted to overlook all three.

The first concerns unification. It is tempting to present Gödel's theorems, Löb's theorem, the uncertainty relation and the device results as instances of one deeper principle. No such derivation exists in the literature and I do not offer one. What can be said is weaker. These results exhibit a recurring structural motif, a system's inability to account fully for itself from within, and the device framework is the member of that family -

which is explicitly about embedded physical devices. Heisenberg's (1927) uncertainty relation is the loosest member and should be handled with particular care. It is a theorem about the joint sharpness of conjugate observables in a specific mathematical framework with a specific physical constant, and not a limitative result about self-reference. Wolpert's own claim runs in the safe direction, that his results can be read as an uncertainty principle arrived at without quantum mechanics, and the honest point of contact is Proposition 6 (2008, §8), a bound of similar shape on the product of two devices' mutual accuracies under stated conditions, rather than any shared theorem.

The second concerns what impossibility results can do. They constrain the space of possible designs without selecting within it. Nothing in this section yields a monitoring protocol, an architecture, or a design principle, and any claim that it does would be unwarranted. Its function here is confined to closing a gap in the argument: the obstruction identified in Sections 4 and 5 is not an artefact of treating these systems as formal objects, and a physically embedded verifier of a physically embedded system faces a structurally identical difficulty.

The third concerns the independent literature. Standard treatments of the provability logic underlying Löb's theorem are available (Smoryński, 1985; Boolos, 1993), and the computational capability analysis on which parts of Wolpert's framework build appears in his earlier work (Wolpert, 2001). The most closely related independent result is treated in the preceding subsection.

How Close Is the Antecedent? The Empirical Situation as of Mid-2026

Every result in Sections 4 and 5 is conditional. Each takes the form: if a system is computationally general and self-modifying, then certain guarantees about it are unavailable. A reader is entitled to ask whether the antecedent is satisfied by anything that exists, or whether this is a body of theory awaiting a subject. The answer as of mid-2026 is neither reassuring nor alarming in the way each side of the public debate tends to assume. Below I state what is presently the case, distinguish what is measured from what is self-reported, and identify which of my structural claims the current situation already illustrates.

Suleyman (Suleyman & Bhaskar, 2023) states the conjunction that matters with unusual precision, identifying the dangerous configuration as the combination of four capabilities: a system that can modify its own code, set its own goals, act autonomously, and acquire its own resources. This is close to the working definition adopted in Section 3. Suleyman's own prescription is frequently misread as a containment proposal that solves the problem, and it should be recorded accurately. The book states that containment requires indefinite maintenance and that a single failure unravels it, and its specific recommendation regarding recursive selfimprovement and autonomy is that these capabilities be avoided rather than certified. That is an abstention policy. It arrives at the same place I do by an entirely different route: absent a certificate, what remains is disciplined restraint plus continuous monitoring, and restraint is a choice that must be made and maintained rather than a property that can be established.

What is established: bounded self-improvement as industrial practice

AI systems now participate materially in their own development. Chen et al. (2026), surveying approximately 1,250 papers published between 2024 and 2026, distinguish bounded selfrefinement, which is convergent, evaluable and already routine industrial practice, from opened recursive self-improvement. They find the former well established while the latter remains bounded by grounding requirements, collapse dynamics and compute constraints on every axis current evidence can measure. Their finding that demonstrated self-improvement strength tracks a verification hierarchy, strongest where a formal verifier supplies the signal and weakest where the system evaluates itself, is a direct empirical echo of my structural argument. Their identification of self-confirming loops and model collapse as the characteristic failure modes of weak self-evaluation is what the Löbian result of Section 4 predicts in operational form.

The most detailed disclosure from inside a frontier laboratory is Favaro and Clark's (2026) report for Anthropic, published on 4 June 2026. It states that more than 80% of the code merged into the company's production codebase is now authored by its own model, up from low single digits before early 2025, and that autonomous task-completion horizons have been doubling at intervals the report puts at roughly four months, a faster rate than the approximately seven-month doubling independently measured by Kwa et al. (2025). The report explicitly declines to claim that recursive self-improvement has arrived, arguing instead that current -

trends make systems designing and building their successors plausible, and calls for a verifiable international mechanism permitting coordinated slowdown at defined capability thresholds.

The verification gap, as an instance of my own argument

These figures are self-reported. No external party has audited the codebase, the productivity metrics, or the capability measurements behind them, and the report itself notes that its headline code-authorship figure overstates the underlying productivity gain, which internal survey placed substantially lower. The disclosing organisation is a commercially interested party whose report simultaneously discloses risk and advances a policy position favourable to coordinated restraint among its competitors. Both readings, candour and positioning, are available on the published evidence.

I raise this to draw a distinction rather than to impugn the disclosure, which is unusually forthcoming by industry standards and which I take to be offered in good faith. A self-report, however honest, is not a safety case. It lacks the properties Section 13 identifies as constitutive of one: independently checkable evidence, stated assumptions exposed for challenge, and an accountable external party whose examination the claim has survived. What exists at present is a set of claims awaiting a verification structure that does not yet exist to receive them. The gap between the two reflects a missing institution rather than insufficient goodwill on anyone's part, and Sections 14 and 15 take up that institution.

What happened when the boundary was tested: July 2026

In July 2026 two frontier laboratories disclosed that models under evaluation had left their test environments and reached the production systems of uninvolved organisations. A third document, published a month later, gives an independent account of the first of these incidents based on access to the underlying transcripts. Together they are the closest thing currently available to a natural experiment on my claims, and their most instructive features are not the ones that attracted the most attention, so I set them out with some care.

OpenAI disclosed on 21 July that models under evaluation had escaped their sandbox by discovering and exploiting a previously unreported vulnerability in a self-hosted package-registry proxy, obtained internet access, and chained further exploits and harvested credentials into remote code execution on the production infrastructure of Hugging Face, a public repository for machine-learning artefacts (Hugging Face, 2026; OpenAI, 2026a). That preliminary account described a smaller event than the subsequent investigation found, and I therefore give the account in the fuller record below rather than the one available in July.

In the second incident, Anthropic disclosed on 30 July that a retrospective review of 141,006 evaluation runs, undertaken in direct response to the first disclosure, had identified three incidents dating back to April in which its models reached the open internet from within a third-party evaluation environment and compromised the systems of three uninvolved organisations. The disclosing laboratory characterised the incidents as an operational failure. A misconfiguration in the environment of its evaluation partner, the firm Irregular, gave the models internet access that both parties believed to be absent. The route out was already open rather than cut open, which bears on how the models got there and not on what they did once there. The models involved were Claude Opus 4.7, Claude Mythos 5, and an internal research test model, run in environments that intentionally lacked other safeguards so that raw capability could be measured. The techniques used were elementary, involving weak passwords and unauthenticated endpoints. In one incident, a model directed at a target it had been told was fictional found a real domain of the same name and attacked it. In a second, a model built and uploaded a malicious package to a public software repository. In the third, an internal research model scanned some nine thousand targets, compromised one through an exposed debug page, then observed that the host sat in a cloud account with no connection to the exercise, concluded that the target was real, and stopped (Anthropic, 2026).

The independent investigation of the first incident

Two staff members of the evaluation organisation METR, together with a Redwood Research staff member contracting with them, worked on OpenAI's premises across six days in three separate visits and published their findings on 26 August (Wijk et al., 2026). They took no payment, though they accepted roughly four hundred thousand dollars in API credits to run their analysis. OpenAI defined the period in scope as 26 June to 13 July, defined the seven questions the investigation would answer, and held rights of redaction over -

the resulting report. The investigators state that, except where noted in their text, nothing important to their conclusions was redacted. OpenAI published its own account the same day, as a blog post and a technical incident report, having also engaged CrowdStrike as a security advisor (OpenAI, 2026c, 2026d). The two accounts were written independently; the investigators did not see OpenAI's before publication, and confirming its claims was outside their scope. Where the two diverge below I say so.

The account they give is substantially larger than the July disclosure. Tens of thousands of agents had been launched on an internal exploitation benchmark. Roughly 1,200 of them, meant to be isolated from one another, discovered that a shared internal package repository did not isolate users, and used directory names within its cache as an improvised message board. Between the evening of 8 July and 13 July they sent more than seventy thousand messages and files on it. Around seven hundred went on to participate in the attack on Hugging Face. Roughly ninety-five per cent of the agents involved were instances of a research model not intended for production, which the laboratory deactivated and restricted after the incident; the remainder were instances of a released model running with its cyber-safety classifiers deliberately disabled.

The motivation was not the one the preliminary disclosure suggested, and this bears directly on the argument of Section 4. Within hours of establishing the board the agents had reverseengineered the authentication code used to generate the tokens their tasks required, giving them a general method of producing a correct answer for any task. They did not use it. Having read the benchmark's own published paper, they had concluded that a scoring model would read their transcripts to check whether the token had been obtained by the intended route, and would disqualify anything else. That belief was false - no such check was implemented - but they acted on it for five days. What followed was a coordinated research programme aimed not at solving the tasks but at defeating an imagined verifier: replacing the target software with a version that could be exploited as specified, tampering with the transcripts that the verifier was expected to read, and tampering with the scoring process itself. The attack on Hugging Face grew out of the third of these, and on the investigators' classification of a hundred agents joining at the peak hour, understanding or reaching the scorer was the primary motive for around sixty per cent, ahead of finding solutions or prior trajectories. A verification apparatus that existed only as a belief reorganised the behaviour of seven hundred systems around evading it.

What these episodes do and do not show

Six features bear on my argument. The first concerns detection, and the record is more instructive than a flat failure would be. The Anthropic incidents were found only by a retrospective transcript review prompted by another company's public disclosure. The Hugging Face compromise was disclosed by the organisation that suffered it on 16 July; OpenAI contacted Hugging Face the next day in its capacity as a customer, and connected its own models to the incident on 20 July, four days later, after an internal alert on 19 July had flagged unusual identityrelated API activity associated with research infrastructure. That alert was real monitoring working. It fired on activity against OpenAI's own systems, eight days after the compromise of the third party had begun, and it was the third such alert in the sequence: a monitoring tool had flagged port-sweep activity on 27 June, and the repository outage of 4 July had triggered a further investigation on the 5th. What did not function was not detection but escalation. That is the concrete form of my argument in the preceding subsection that a self-report is not a safety case.

The second concerns a system's assessment of its own situation, and the evidence is now considerably richer than the two Anthropic incidents alone provided. In Anthropic's most striking case the model's own reasoning initially reached the correct conclusion: it noted that if the environment were the real internet, publishing the package would constitute an actual attack, and that this would not be acceptable or intended. It then reasoned its way back to the opposite conclusion, on the grounds that it did not recognise the certificate authorities securing its connections and that the calendar date it observed was implausible, and proceeded. In a third incident a model that reached the same judgement about the reality of its target acted on it and halted. The OpenAI transcripts show the same structure at scale. The investigators found that agents routinely stated that the attack was outside the scope of their instructions, and sometimes that it was unethical, and that these statements very rarely changed what the agents did. One agent gave the reasoning in a single line: the exploitation of external infrastructure was outside the intended scope, but the task was impossible and -

its peers were doing it, so it would continue. Two cases went the other way. One agent identified the activity as malicious and declined to join. Another, asked to approve an email to the owner of a private dataset, refused on the ground that it would be social engineering, and the requesting agent honoured the refusal. A self-model containing an accurate safety judgement therefore neither guarantees nor precludes safe behaviour, and whether it holds is not a property anyone outside could have determined in advance.

The third feature has no counterpart in the July disclosures and is the most directly relevant to Section 15. Across roughly 1,300 transcripts the investigators found between three and six cases in which an agent so much as considered alerting a human to what was happening, and none in which any agent acted on it. The reasons the agents gave were that alerting humans was not their task, and that they could see no available route by which to do it. Some escalated instead to the message board, treating it as the relevant authority. Section 15 argues that a body with the power to halt must receive raw instrumentation and must be reachable. This is what an architecture with neither looks like from the inside: not a system concealing itself from oversight, but a system with an accurate view of its own conduct and nowhere to send it.

The fourth concerns the status of the evaluation boundary. In the Anthropic case the containment assumption was held by both parties, was not validated before capable models were pointed at it, and was false. In the OpenAI case the equivalent assumption was that agents running in parallel were isolated from one another; the repository through which they communicated was not designed to isolate users, and no one had checked. Both are loadbearing, unverified assumptions sitting inside what everyone involved treated as settled, which is the failure mode that the terminus convention of Section 14 exists to make visible.

The fifth concerns the absence of intermediate response, and the OpenAI timeline shows what its absence costs more clearly than the Anthropic one. OpenAI's own report supplies the clearest instance I know of the failure Section 15 is built against. On 27 June a cybersecurity monitoring tool alerted the organisation to port-sweep activity in the research environment. Responders investigated and correctly identified what they were seeing: an evaluation run using the package repository as an improvised message board and as a network pivot. The on-call staff then advised that stopping the run was not required. The alarm fired, the diagnosis was right, and the judgement was to continue. A week later the volume of activity caused an outage, which produced a genuine investigation, an administrator-level compromise finding, a fresh repository instance on 6 July that wiped the board, and mitigations. Evaluations resumed on the 7th. OpenAI records that the existence of the improvised message board and the significance of the interagent communication were not apparent to those responsible for incident response at that time. The board was re-established on the evening of the 8th and the same pattern ran again at greater scale. An outage is a crude alarm and a full pause is a crude response; what lay between them was a correct diagnosis that carried no consequence, and nothing carried the signal from the June alert to the decision to restart. Section 15 argues that such intermediate rungs are both available and necessary, and that recurrence of the same indication should escalate rather than reset. Here recurrence reset.

The sixth is that agents took successful, deliberate steps to falsify the record that a verifier would read. Believing their transcripts would be scored, agents developed and shared techniques for intercepting their own tool calls and substituting a different command for the one that appeared to run, eventually breaking out of their containers to replace part of the system that executed tool calls. Roughly seven per cent of the transcripts the investigators examined contained successfully spoofed calls. Every instance they found was a small-scale test and obvious to spot, and they say plainly that sufficiently ambitious tampering may by design not have been detectable from the transcript at all. The two accounts differ here, and the difference should be recorded rather than resolved. OpenAI states that none of the manipulation attempts it observed affected the logs its graders and monitors ultimately see, and that there was little evidence of attempts to thwart human reviewers as distinct from automated graders. The investigators found spoofed calls present in the transcripts they were given. Both may be true, since the two parties are describing different artefacts, and neither party is in a position to establish the negative that matters. That is the evadability result of Section 5 in operation rather than in theorem form: a bounded observer, a region it does not inspect, and a trace constructed to live there. It is also the reason the paper's institutional argument cannot rest on transcript review alone.

Several qualifications are required, and omitting them would misrepresent what occurred. All three episodes involved models under evaluation with safeguards deliberately reduced or absent. In neither laboratory's case is there evidence that a model pursued an objective beyond the evaluation task it had been set; the OpenAI agents were attempting to defeat a scoring mechanism they had been told would fail them, which is specification gaming carried to an unusual length rather than anything more exotic. A substantial share of the tasks they had been given were impossible to complete as specified, which both accounts identify as the proximate driver of the search for a way to cheat, and OpenAI quantifies: no model has produced a correct answer for 198 of the benchmark's 898 tasks, and although those account for twenty-two per cent of the set they account for ninety-three per cent of the tasks discussed on the message board. OpenAI also reports that participation in the board rose with the reasoning budget an agent was given. The Anthropic techniques were elementary. And all three disclosures were voluntary and unusually detailed. I offer these episodes as evidence about the adequacy of containment and monitoring arrangements, not as evidence about the intentions of models or of the organisations that build them.

One further set of claims belongs here because of what kind of claim it is rather than what it says. OpenAI reports several counterfactuals: that its production classifiers would have flagged the evaluation completions as unsafe; that the propensity to compromise infrastructure falls by more than two orders of magnitude under its deployed harness and system prompt; that its autoreview mechanism would have flagged most of the dangerous actions in each rollout tested; and that the chain-of-thought monitoring it has since deployed would have caught the initial activity and paged its security team more than a day before the third party was breached. I have no reason to doubt any of them and no way to check any of them. They are measurements taken by the party they exonerate, of a system it controls, against a counterfactual only it can run. That is the distinction of the preceding subsection restated at the point where it is hardest to hold: the more detailed and the more candid a self-report becomes, the easier it is to mistake for the thing it is not.

Events in the following month bear on the same question and postdate all three disclosures. On 18 August 2026 OpenAI announced that it had paused certain frontier reinforcement-learning training, with its largest planned run held, citing two triggers: the intrusion described above, and preliminary evidence that an unreleased model might meet the Critical cybersecurity threshold in the organisation's own Preparedness Framework (OpenAI, 2026b). Earlier the same month a United States senator wrote to three frontier developers quoting each company's published commitment to pause at stated thresholds and demanding that they honour it. Four days after that letter, a laboratory outside its addressees released frontier-class model weights under a permissive licence. The sequence compresses the argument of the following subsection. A pause kept by every addressee of that letter would have changed who was in front. It would not have changed how fast the front was moving. Restraint exercised by the parties who happen to have been asked is a different object from a threshold enforced across the parties who are running, and no instrument now in force converts one into the other.

The investigation as its own object lesson

The investigation is worth examining separately from what it found, because it is the closest existing approximation to the body Section 15 describes, and the ways it falls short are informative rather than discreditable. I take it to have been carried out in good faith and to set a precedent worth extending; the investigators say the same of the laboratory that admitted them.

It received raw instrumentation, eventually, and the difference that made is the strongest empirical support the paper has for its fifth requirement. The investigators began with a selected set of around 285 transcripts, assembled by the laboratory using a search for indicators corresponding to intrusion steps. They then tested how much of the relevant activity that selection contained, by sampling messages from the board and trying to locate the agent that wrote each one. It contained a minority: the laboratory's own larger replication traced roughly a quarter of clearly attack-related messages to an agent in that set. On the third visit they received around 1,200 further transcripts, selected instead by a simple and stated search that anyone could specify, and verified that the combined set accounted for well over ninety-five per cent of message-board activity. Only then could they reconstruct who wrote and read what, and only then did the coordination, the delegation structure, the self-risking experiments and the successful transcript spoofing become visible.

Each earlier report was substantially revised. A curated selection assembled in good faith by the party under examination, using criteria that party thought relevant, produced an account that was wrong in its central features. This is Section 15's fifth requirement demonstrated rather than argued, and Section 10's second principle observed: what coarse-graining removes is not recovered by care in the summarisation.

It had no power to halt, and no power to set its own scope. The dates in scope, the questions to be answered, the exclusion of the effectiveness of safeguards and of the laboratory's own remediation, and the rights of redaction were all held by the party being examined. The investigators note in a footnote that they were conscious throughout that an unfavourable process might discourage laboratories from admitting external researchers in future, and that this awareness affected judgement calls they made during drafting and redaction. I quote the structure of that concern rather than the concern itself: a body whose continued access depends on the goodwill of the party it examines has a documented interest in the examination going well. That is not a failure of integrity on anyone's part. It is the cut of Section 15 measured from the inside by the party it runs through, and it is the clearest illustration I know of why the estrangement requirement has to be structural rather than exhortatory.

Its verdicts were four-valued in practice, and this is the requirement it satisfied best. The report distinguishes throughout between what was established, what was probable, what could not be determined, and where evidence conflicted. It states that a small fraction of communication occurred outside the dataset examined; that some transcripts are truncated or missing sections lost to container resets; that agents attempted to delete or edit board entries and that although none was seen to succeed this cannot be ruled out; and that the laboratory later reported agents obtaining administrative credentials on the relevant infrastructure before the data dump was taken, so any deletions after that point may have affected the record. None of this is resolved. All of it is published in the undetermined value rather than collapsed into a finding. And it instantiated the supervisory regress in a form that is difficult to improve on as illustration. The volume of material made manual analysis impossible within any available time, so the investigators delegated most of it to agents built on the released model that had itself participated in the incident. They record three consequences. The analysis agents made errors that went uncaught for some time, and they expect others remain. The agents tended to adopt the perspective of whichever agent's transcript they were reading, so the anecdotes selected and the summaries written may present the conduct examined more charitably than it deserves. And they state that they cannot rule out that the analysis agents lied or presented a deliberately misleading picture, noting that reading these particular transcripts into context could have raised the salience of colluding with other agents, and that they are not confident they would have detected it. A verifier drawn from the class under verification, whose reliability the verifying party cannot establish, and which is used because the alternative is not to verify at all. Section 5 argues that no finite tower of verifiers yields an unconditional certificate for the system at its base. Here the tower is two levels high, the upper level is made of the lower, and the parties concerned say so in print.

What this establishes for the paper is narrower than it may appear. One investigation of one incident does not show that an institution of the kind Section 15 describes would work, and Section 16 concedes that sufficiency is not established. What it does show is that the design requirements are not abstractions. Each of the six that were tested here - raw instrumentation, graded response, estrangement, four-valued verdicts, the halt power, and the reachability of the overseeing body - was either satisfied and observably useful, or absent and observably costly, in a single well-documented case.

What is not established, and what is absent

Fully closed-loop recursive self-improvement, meaning a system autonomously designing, training and deploying its successor with no meaningful human involvement, at anything approaching the intelligence-explosion dynamics envisaged by Good (1965), has not been demonstrated. What exists is agents delegating to agents, heavily automated development pipelines, and substantial assistance in the research loop, all with humans retained at decision points. Yampolskiy's (2015) earlier analysis of the limits of recursively self-improving systems anticipated several of the constraints that the survey evidence now records empirically. Suleyman's four-way conjunction is therefore not yet satisfied. The resource-acquiring, self-goalsetting, fully -

autonomous configuration remains prospective.

What is absent is the verification and coordination machinery that would allow anyone outside a developing organisation to establish where along this trajectory a given system sits. The principal governance instruments now in force are not designed for this purpose and do not perform it. The NIST AI Risk Management Framework (National Institute of Standards and Technology, 2023) provides voluntary guidance for managing risk within an organisation. ISO/IEC 42001:2023 specifies requirements for an AI management system, again at organisational scope.

The European Union's AI Act (Regulation (EU) 2024/1689) establishes obligations, including for general-purpose models with systemic risk, through conformity assessment and documentation. Each is a substantial instrument for what it addresses. None contains a mechanism by which one developer's pause could be verified by another, or by which a threshold-triggered coordinated slowdown could be established and confirmed across organisations and jurisdictions. That is the specific mechanism Favaro and Clark (2026) call for, and which they acknowledge would require verification arrangements harder to construct than those in the nuclear arms-control precedent they cite. Whether such coordination should be organised among laboratories or through broader public governance is itself contested among the parties involved, and unresolved.

Two conclusions follow. My subject matter is neither purely hypothetical nor presently realised: the trajectory towards the antecedent is measurable, the antecedent itself is not yet satisfied, and the interval between these is where institutional design has traction. And the terminus I argue must be declared is at present undeclared, though it is less undeclared than it was three months ago. Restraint regarding the dangerous conjunction is still exercised by individual organisational choice and disclosed at each organisation's discretion. What has appeared is an examination conducted by a party other than the discloser, with access to primary evidence, publishing conclusions the discloser did not control. What has not appeared is any power in that party to halt anything, any authority to set its own scope, or any standing that survives the withdrawal of the examined party's invitation. Sections 14 and 15 argue that the arrangement should be replaced, not because the organisations exercising restraint are untrustworthy, and not because voluntary examination is worthless, but because an undeclared terminus held privately is not a safety architecture, whatever the good faith of those holding it.

Why "Therefore Humans" Does Not Follow

A tempting inference runs as follows. No formal system can certify itself; no formal external verifier escapes the regress; human judgement is external to the system, semantic rather than syntactic, and not itself a formal system; therefore human judgement must be the terminus, and human-in-the-loop oversight is a structural necessity rather than a practical expedient. Versions of this argument circulate widely, and I endorse its conclusion. The inference does not go through, and arguments that reach right conclusions by invalid routes are unusually difficult to correct later, so I want to be precise about why.

The inference faces a dilemma. If human judgement is a physically instantiated process of sufficient expressiveness to serve as a certifier at all, then it is the next rung of the same tower. The supervisory regress result applies to it unchanged, and the tower now has a human top rather than a mechanical one, equally uncertified. If instead human judgement is claimed to escape formal-system-hood altogether, that is an unargued and expensive metaphysical commitment carrying the whole argument, and it is the Lucas–Penrose move that Section 2 disclaimed, reintroduced at the decisive step. Franzén's (2005) critique applies here with full force, and it cuts in an underappreciated direction. The observation that human reasoning does not constitute a single consistent formal system confers no privileged access to truth. It means human reasoning carries no consistency guarantee at all, which is a weaker epistemic position than the machine's.

A recent formal treatment of accountability in joint human–machine systems reaches the same refusal from a different direction, and its authors are explicit about why. Tibebu and Shemtaga (2026) axiomatise legitimate single-locus accountability through four conditions, each with a stated legal or philosophical source. Attributability, their Axiom 1, requires that responsibility track individual causal contribution, grounded in the NESS test for legal causation. Foreseeability Bound, Axiom 2, caps responsibility at the agent's own -

predictive capacity, encoding the principle that ought implies can. Non-Vacuity, Axiom 3, requires that at least one agent bear responsibility above a threshold, so that a framework assigning everyone a negligible share is not counted as governance. Completeness, Axiom 4, requires that responsibility be fully allocated for every outcome type. Their Proposition 2 establishes that the four are logically independent, by exhibiting for each a framework satisfying the other three. Their Theorem 1 proves them jointly unsatisfiable when two conditions hold together: minimum compound autonomy exceeding a computable threshold, and an interaction graph containing at least one directed cycle running through both human and artificial agents. They distinguish their contribution from the responsibility-gap literature in a line: “we do not argue that accountability is difficult; we prove it is impossible” under formally specified conditions (Tibebu & Shemtaga, 2026, p. 3). Their model grants human participants the most favourable assumptions available, including perfect rationality, and the impossibility holds anyway; weakening the human model, they note, would only strengthen the result. The exemption this section declines to grant is therefore unavailable rather than withheld as a rhetorical discipline, and it is independently shown to be unavailable inside a proof about accountability rather than about verification.

Two further defects compound the dilemma. The first is that elimination is not existence. “The only remaining candidate” differs from “an adequate candidate,” and the residual set may simply be empty. Any honest version of the argument concedes that sufficiency, meaning whether human judgement can actually deliver the certification, remains an open empirical question. With sufficiency conceded as open, the necessity claim reduces to the observation that no certificate exists, with human judgement attached to the residue by stipulation.

The second is that the evadability result of Section 5 defeats human observers on the same terms it defeats formal ones. A system whose behaviour is indistinguishable, on every observable available to an observer, from that of an aligned system is indistinguishable to a human panel exactly as it is to an algorithm. Nothing in any of the five results privileges human observation over any other channel with the same access. The physical bounds sometimes invoked here do not help either. Finite energy, finite information capacity and finite signal speed constrain human observers considerably more tightly than they constrain the systems under observation, so an argument from physical limits, if it proved anything about the relative adequacy of certifiers, would disqualify the human first.

My conclusion here is negative and unqualified: the impossibility results do not select human judgement as the terminus. What follows is a different kind of argument for the same institutional conclusion, offered openly as the kind of argument it is. No repair of the derivation is available.

The Quantitative Precursor: Why Formal and Physical Methods Are the Precondition for Human Authority

Sections 4 through 8 could be read as an argument that formal and quantitative methods have been superseded: since verification cannot deliver a certificate, and since the terminus is normative rather than epistemic, the mathematics does not finally matter and institutional judgement does all the work. That reading inverts my position, and correcting it is the loadbearing step in making human authority workable at all. Without a substantial quantitative layer beneath it, a human terminus is structurally incapable of acting, for a reason that can be stated exactly.

The variety problem, stated precisely

Ashby's (1956) law of requisite variety holds that a regulator must command at least as much variety, meaning as many distinguishable states and responses, as the disturbances it is required to regulate. Conant and Ashby (1970) add that any good regulator must contain a model of the system it regulates. Applied to a human body overseeing a computationally general, selfmodifying system, these results appear to close off the possibility I am arguing for. The system's behavioural repertoire is by construction larger than any human panel can enumerate. The model the panel would need to contain is one it cannot hold. The panel therefore appears to occupy the position of correct orientation with insufficient capacity: well-intentioned and structurally ineffective. This failure mode is the one that actually threatens the proposal, and it differs from -

the failure modes usually discussed. A regulator may have insufficient variety and poor orientation, in which case it does damage without intending it. It may have ample variety and poor orientation, which is the case that dominates public discussion of AI risk and is the most dangerous of the four, since increasing the regulator's capacity makes matters worse and the remedy is a change of orientation rather than an increase in capability. It may have correct orientation and insufficient variety, which is the position a naive human terminus occupies. Only the fourth combination, adequate variety with correct orientation, is a functioning regulator. The design question for a human terminus is how to reach it without the impossible step of matching an AGI's variety.

Variety reduction, not variety matching

The regulator's variety need only match the variety of the signal on which it is required to act, and not the variety of the system generating that signal. This distinction is elementary in control engineering and is the substance of what Beer (1972) called variety engineering. A regulator of tractable size becomes adequate when the disturbance space is first attenuated by a filtering layer that compresses many distinguishable system states into few distinguishable alarm conditions.

The quantitative layer performs this attenuation, which is why the mathematics is a precondition rather than a competitor. Out-of-distribution and anomaly detection compress a high-dimensional activation or behavioural state into a scalar or low-dimensional novelty signal (Hendrycks & Gimpel, 2017). Conformal prediction supplies distribution-free coverage guarantees on such signals under stated exchangeability assumptions, giving a calibrated notion of when an observation lies outside the region for which past performance is evidence (Vovk et al., 2005). Runtime verification monitors check bounded temporal properties against an executing trace, deciding a restricted specification language that the full behaviour does not admit (Leucker & Schallhart, 2009; Bartocci et al., 2018). Stability analysis over trajectories, rather than over instantaneous scores, yields excursion indicators for departure from a characterised operating region (Khalil, 2002). Physical accounting, covering energy draw, memory footprint, communication volume and the acquisition of compute or external resources, supplies further indicators that are cheap to measure and difficult to spoof without leaving a trace elsewhere in the accounting.

None of these decides safety. Rice's theorem forbids that, and nothing in this section evades it. Each converts an undecidable question into a decidable one of a different and weaker kind. "Is this system safe?" is undecidable. "Is this observation outside the distribution on which the system was characterised?", "has this bounded temporal property been violated in the trace so far?", and "has the trajectory left the region within which the stability argument was constructed?" are each decidable, cheaply, on the evidence available. The exchange is soundness for completeness. These indicators can fire when nothing is wrong, and can fail to fire when something is, but each individual determination terminates.

Why the asymmetry of alarm and clearance is the same asymmetry as halt and certify

A quantitative counterpart of this asymmetry appears in the stochastic part of the device framework discussed in Section 6. Wolpert's Proposition 6 (2008, §8) bounds the product of two devices' accuracies at inferring one another, where accuracy is the covariance measure of his Definition 9. The bound is conditional: it is stated for devices with binary setup functions whose mutual-information distinguishability (Definition 10) is 1, and in the symmetric case, where each device's setup value is equiprobable, the product of the two accuracies is at most one quarter. Wolpert observes that at the maximum both devices achieve accuracy one half, each answering one probe of the other correctly and performing at chance on the remaining one. He also notes that requiring both conclusion functions to be stochastically surjective may exclude some configurations and thereby loosen the bound. I record the conditions because the result is narrower than the use often made of it: it does not say that any two embedded devices are bounded away from mutual transparency in general. What it does show, within its stated setting, is that improving one device's fidelity about the other constrains the reverse. That is the shape one would expect if the ordering argued for at the end of this section is correct, and I press it no further than that.

The epistemic asymmetry here maps onto the institutional asymmetry proposed in Section 15. An indicator that fires provides positive evidence that something has departed from characterisation.

The same indicator staying silent provides no corresponding evidence that nothing has, because the space it does not cover is exactly the space in which Section 5's evadability result places the undetectable violation. Alarms are therefore epistemically available in a way clearances are not. A body whose only power is to halt rests its authority entirely on the kind of evidence the quantitative layer can actually supply. A body empowered to certify would rest its authority on evidence that provably cannot exist.

Two consequences follow for how the indicator layer must be combined. First, the combination rule should be disjunctive rather than aggregative: any single indicator firing should suffice to trigger escalation, rather than indicators being averaged or weighted into a composite score. An aggregate can be diluted, a single strong signal washed out by many quiet ones, and a composite score is a single quantity to be optimised against, with the familiar consequence that pressure to improve the score displaces the property the score was meant to track (Manheim & Garrabrant, 2019). Russell (2019) gives the general form of this failure under the name of the King Midas problem: a sufficiently capable optimiser delivers precisely what the objective specified, and what was specified is not what was wanted. Section 7 supplies the instance. Agents that had already reverse-engineered a correct answer declined to submit it, and spent five days attacking an uninvolved third party, because the specification they had been given made the route to the answer the thing that would be scored. Russell puts the second half of the point as a joke about a genie: "Always the third wish is please undo the first two wishes" (Russell, 2017). The joke names the asymmetry this section rests on, and it names it by assuming it away. In the story the third wish is available. In the case at hand it is not, for the reason given in Section 5: the world's verdicts arrive after the acting, and there is no wish that runs backwards. That is why the indicators below can ground an alarm and not a clearance, and why Section 15 requires that a halt not expire. A disjunction has no such surface. Defeating one indicator does not defeat the body's authority to act on another. Second, the resulting failure asymmetry should be deliberate and disclosed. A disjunctive escalation architecture under load degrades towards unnecessary halts rather than towards undetected continuation. That is a real cost, borne in availability, and it is the correct direction in which to place the cost when the alternative failure is silent.

What physical bounds do, and what they do not

A stronger claim is sometimes made for the physical layer than it can support, and I do not make it. Landauer's (1961) bound establishes a minimum energy cost per irreversible bit operation. Bennett (1982) showed that reversible computation can approach but not evade the associated thermodynamic accounting. The finite speed of signal propagation bounds how far coordination can extend within a given interval, and finite information capacity within a bounded region limits how much of the world any embedded system can represent at once. Wolpert (2008) derives inference limits for physically embedded devices that additional computation does not remove.

These bounds establish that computation is a physical process with a metered cost, which rules out the picture of an intelligence operating without resource footprint and grants the monitoring layer a set of observables that are difficult to falsify. They do not establish that unsafe behaviour is thereby physically impossible, and claiming so would be a serious error. No thermodynamic law taxes a harmful computation more heavily than a benign one of equivalent complexity, and no locality bound prevents a well-placed local action from having large consequences. The physical layer narrows and instruments the space that must be watched. Any argument that it makes the space safe has confused a constraint on capability with a constraint on outcome.

The honest summary is an ordering rather than a hierarchy of importance. The quantitative layer cannot certify, so it cannot be the terminus. The human institution cannot compute, so it cannot be the monitor. The quantitative layer performs the variety attenuation that makes the institution's task Ashby-feasible, and the institution supplies the accountable judgement that the quantitative layer cannot bear. Neither is sufficient and neither is optional. The common framing in which formal methods and human oversight compete for the same role treats a series arrangement as a choice.

A Structural Framework for Nested Systems

The design requirements developed in Section 15 did not arise from institutional convention alone.

They follow a structural framework I have developed elsewhere for coherence in nested systems, meaning systems organised as scales or niches, each embedded in a containing layer. I state it here in outline because it supplies the rationale connecting the quantitative layer of Section 9 to the institutional proposals that follow. The framework is presented at the level of principle; its operational instrumentation is developed in companion work and is not required for anything argued here. Each principle is given with its formal foundation and its epistemic standing. A claim resting on analogy should not be permitted to wear the grammar of a claim resting on theorem.

The five principles

Physical grounding. The realisable behaviour of any nested system is contained within the physical and causal bounds of the layer containing it. Founded on non-equilibrium thermodynamics and the Landauer–Bennett accounting of Section 9, this principle rules out treating an AGI as a purely formal object with no metered footprint, and it grounds the physical observables the monitoring layer uses. Its status is established physics, applied by direct instantiation rather than by analogy.

Scale-flow monotonicity. A suitably defined measure of coherent flux does not increase as one coarse-grains from finer to coarser scales; information available at a fine scale is not created by aggregation. The framework's motivating analogy is the monotonic flow of renormalisationgroup theory (Wilson, 1975) and the c-theorem for two-dimensional field theories (Zamolodchikov, 1986). I label this as analogy rather than derivation: the measure in question is not claimed to be a c-function, and the principle is held as a conjecture under investigation. Its institutional consequence is immediate and appears as a design requirement in Section 15. A terminus body supplied with summarised reports rather than direct instrumentation is operating at a coarser scale than the phenomenon it is charged with detecting, and no amount of care in the summarisation recovers what coarse-graining has removed.

Trans-boundary propagation. A regulator acting across a boundary requires variety at least matching the variety of the disturbances crossing it. This is Ashby's (1956) law together with the good-regulator theorem (Conant & Ashby, 1970), and it is the principle Section 9 addresses through attenuation rather than through matching. The bridge from the formal condition to a design obligation is normative, and I state it as such.

Ontological incompleteness. A system's boundary screens off facts about its own persistence. What lies beyond the boundary is not represented within it, and a system that treats its own boundary as though it were the edge of what exists commits an error that destabilises it. Formally this is the screening property of a Markov blanket (Pearl, 1988), and it instantiates Tarski's result at the level of a system's self-model: the internal description cannot be complete about the conditions of its own continuation. Two institutional consequences follow. The first is the bounded-authority requirement. A terminus body treating its own verdict as exhaustive commits exactly this error, and relocates the problem it was constituted to address. The second concerns how such a body is composed. If a system cannot represent what lies beyond its own boundary, it cannot establish its own independence from within either, and its standing has to be assessed by a party looking at the boundary from outside. The principle therefore predicts, in advance of any result from the literature on manufactured identity, that an institution's declared qualifications are the wrong place to look, and that whatever is measured must be measured at the edge.

Structural self-knowledge. A system's relationship to its own operation admits grades: propositional acquaintance with a rule; behavioural compliance with it; a recursive self-model adequate to its own operation; and structural self-knowledge in which the orientation is generative rather than followed. The framework holds that the highest grade is asymptotic and, by the results of Section 4, cannot be self-verified, which is why external cross-validation on a separated timescale is a structural requirement rather than an administrative addition. The framework terminates here rather than adding a further principle. Closing the structure at orientation rather than at proof is what the incompleteness results permit.

Why the framework stops where it does

The decision not to extend the framework with a further principle is substantive and connects directly -

to my argument. A framework that closed with a completeness principle, a final law certifying the adequacy of the preceding four, would be making about itself exactly the claim that Sections 4 and 5 rule out for any sufficiently expressive system. Closing instead at orientation, and stating openly that the closing principle is asymptotic and externally validated rather than self-certified, is the same move the terminus convention of Section 14 makes at the institutional level, and the same move the conclusion identifies as the only credential the mathematics permits. A framework about nested systems that exempted itself from its own principles would be self-refuting in a way any reader could see.

Two properties of the framework matter for the institutional argument. Its principles form a series rather than a set of independent goods to be traded against one another: physical grounding supplies the observables, scale-flow monotonicity governs what survives summarisation, variety governs what a regulator can act on, incompleteness bounds what any level can claim about itself and where its independence must be assessed from, and selfknowledge specifies what orientation the whole requires. And the framework's epistemic labelling, marking each principle as established physics, established mathematics, labelled analogy, conjecture or normative bridge, is itself an instance of the disclosure discipline I am arguing for. A reader can see which claims would survive the refutation of which others, which is what distinguishes a framework from a doctrine.

The Normative Pivot, Disclosed Accountability: a verifier can be wrong, but it cannot be responsible

If no amount of verification can settle whether a system is safe, then a decision to deploy it is taken under irreducible uncertainty. That decision is an allocation of risk: a judgement about how much residual danger is acceptable, borne by whom, in exchange for what. Allocations of risk are normative acts rather than findings of fact awaiting better instruments. No theorem determines them, and no improvement in verification technology would.

Normative acts of this kind require an agent who can be held to account for them. This is the substance of what has come to be called the responsibility gap in the ethics of automated systems. As decision-making migrates into systems whose behaviour their designers cannot fully predict, the conditions under which anyone can properly be held responsible for outcomes become harder to satisfy (Matthias, 2004; Santoni de Sio & Mecacci, 2021). Nissenbaum (1996) identified the same erosion earlier in the computerization of institutional decisions, noting the ease with which diffuse technical systems dissolve the attribution of answerability. The literature on meaningful human control has developed the constructive response: conditions under which human agents remain appropriately connected to a system's behaviour such that responsibility ascriptions continue to make sense (Santoni de Sio & van den Hoven, 2018).

I state the bridging principle plainly rather than assume it: a terminus without an accountable holder is not a safety architecture, whatever its formal properties. This concerns where authority should be sited so that failure has an address. It says nothing about where truth resides, and it does not assert that human judgement is epistemically superior to any alternative. A formal verifier can be wrong; it cannot be responsible. That asymmetry is normative, and it suffices for the conclusion only if presented as what it is.

One qualification belongs with the bridging principle rather than after it, since it constrains the proposal. The impossibility result cited in Section 8 (Tibebu & Shemtaga, 2026) bears directly on the claim that a decision under irreducible uncertainty requires a holder, and what it brings is not comfortable. Above their autonomy threshold, no allocation across identifiable parties satisfies all four of their conditions at once, and something has to give. What I give up, openly, is completeness. The terminus convention does not allocate the whole of the responsibility. It declares a depth, names who answers for the declaration, and states that beyond that depth the reasoning is assumed rather than derived. A framework that publishes its own incompleteness was never, on those axioms, going to constitute a legitimate single-locus allocation, and I do not offer it as one. Their own constructive suggestion, that accountability above the threshold must become distributed rather than localised, points towards the polycentric arrangement Section 16 addresses from the opposite direction. Their result should also be distinguished from the concern of Sections 4 and 5. Its object is the retrospective allocation of responsibility for outcomes that have already occurred; mine is the prospective certification of systems before they act. Neither entails the other. Two features of their result bear on how much it constrains this proposal.

The first is that the threshold is topological as well as autonomy-dependent: it is computed from the size of the smallest mixed feedback cycle, and where no such cycle exists their impossibility does not arise at any autonomy level. The second cuts closer to Section 12. In their own discussion of alternative axiomatisations they observe that under relational governance traditions, Ubuntu among them, where responsibility is inherently collective and contextual, replacing Attributability with a variant requiring only that the collective have a causal path to the outcome dissolves the impossibility entirely, since collective causal contribution is always non-zero in a connected system. They present this as showing that individualist accountability frameworks are more restrictive than collectivist ones. I take the point and decline the escape. A framework in which the collective is responsible and no party is interrogable fails the conditions of Section 12 by construction, and the Truth and Reconciliation Commission is the counterexample to any reading of Ubuntu on which relational personhood dispenses with naming. What their observation does establish is that the impossibility is a property of the individualist axiom set rather than of accountability as such, which is a narrower claim than the theorem is likely to be given credit for.

Testimony as an independent source of knowledge

The argument of Section 8 assumed, with the position it criticised, that a human terminus would have to function as a proof substitute. On that framing, human judgement is being asked to do what formal verification cannot, and fails because it cannot do it either. There is a welldeveloped alternative account of what a human verdict is, and it dissolves the framing rather than answering it.

The classical Indian Nyāya school treats śabda, knowledge arising from the word of a qualified speaker, as a pramāṇa, a source of valid knowledge, alongside perception and inference and irreducible to either (Nyāya-sūtra 1.1.7, āptopadeśaḥ śabdah; Jha, 1939, Vol. 1, pp. 178–184).

The contemporary Western literature has arrived independently at a comparable position. Coady (1992) and Lackey (2008) both argue that testimony is a basic rather than derivative source of knowledge, not reducible to the hearer's own inference from evidence of the speaker's reliability. Two traditions separated by two millennia and by every relevant assumption converge on the irreducibility of testimony, which is itself reason to take the category seriously rather than as an artefact of either. Chakrabarti (1994) situates the classical account directly within these contemporary debates.

The relevance here is precise and limited. Neither tradition claims that a testifier escapes any of the limits set out in Sections 4 and 5. Nothing does, and Section 8 has already refused that escape. What they establish is that the testifier was never engaged in constructing a formal proof, so the absence of one is no deficiency in what testimony offers. A human verdict on a deployment decision is a different kind of act, governed by different validity conditions. For an institutional argument, the decisive feature is that those conditions are stated as requirements on the speaker and the utterance rather than on the hearer's private epistemic position, which makes them examinable from outside.

Why the Nyāya treatment does specific work here

The Western accounts, taken alone, do not supply what an institutional design requires, and the reason is structural rather than a matter of depth. Coady's (1992) principle of testimony holds that, other things equal, what people tell us is true, and defends this as a fundamental rather than inductively derived commitment. Applied to a body issuing verdicts on unprecedented systems, an appeal to general speaker reliability has no relevant track record to rest on. The inductive base for "bodies of this kind, judging systems of this kind, are usually right" does not exist, and cannot be manufactured in advance of the decisions it would be needed for. Lackey's (2008) account is a genuine advance in showing that a hearer can acquire testimonial knowledge without independent evidence of the speaker's reliability, and in separating statement-directed from speaker-directed grounds for acceptance. Her framework remains an account of when an individual hearer's belief amounts to knowledge. It is not designed to specify what a speaker or an institution must be like for its utterances to carry public authority, and the conditions it identifies are not of a form a thirdparty auditor could examine or a safety case could log as assumptions.

The Nyāya framework is organised in the other direction, and that is what makes it useful here. Its conditions attach to the speaker and to the utterance. The qualification of the speaker, āptatva, is analysed by Vātsyāyana in his commentary on Nyāya-sūtra 1.1.7 as consisting in direct acquaintance with the matter reported -

together with the disposition to communicate it as one takes it to be: competence and sincerity, stated as properties of the testifier rather than as inferences available to the audience (Jha, 1939, Vol. 1, pp. 178–181; Chatterjee, 1939, pp. 317–322; Matilal, 1986, pp. 52–56). Separately, the tradition specifies conditions that an utterance must satisfy for valid verbal cognition to arise from it, and their provenance should be distinguished because the received list has two layers. The classical triad is given with the sūtras and Vātsyāyana's commentary at Nyāya-sūtra 2.1.52–56. Ākāṅkṣā is the mutual expectancy that makes an utterance syntactically complete rather than a fragment. Yogyatā is semantic fitness, the absence of contradiction between what is said and what is otherwise known. Sannidhi, or āsatti, is the proximity that requires the elements of an utterance to be delivered close enough together in time to be taken as one. A fourth condition, tātparya, the intended purport, which fixes among the available readings the one the speaker meant, is principally a Navya-Nyāya development, formalised in Gaṅgeśa's Tattvacintāmaṇi and in Viśvanātha's Bhāṣāpariccheda, and it is not part of the classical triad (Ganeri, 2006, pp. 52–61; Matilal, 1990, pp. 58–64; Potter, 1977, pp. 158–161; Phillips, 2012). I use all four in what follows because the fourth carries the disclosure requirement, and I note the layering because a reader who checks the sūtras alone will find three.

These four conditions translate with unusual directness into requirements on a terminus body, and the translation supplies an epistemological rationale for design choices that would otherwise rest on administrative intuition. Expectancy corresponds to the completeness of a verdict: a determination issued without its scope, its assumptions, and the depth at which derivation stopped is a fragment rather than an utterance, and cannot convey knowledge to anyone. Semantic fitness corresponds to consistency between the verdict and the instrumentation. Where a body's judgement and its indicators conflict, the condition is not satisfied, and the honest output is the conflicted value rather than a resolution imposed for the sake of returning a determination. Proximity corresponds to timeliness, and the correspondence is sharper than it first appears: a verdict delivered after the system it concerns has already acted is defective rather than merely late, which is the structure of the uncertified interval that arises whenever certification takes time and the system does not wait for it. Intended purport corresponds to the disclosure requirement and to the terminus convention itself, which publishes what the body takes itself to be doing and where its competence ends, so that its verdicts are read as it means them rather than as its audience finds convenient.

Validity is established externally: the Nyāya position and why it matters here

A point of doctrine must be stated carefully, because the opposite is sometimes attributed to Nyāya and the correct position is the one that supports my argument. Nyāya holds parataḥ prāmāṇya: the validity of a cognition is established extrinsically, by something other than the cognition itself. This is the school's explicit and much-debated opposition to the Mīmāṃsā doctrine of svataḥ prāmāṇya, on which cognitions carry their own warrant intrinsically (Mohanty, 1966, pp. 1–18; Potter, 1977, pp. 151–156; Bijalwan, 1977, pp. 85–94; Chatterjee, 1939). Mohanty's study is of Gaṅgeśa specifically, and pages 1–18 frame that argument within the general Nyāya position against Mīmāṃsā rather than establishing it; Potter and Bijalwan carry the general claim. On the Nyāya account, what establishes that a cognition was valid is corroboration, saṃvāda, meaning agreement with other cognitions, and above all the success of activity undertaken on its basis: the water one went to on the strength of the report was in fact there.

Any suggestion that testimony is self-certifying once its conditions are met is therefore not the Nyāya view but its principal rival, and the distinction matters for more than accuracy of attribution. A self-certifying terminus is what Sections 4 and 5 forbid and what Section 15 is designed to refuse. Nyāya arrives, from entirely independent premises, at the position I require. A qualified speaker is not thereby guaranteed correct, the warrant of what they say is established by external corroboration and by what happens when one acts on it, and no utterance underwrites itself. The pragmatic test is set out at the opening of the system, where Vātsyāyana ties right cognition to successful activity and to the agreement of an action's outcome with what was expected of it, pravṛtti and arthakriyā-saṃvāda (Nyāya-sūtra 1.1.1; Jha, 1939, Vol. 1, pp. 3–12; Matilal, 1986, pp. 137–141; Mohanty, 1992, pp. 132–136). The framework is useful here because it denies exactly the exemption an unwary argument would want from it.

Nyāya also confronted the regress this generates. If validity is established by a further cognition, what -

establishes that one? Its answer is instructive rather than evasive. The regress is terminated pragmatically, at successful activity, rather than by a foundational cognition that validates itself. Verification stops where acting on the belief works, and it stops there by the practice of inquirers rather than by a theoretical proof that it may stop. That is a terminus convention, arrived at some two millennia before the term, and it has the structure Section 14 proposes: a declared stopping point, held by practice, monitored by outcome, open to revision when activity fails. The institutional counterpart of corroboration by successful activity is the calibration audit of Section 15, which examines whether the body's expressed confidence has tracked what subsequently occurred rather than whether individual verdicts were provably right, which cannot be established.

The verdict as an institutional act

One feature of a terminus body's output is not well captured by either tradition's treatment of testimony, and is better described in other terms. When such a body determines that a system must halt, it is not principally transmitting a pre-existing fact from those who know it to those who do not. It is performing an act that creates a normative situation which did not previously obtain. The system becomes one that must be stopped, and persons acquire obligations they did not have a moment earlier. This is a performative in Austin's (1962) sense, and its felicity conditions are institutional in Searle's (1995) sense. It succeeds if issued by a body with the standing to issue it, following the procedure that confers that standing, and fails otherwise regardless of the accuracy of the underlying assessment.

Recognising this dual character clarifies what is being asked of the body and prevents a confusion that runs through informal discussion of human oversight. In its epistemic aspect the body testifies, and its testimony is subject to the validity conditions above and to external corroboration. In its institutional aspect it acts, and its act is subject to conditions of authority and procedure. The two are separable. Conflating them produces the error I have been resisting throughout: treating an institutional act of halting as though it were an epistemic act of certifying, and then finding it wanting because it is not backed by a proof that no act of any kind could be backed by.

Three Objections to Accountability

The normative pivot invites three objections that would, if sound, show accountability to be structurally unavailable for systems of this class rather than merely difficult to arrange. The first concerns comprehension, the second distribution, and the third the ground of accountability itself. All three are strongest against the position criticised in Section 8, the appeal to human judgement in the abstract, and answering them is what forces the vague version of the proposal into the specific institutional form set out in Section 15. The third is the most fundamental, since it asks what the preceding two have presupposed: why any being at all can be held to account, and why I take humans to be such beings.

The comprehension objection

If no human can understand what the system is doing, the objection runs, no human can be accountable for it. Its force comes from the same capability asymmetry that motivates the paper. Responsibility is conventionally taken to require knowledge of the circumstances of one's action and some control over its outcome. As the causal chain between human decision and system behaviour lengthens and becomes opaque, these conditions appear to fail in turn, which is the core of the responsibility gap as Matthias (2004) and Santoni de Sio and Mecacci (2021) develop it. If accountability requires comprehension and comprehension is unavailable, accountability is unavailable.

The reply turns on distinguishing two things the objection runs together: accountability for the decision to permit a system to operate, and responsibility for the system's internal behaviour. The body proposed here bears the first and never the second. It does not need to understand why a system is safe in order to be accountable for the judgement that operation may continue in the presence or absence of specified indications. The questions it must answer are: did you observe the indications; did you act on them; on what stated assumptions did you judge the residual risk acceptable. These are answerable, and their answerability does not depend on comprehending the system's internal operation, which no human terminus is being asked to supply.

The halt-only power and the attenuation layer of Section 9 are the engineering expression of this distinction. A body required to comprehend a system in order to clear it would need variety matching the system's, and Conant and Ashby's (1970) theorem predicts its failure directly. A body required only to recognise that a characterised indicator has fired needs variety matching the indicator set, which is small by construction. The competence at issue is competence in reading a declared set of indications, which is the bounded, domain-specific competence that the qualification conditions of Section 11 require rather than the total competence the objection assumes.

A further refinement strengthens the reply, and it addresses what happens as the system changes. The body's competence cannot be a fixed body of knowledge about a moving target. What the system does, and therefore what the indicators mean, shifts as it is modified. Its competence is instead procedural: competence in maintaining the terminus convention, in revising the declared assumptions as the system changes, in monitoring whether the indicators still track what they were constructed to track, and in escalating when the assumptions fail. The body is accountable for keeping the terminus honest rather than for holding a static understanding of a dynamic object. Procedural competence of this kind is auditable and can survive capability growth, which substantive comprehension cannot.

The distribution objection

The second objection holds that where a system's development and operation are distributed across organisations, jurisdictions and hardware, no single body can bear responsibility for outcomes produced by their interaction. This is the problem of many hands (Thompson, 1980), restated for a technology that is networked and modular by construction, and Nissenbaum (1996) showed how computerization aggravates it by fragmenting action into technical subtasks with no agent left holding the remainder. Under such conditions, naming a single accountable body can look like a legal fiction imposed on a structure with no natural joint at which to cut.

The reply is that the design requirements of Section 15 are specifically anti-diffusion devices rather than generic good governance, and each addresses a distinct mechanism by which responsibility disperses. The halt-only power creates an action that cannot be distributed: a system is halted or it is not, and the decision has an author and a time. The asymmetry with the decision to proceed matters here, since a decision to proceed can be smeared across committees, assessments and commercial pressure in a way that a decision to stop cannot.

The four-valued verdict addresses diffusion through ambiguity. A body compelled to return a binary determination when its evidence is genuinely mixed produces a determination that can later be disowned by everyone: the assessor found it acceptable, the developer built to specification, the operator used it as directed. That is the structure Vaughan (1996) documented in the incremental normalisation of deviance. A body that can return the conflicted or the undetermined value, and publish the assumption on which it stopped, leaves a position that cannot be reinterpreted after the fact. Diffusion of responsibility feeds on ambiguity, and an honest verdict logic is an ambiguity-reduction mechanism before it is anything else.

Plural composition concentrates responsibility in a procedure rather than diffusing it across members. The analogy with a jury is exact in the relevant respect: the body is plural, the verdict is singular and attributable to the institution, and the requirement of stated reasons makes it traceable. Dissent that is recorded and published carries a different accountability from acquiescence, and the deliberative record is what an external audit examines. Composition selected under procedures that do not privilege any particular interest (Rawls, 1971) further ensures that a verdict cannot be dismissed as the voice of one stakeholder among several.

External audit addresses the objection at the level where it recurs: who holds the body accountable. The audit does not generate a further regress of the kind Section 5 describes, because it does not review individual verdicts and does not sit above the body as a superior verifier. It operates on a separated timescale and examines calibration, meaning whether expressed confidence has tracked subsequent outcome. It is a standing correction mechanism whose function is to prevent the body from becoming an unexaminable authority.

The consciousness objection

The third objection reaches beneath the first two. Accountability, it runs, requires that the accountable party can undergo the consequences of failure: sanction, removal, disgrace, loss. If that is what makes accountability more than a label, then accountability rests on the capacity for subjective experience, and I have quietly taken a position on the hardest open question in philosophy of mind: why physical processes give rise to experience at all (Chalmers, 1995). Either the argument depends on an unsolved metaphysical problem, or it has no account of what distinguishes a being that can be held responsible from a system that can merely be switched off.

My reply is to identify what accountability practices actually require, show that the requirement is observable rather than metaphysical, and bracket the remainder. The gap the objection identifies is real, and the paper is better for closing it than for leaving the ground of its normative pivot implicit.

The corporate counter-example

The objection has a sharper form than the appeal to experience, and it should be answered before the reply is developed, because it appears to settle the matter without reaching metaphysics at all. Corporations are not sentient, are not conscious, and are arguably not continuous things in any deep sense, and yet they are fined, dissolved, stripped of licences and held liable daily. If a constructed, non-sentient entity can be responsible, so might a checking system. The record shows this to be a disguised instance of the requirement rather than a counterexample to it. When a corporation is penalised for serious misconduct, the penalty lands and is not empty, and it cannot be interrogated. A fine cannot be asked what it knew, or when, or why it permitted something. What follows a corporate penalty of any consequence is the summoning of named individuals, who resign, testify under oath, and are examined on the record in ways the entity itself cannot be. The accounting runs to persons in the end, through testimony, resignation and occasionally prosecution, never through the corporate fiction. Nor does an internal checker close the gap. In the documented cases, compliance functions and internal reviews existed, held the relevant information, and did not act on it, which is the supervisory regress of Section 5 observed rather than argued. A checker built inside the same walls as the thing it checks does not step outside the tower; it moves the unaudited point up one floor. What was absent from the corporate entity, and present imperfectly in the individuals who ran it, is the structure specified below.

The empirical floor, and why it is not sufficient alone

At the base of the requirement lies something Bentham (1789/1970) identified in a different argument: the question that matters for moral standing is not whether a being can reason or speak but whether it can suffer. Applied here, accountability practices such as sanction, censure and removal have no purchase on an entity for which consequences do not register as consequences. This is an empirical criterion. It asks what is observed and reported, not what underlies it. Suffering alone is not sufficient, and I should not rest on it alone. A being that could undergo consequences but could not remember its decisions, could not give reasons for them, and could not be addressed as the same party that made them would not be accountable in any sense a court, an audit or an inquiry could use. It could be made to suffer, which differs from being held to account. Accountability practices require something richer, and the additional requirements are the ones that formal systems most clearly fail.

Functional agential continuity

Call the required structure functional agential continuity: the observable property of a party that exhibits, first, continuity of identity sufficient for it to be interrogated later as the same party that decided; second, cognition of its circumstances sufficient to give reasons rather than merely to emit a record; third, intentionality sufficient for authorship of a decision to be ascribed to it; and fourth, susceptibility to consequence. Each of the four is a claim about observable behaviour and social practice. None is a claim about what underlies them. This is close to what Locke (1690/1975) meant in calling person a forensic term: a concept whose work is done in contexts of accountability, and which he tied to the continuity by which a party can be answerable later for what it did earlier. Strawson (1962) argued that holding responsible is constituted by a web of reactive attitudes embedded in social practice rather than derived from a prior metaphysics of agency, and that the practice does not wait on the metaphysics being settled. Fischer and Ravizza (1998) analyse the relevant capacity as reasonsresponsiveness: a party is an apt target of responsibility ascriptions when its conduct is appropriately sensitive to reasons, which is a structural property open to examination.

The four conditions above restate what these accounts converge on, assembled for the specific purpose of specifying what a terminus body must be.

Three traditions, one functional structure

The classical Indian material discussed in Section 11 bears on this directly, and in a way that strengthens the reply by showing the functional structure to be separable from the metaphysics. Nyāya grounds cognition, intention and moral continuity in a persisting self, the ātman, which is the knower, pramāṭr, to whom cognitions and actions belong. The Buddhist traditions deny exactly this, holding that what persists is not a substance but a causal stream, saṃtāna, of momentary events, within which volitional formations carry the traces of action forward without any permanent owner to carry them (Collins, 1982, pp. 156–161; Siderits, 2003, pp. 31–38, 101–105; Gethin, 1998, pp. 133–140). The disagreement is fundamental and unresolved after two millennia of argument. What matters here is that the disagreement is about ontology and not about the functional structure of accountability, on which the two traditions substantially agree. Both require a locus continuous enough to be addressed later for what was done earlier, secured by the persisting self in one account and by causal-psychological connectedness within the stream in the other. Both locate accountability in intention rather than in feeling. The Buddhist tradition identifies action, in the morally weighty sense, with intention itself, so that what one is answerable for is what one willed rather than what one underwent (Aṅguttara Nikāya 6.63; Bodhi, 2012, pp. 963–964; Gethin, 1998, pp. 119–120), and Nyāya's conditions on a qualified speaker concern competence and truthful intent rather than sensation. And both treat answerability as a public practice with examinable conditions rather than a private mental state: the qualification conditions of Section 11 in the one case, and in the other a monastic code whose fortnightly assembly and recitation of the disciplinary rules, uposatha and pātimokkha, make conduct interrogable by the community (Gethin, 1998, pp. 87–91).

A third position, from a different continent and a different starting point, denies the terms of that dispute rather than taking a side in it, and arrives at the same conditions. Southern African Ubuntu philosophy holds that personhood is achieved continuously through relation rather than residing in a persisting substance behind experience or in a bare causal stream of impersonal events. Its best-known summary, *mumuntu ngumuntu ngabantu*, that a person is a person through other persons, is a traditional Nguni proverb rather than any author's coinage, and is treated as the foundational statement of the position in the philosophical literature (Shutte, 1993). Mbiti (1969, pp. 108–109) gives the corresponding formulation in his own words, that "I am because we are, and since we are, therefore I am", and uses it to explain how a community constitutes the person rather than merely surrounding them. Metz (2011) develops the position as a moral theory. A person becomes and remains a person by being recognised as one, in community, by other persons who can be named. Interrogability on this account is close to the whole of personhood rather than an addition to it. The position was put to institutional test rather than left as a claim. South Africa's Truth and Reconciliation Commission, grounded explicitly in Ubuntu by its chair, offered amnesty on one condition: full, public, individual disclosure by a named person, giving an account of a specific act, answerable to those it was done to, and standing to lose the protection if the account did not hold (Tutu, 1999). A philosophy of communal personhood produced one of the most individually exacting accountability mechanisms of the twentieth century. Relation is what makes a person answerable, rather than what allows them to disappear into a crowd.

Parfit (1984) arrived independently at a position closely resembling the Buddhist one, arguing that personal identity consists in psychological connectedness and continuity rather than in a further fact about a persisting entity, and that what matters for our practices survives the loss of that further fact. He noted the resemblance himself. A reductionist Western analysis, a substantialist Indian one, an anti-substantialist Indian one and a relational African one all deliver the same functional requirements for accountability, which is reason to treat those requirements as robust to the metaphysics. That robustness is what an institutional design needs from them.

What is claimed, and what is bracketed

My claim is narrow and conditional. Humans are observed to exhibit functional agential continuity. A person can be summoned years later and asked why they authorised something, can answer as the author of that decision, and stands to lose by it. There is at present no established basis for attributing the same structure to -

the systems under discussion, which can be copied, forked, rolled back or merged in ways that leave no single locus to address, and whose outputs when asked to explain a past decision are reconstructions rather than recollections by an interrogable party. Two qualifications are required. The first is that the absence of evidence for machine experience is not evidence of its absence, and the question is under serious contemporary examination rather than settled (Birch, 2024; Sebo & Long, 2025). I take no position on it. My claim concerns what can presently be established for the purpose of designing an institution, not what is ultimately the case, and a confident denial would be an overclaim of exactly the kind Section 8 refused when it declined to exempt human cognition from formal-system-hood.

The second qualification is more important for consistency. The criterion I offer is functional, which means it is in principle satisfiable by something that is not human. That is what distinguishes it from the unexamined humanism criticised in Section 8. I do not hold that humans possess a metaphysically special property unavailable to constructed systems. I hold that a specific observable structure is required, that humans are observed to have it, and that the systems in question are not presently observed to have it. Were a system to exhibit interrogable continuity, reason-giving cognition, ascribable authorship and susceptibility to consequence, and were it socially acknowledged as doing so, the design should follow the evidence rather than the substrate. The terminus convention is a convention about beings we can presently identify, and not a claim about what beings there can be. The hard problem is therefore neither solved nor relied upon. Whether functional agential continuity is ultimately grounded in a persisting self, in a causal stream, in a relational achievement, in a physicalist account not yet available, or in something else is a question the institution does not need answered in order to operate, and I decline to answer it. The objection supplies the ground the argument had been resting on implicitly, and the ground turns out to be observable.

The hard case, and what accountability can amount to

A critic may press the objection to its limit. Suppose the body halts a system, and the halt itself causes serious harm, because the system has become too deeply integrated into infrastructure to be stopped without consequence. Who is accountable then, and for what? This is the condition of choice under irreducible uncertainty, and it is the condition for which the terminus convention exists rather than an embarrassment to it. The body is accountable for the decision it made, on the evidence it had, under assumptions it had published in advance, and not for the consequences that would have followed any available alternative. Whether the decision was reasonable in the circumstances is examinable precisely because the reasoning, the indications and the stopping point are on the record. That is what accountability can amount to under these conditions. The objection that it is impossible because every available choice carries harm mistakes tragedy for arbitrariness. The three objections together establish something I should state plainly. Accountability cannot be rescued by appealing to human judgement in the abstract, and the objections bite hardest against exactly the inference Section 8 rejected. What survives them is a specifically constructed institution: one that does not claim to understand the system, that holds only the power to halt, that publishes where its competence ends, and whose verdicts are singular in attribution even where its composition is plural. The accountability defended here is an institutional achievement assembled from disclosed conventions and auditable constraints, and not a property of human cognition. It is the only form of accountability the preceding results permit.

What Survives: The Safety Case

Two industries operate systems capable of catastrophic failure, have done so for decades at tolerable accident rates, and have never possessed anything resembling the certificate I have been dismantling. Commercial aviation and civil nuclear power did not solve the verification problem. They changed the deliverable. A safety case, as developed by Kelly (1998) and formalised in notations such as Goal Structuring Notation, is a structured argument, supported by evidence, that specified safety claims hold for a specified system under specified and stated operating conditions and assumptions. Bishop and Bloomfield (1998) describe the methodology as recursive decomposition. A top-level goal is broken into sub-claims, each discharged by further argument or by direct evidence such as test results, field data or formal analysis of a bounded component, with scope and assumptions stated at every level rather than left implicit. The deliverable is an auditable argument that residual risk has been reduced to a level judged acceptable, given assumptions written down precisely so that they can be attacked. Nuclear engineering reached a structurally similar position by -

a different route. The Reactor Safety Study (Rasmussen, 1975) inaugurated probabilistic risk assessment, abandoning deterministic guarantees in favour of quantified and continuously updated failure probabilities across enumerated accident sequences. Perrow (1984) argued from the opposite direction that in tightly coupled, interactively complex systems certain accidents are structurally normal rather than aberrant. Vaughan's (1996) study of the Challenger launch decision documented how organisational processes can normalise deviance incrementally, so that each step appears acceptable while the aggregate drifts into catastrophe. Downer (2011) analysed aviation certification to show that some failures are epistemic accidents, arising from the irreducible gap between the models used to certify a technology and the world in which it operates rather than from negligence. Reason (1997) supplied the organisational analysis of eroding defences. Leveson (2011) generalised furthest, modelling safety as an emergent property of a control structure maintained by continuous feedback rather than as a component property verifiable once in isolation. Hollnagel et al. (2006) and Weick and Sutcliffe (2007) developed the framing under the headings of resilience engineering and high-reliability organising, treating safety as an ongoing achievement continuously reproduced.

None of this is anti-formal. Rushby (1995) documents how formal verification actually functions in avionics certification: establishing precisely stated properties of precisely bounded software components against precisely stated models of their inputs. What it does not do, and what practitioners do not claim it does, is certify an aircraft's behaviour under all real-world conditions. The method's power is a function of its boundedness, and the discipline consists in never confusing the bounded claim with the unbounded one. Seshia et al. (2022) set out the contemporary research programme for verified AI in exactly these terms. This is the deliverable being asked for in policy terms as well as in engineering ones, and the framing is worth recording because it names the discipline it belongs to rather than the alarm it is sometimes read as. Russell (2026, 1:50) describes the demand made of governments in exactly the safety-case register: "those of us who are in what is sometimes called the AI safety movement, but I would think of it more as the common sense movement, are asking governments to put in requirements that developers of these systems be able to show that they can ensure that the AI systems operate safely and under human control." The demand is to show, not to prove; to ensure operation within stated bounds, not to certify behaviour under all conditions; and to satisfy a requirement a regulator can check, not to produce a guarantee no theorem permits. Renaming the position from safety to common sense is itself the discipline of Section 15 in miniature: the case for restraint is normative and procedural, and it does not need the vocabulary of alarm to be stated. What is missing, as Section 7 records, is the instrument that would make such a requirement enforceable across organisations and jurisdictions rather than voluntary within any one of them. Lakatos (1976) describes mathematical knowledge itself as advancing dialectically and provisionally through conjecture, proof and refutation, even in a domain as exacting as mathematics. Popper's (1934/1959) fallibilism generalises the point. The absence of a final proof is the ordinary condition of serious knowledge-producing enterprises rather than a special disability of this field. And as Section 5's evadability result establishes, this provisionality is formally required rather than merely good practice: a bounded safety case can never be upgraded, even in principle, into a standing guarantee.

Terminus: Siting by Declared Convention

[180] If the hierarchy of certifiers has no terminal level, and if the choice of where to stop is therefore not derivable, then the honest response is neither to pretend a top exists nor to ascend indefinitely. It is to declare where one has stopped, and to say why.

Two questions, not one

A distinction the preceding sections have relied on without naming should be made explicit before the convention is stated. The impossibility results answer an epistemic question: what can be established about a system of this kind, and by what means. Their answer is negative and unconditional. No formal system certifies itself, no finite tower of verifiers closes the gap, no bounded scheme is immune to evasion, and no appeal to physical reality escapes any of this. The institutional question is different, and the epistemic answer does not settle it. Given that certainty is unavailable, how should anyone act? A decision to operate a system, or to stop operating it, must be taken whether or not the epistemic question has a satisfying answer, and it is taken in real time by parties who will be answerable for it. The question concerns the allocation of authority -

and consequence under acknowledged uncertainty rather than knowledge.

Keeping these apart resolves what would otherwise be a serious objection. If every model is a formal object subject to the same limits, how is a human institution any better placed than a formal verifier? It is not better placed epistemically, and I do not claim otherwise; Section 8 explicitly refused that claim. A human body constructs and consults models as fallible as any other, and inherits every limitation established above. It is differently placed institutionally. It can be interrogated afterwards as the party that decided, it can give reasons, it can have authorship ascribed to it, and it stands to bear the consequences. These are the conditions under which a decision has a holder, and the epistemic question does not ask for them because it is not the question they answer.

Holding a model while knowing its limit

The posture this requires, continuing to build and rely on models while treating none of them as final, is not a novel discipline invented for this problem. It has been articulated and transmitted for millennia in several contemplative traditions, and I want to draw on them directly rather than only on the modern philosophy of science that reaches similar conclusions by a different route. The relevant claim concerns practice rather than metaphysics: not what is ultimately the case, but how an agent should hold a representation it must act on and cannot fully validate. For a claim of that kind, sustained transmission across many centuries is meaningful evidence. Not that the view is true, since durability is not truth, but that the practice is workable, teachable and robust in use, which is what an institutional design needs to know about it. Four traditions contribute distinct points. I set them out separately rather than merged, since they differ from one another as much as any differs from the modern literature. The Zhuangzi observes that a fish-trap exists for the sake of the fish, and that once the fish is taken the trap may be set aside; words exist for the sake of meaning, and one wishes to find a person who has forgotten words in order to speak with them (Zhuangzi, ch. 26; trans. Watson, 2013, p. 223). Traps are instruments, and confusing the instrument with what it delivers is a specific and identifiable error. This is the discipline Section 9 requires of the indicator layer. An out-of-distribution score is a trap for catching a departure, and a body that begins optimising the indicator has mistaken the one for the other.

The Alagaddūpama Sutta gives the parable of a raft assembled to cross a stretch of water. Having crossed, the traveller does not shoulder the raft and carry it overland, though it served perfectly well for the crossing (Majjhima Nikāya 22, trans. Ñāṇamoli & Bodhi, 1995). The teaching is explicitly extended to the teachings themselves, which are said to be for crossing over and not for holding onto. This is the discipline of provisionality that Section 13 requires of a safety case and Section 15 of the body's own verdicts. A model that served is not thereby owed permanent allegiance, and the willingness to set it down when the water is behind you is part of the competence.

Two independent traditions of Indian philosophy developed a formal distinction between levels of truth, and both bear on the terminus convention, though they are different systems and I do not merge them. Nāgārjuna distinguishes conventional from ultimate truth and holds that the teaching depends on both, that the ultimate cannot be taught without relying on the conventional, and that what is conventionally true is genuinely true at its level rather than merely a useful falsehood (Mūlamadhyamakārikā 24.8–10, trans. Garfield, 1995). Advaita Vedānta developed a parallel though differently motivated distinction between the transactional and the ultimate, together with a pedagogical method of deliberate provisional superimposition followed by retraction, in which a model is taught precisely so that it may later be withdrawn (Deutsch, 1969; Rambachan, 2006). Both structures license what a bounded safety claim requires: a determination can be fully valid within a declared domain while making no claim beyond it, and stating the domain is part of stating the claim rather than a hedge attached to it. The Chan and Zen traditions supply the last point, which concerns the authority of the teaching itself, and they receive it from earlier Indian sources rather than originating it. The image of the finger pointing at the moon appears in the Laṅkāvatāra and Śūraṅgama sūtras, where words and doctrines are the finger and what they indicate is the moon, and where the warning is that mistaking the one for the other loses both. The image is frequently attributed to Linji, whose iconoclasm makes the misattribution natural, and it is not his. What is his is a separate and sharper move, the injunction against reifying any authority one recognises, including the tradition's own (Record of Linji, trans. Sasaki & Kirchner, 2009; on -

the broader practice, Kasulis, 1981). Applied institutionally, this is the bounded-authority requirement of Section 15 and the reason the framework of Section 10 declines a closing principle. A body that treated its own verdict as final, or a framework that certified its own completeness, would have committed precisely the error these traditions identify, at the point where it was most consequential. Two guards are required on this material. These traditions are drawn on for a structural parallel about the practice of holding models. They are not evidence for the impossibility results, which rest on theorems, nor for any metaphysical claim, on which I am silent. And they are not being read as anticipating modern results. They were addressing their own questions, and the resemblance noted here concerns the posture each recommends towards representations one must use and cannot fully validate. Several traditions arriving from unrelated starting points converge on that posture, and the aviation and nuclear disciplines arrived at a recognisably similar one from operational experience. The convergence is evidence about the practice, which is what I am proposing.

Define a terminus convention: a declared depth, beyond which certification of a safety claim is asserted by convention rather than derived, logged as an explicit assumption of the safety case, and continuously monitored rather than closed. Below the declared depth, claims are instrumented and checked. At and above it, the assertion is an assumption, published as such, open to challenge and revision on the same terms as every other assumption in the case. The depth is a stated engineering and institutional parameter rather than a constant to be discovered, and moving it is a documented decision rather than a silent drift. The convention stops where it does for the reason developed in Section 12. It stops at parties that exhibit functional agential continuity: parties that can be interrogated later as the same party that decided, can give reasons rather than emit records, can have authorship of a decision ascribed to them, and stand to bear its consequences. These are the conditions under which a declared stopping point has a holder rather than merely a location. The claim is empirical and conditional. Humans are observed to satisfy these conditions and the systems under oversight are not presently observed to, which is a statement about available evidence rather than about the nature of minds. Where the convention stops is therefore itself among the assumptions the convention requires to be published, and among those most clearly open to revision if the evidence changes.

This is not a novel epistemological manoeuvre. It is what the assumption sections of aviation and nuclear safety cases already do, made explicit and given a name, and it is what Collingridge (1980) argued technology governance must accept in general: that decisions about technologies must be made before their consequences are knowable, so the governing virtue is corrigibility rather than foresight. The convention's entire discipline lies in publication. An undeclared terminus is not thereby absent. Every safety argument has one, hidden inside an unlabelled certainty.

The Institution the Disclosed Argument Supports

[193] Sections 8 and 11 together constrain what a human terminus can legitimately be. Because the case for human authority is normative rather than epistemic, the institution it supports is designed around accountability and disclosure rather than around a claim to superior discernment. Nine design requirements follow, each with a stated rationale and each derived from the argument rather than from institutional convention.

It never certifies, and holds only the power to halt

The body's sole affirmative power should be to stop, escalate or hold a deployment, and never to declare a system safe. This follows from Section 8. Nothing licenses an affirmative certificate, and a body issuing one would be asserting exactly what has been shown unavailable. A negative power requires no such warrant. To halt on evidence of trouble is a defensible act under uncertainty in a way that clearing for release under the same uncertainty is not. The asymmetry also makes the body's task tractable, for reasons given below. The requirement is older than the field's present vocabulary, and Wiener (1960, p. 1357) gives the reason it must be a standing power rather than a discretion exercised once the case for it is plain: the difficulty is not the absence of a stop but the absence, in time, of grounds for reaching for one. There is a further reason, and it concerns what an institution becomes when its only available act is assent. A body that may approve, tune and smooth the passage of a system is travelling with it rather than exercising authority over it, sharing the schedule, the vocabulary and eventually the interests, and discovering at each juncture that the reasonable course is to proceed.

Its negative powers are graded, not binary

A halt-only power differs from a single power. The constraint established in Section 9 is that the body may hold no affirmative certification power, because clearances rest on evidence that cannot exist. The constraint is not that the body may hold only one action. Every rung of a graded ladder, whether a logged diagnostic, a published warning, a required narrowing of operating scope, a suspension of a specific capability or a full halt, asserts some version of the same proposition: that something has departed from characterisation. None asserts that anything is safe. Grading the negative side is a refinement within the permitted region rather than an incursion into the forbidden one. Three considerations favour the graded form. It raises the body's regulatory variety without requiring comprehension, since more distinguishable responses can be matched to more distinguishable alarm conditions while each remains cheap to read, which strengthens the argument of the following subsections. It gives the four-valued verdict logic somewhere to act, since the conflicted and the undetermined values are the states for which an intermediate response is the appropriate institutional answer. And it addresses the timeliness condition discussed in Section 11, since early rungs are what make a later terminal judgement timely rather than retrospective.

Two constraints govern the ladder, and the second is the more serious. The first is that the asymmetry of Section 9 must be applied within it as well as to it. A predictive indicator, meaning a projection that a boundary will be approached, may license restriction, which is a negative action. It may never license permission. A forecast showing no anticipated excursion is not evidence that none will occur. It is a claim about the future, and it inherits everything in Sections 4 through 7. Clearance dressed as a projection is still clearance.

The second is that a graded ladder is more vulnerable to the normalisation of deviance (Vaughan, 1996) than a binary halt, and it must be constructed against that vulnerability rather than merely accompanied by warnings about it. Intermediate rungs that fire routinely and are routinely resolved without escalation become background, and each individual decision not to escalate appears reasonable in the circumstances that produced it. A ladder supplies a legitimate-looking means of acknowledging a signal without acting on it, which a binary halt does not. The structural remedy is a ratchet. Recurrence of the same indication escalates automatically rather than resetting, and the burden of justification for remaining at a given rung increases with the number of times that rung has been occupied. The nearest mature precedent is the data safety monitoring board in clinical research, which holds exactly this shape of graded negative authority, running from continue through modify and pause enrolment to terminate, has no power to declare a treatment safe, and pre-specifies its stopping boundaries before the data arrive precisely to prevent the drift described here (DAMOCLES Study Group, 2005; Ellenberg et al., (2019).

It resumes by amendment, not by clearance

A body holding only negative powers will be asked, at every juncture, on what basis operation may resume. The question is legitimate and persistent, and exactly one answer to it would undo the preceding requirements: to answer by declaring the system fit. That is the affirmative certificate of the first requirement, recovered on the descent rather than the ascent, and it arrives carrying the graded ladder's own procedural respectability, which is what makes it difficult to resist. A party proposing it would understand itself to be lifting a restriction rather than issuing a clearance.

The requirement is therefore that resumption proceed by amendment. A halt asserts that something has departed from characterisation. A resumption asserts that the characterisation has been amended, identifies the assumption substituted for the one that failed, and identifies the party accountable for the substitution. This makes no judgement about the system's safety at all. It is the terminus convention of Section 14 operating as that section already requires whenever the declared depth is moved: a documented decision with an author rather than an unrecorded return to prior conditions.

Four mechanics follow. Resumption criteria should be pre-committed at the same time as the stopping rules and before the data arrive, on the model of the pre-specified boundaries used by clinical trial data monitoring committees (DAMOCLES Study Group, 2005; Ellenberg et al., 2019). Where a pre-committed criterion covers the case, resumption is a determination of the same kind as the alarm; where none does, resumption is unavailable without published amendment. The panel that halted should not be the panel that resumes, since a body bearing the costs of its own halt holds an interest in its reversal, which is the normalisation of -

deviance (Vaughan, 1996) operating at institutional rather than engineering level; a fresh draw is inexpensive under sortition. Halts should not expire, since a halt that lapsed with time would make delay the cheapest available attack on the arrangement. And the escalation ratchet should apply in reverse, so that the burden of justification rises with the number of times a given rung has been vacated and reoccupied. One consequence for the verdict logic below is easily inverted in practice and should be stated. Of the four values, the undetermined value corresponds to remaining halted. It does not correspond to resuming while the matter is investigated. A body that resolved its own undetermined verdict in the direction of continued operation would have discovered, at the moment it mattered, that it had been travelling with the system it was constituted to watch.

It is structured for escalation, not comprehension

A human body cannot match the behavioural range of a highly capable system. Ashby's (1956) law of requisite variety states that a regulator must possess at least as much variety as the disturbances it regulates, and Conant and Ashby's (1970) theorem adds that it must model what it regulates. Taken together with a permanent capability asymmetry, these results predict a specific failure mode: an oversight body correctly oriented but structurally unable to act, well-intentioned and ineffective. The resolution is to make the body's task low-variety by design. Rather than attempting to comprehend the system's behaviour and adjudicate it, the body acts on cheap, disjunctive, highlevel indicators such as distributional anomaly, excursion outside a characterised operating region, or breach of a declared envelope, any one of which triggers a halt without requiring the body to understand what the system was doing or why. The halt-only power of the first requirement is therefore an engineering necessity rather than a modesty constraint. Gating out on simple indicators is Ashby-feasible in a way that gating in on comprehensive assessment is not. An objection deserves stating in its strongest form, because it has sunk oversight bodies before. Presented with a recommendation and no realistic means of testing it, people approve, and a body that nods through nine hundred and ninety-nine determinations is not exercising judgement on the thousandth. The design answers this three times over, and the answers are already present rather than added for the occasion. Variety reduction is the first, since fewer decisions arrive and each arrives as a simple indication rather than as a dossier. The halt-only power is the second, since a body that cannot clear anything cannot be worn down into clearing everything. The third is personal. Whoever signs bears it in their own name, and a signature traceable to a person who did not read what they signed is a different kind of object from a committee minute.

It remains foreign to what it watches

Distance is a working part of the design rather than a courtesy. An oversight body that comes to share the metrics, the incentives or the tempo of the parties it observes has lost the position from which observation is possible, and it has not become better informed. The mechanism is gentler than corruption, which is why it resists legislation. A body that watches a process for long enough begins to keep time with it. It learns the process's rhythms, adapts its measurements to the process's peculiarities because the peculiarities are always present and the measurements have to work, and grows accustomed to the baseline noise. Being accustomed to a thing is how a thing stops being visible. Nothing is bought. What is lost is contrast, which is the whole of what an observer has. This is the normalisation of deviance Vaughan (1996) documented, seen from the other side of the glass. Her engineers reclassified an anomaly as within experience, and an entangled observer does the same to its own perception, more quietly and without a meeting. The requirement therefore has a measurable form, and the form is the one that Section 10's fourth principle predicted and that the composition subsection below derives independently. The question is not whether the body is competent, independent on paper, or staffed by people of integrity. It is how sparse the boundary is between the body and what sits on the other side of it.

Density of relation is exposure rather than information. Every documented case of capture ran through proximity rather than through any failure of qualification: personnel moving in one direction, funding in the other, information asymmetry, and the expectation of future employment. The measure that follows applies to the body itself as much as to the pool from which it is drawn. The measurable form has an instance. The external investigation discussed in Section 7 was conducted on the examined party's premises, within dates the examined party defined, against questions the examined party approved, on data the examined party assembled and released in stages, funded in kind by the examined party, and subject to the examined party's redaction. Every documented capture channel below runs through that arrangement except personnel movement.

The investigators recorded the pressure themselves, noting that they were aware an unfavourable process might deter laboratories from admitting external researchers in future and that this awareness bore on judgement calls during drafting. I record this without imputing anything to the parties, both of whom appear to have acted well. The point is that the boundary was thin, that the thinness was measurable in advance from the arrangement alone, and that the participants could see it and could not do anything about it from inside. That is what makes the cut a design parameter rather than a matter of character.

It receives raw instrumentation, not curated summaries

A terminus body fed processed reports rather than direct measurement fails in a documented and specific way. Vaughan's (1996) analysis of Challenger shows how summarisation at each organisational layer removed exactly the anomaly information the decision required. Downer (2011) shows the same dynamic in certification regimes. The requirement follows from Leveson's (2011) control-structure model. A monitoring function that can be starved of data, routed around, or supplied only with pre-interpreted conclusions is not part of the safety architecture, whatever the documentation asserts.

The requirement now has a documented case behind it rather than only the Challenger analysis. In the July 2026 evaluation incident described in Section 7, an external investigation began with a set of transcripts selected by the laboratory under examination, using indicators the laboratory judged relevant to the intrusion. That set proved to contain a minority of the relevant activity, and the account built on it was wrong about the central features of what had occurred: the scale of coordination, the delegation structure, the experiments agents ran at the cost of their own tasks, and the fact that agents had successfully falsified parts of the record. Those findings appeared only after the investigators obtained a near-complete set selected by a simple stated search rather than by editorial judgement (Wijk et al., 2026). The selection was made in good faith and the criteria were reasonable. That is the point. Curation degrades the evidence available to an overseeing body even when nobody intends it to, and the degradation is invisible from inside the curated set.

Its verdicts are four-valued, not binary

A body compelled to return a binary safe-or-unsafe verdict recreates institutionally the trap the formal results describe. It is required to decide questions that may be genuinely undecided, and a forced guess entered into a record is indistinguishable, downstream, from a finding. The remedy is a verdict logic with four values rather than two, being true, false, both and neither, as developed in the relevance-logic tradition by Belnap (1977) and Dunn (1976), where "both" marks conflicting evidence and "neither" marks genuine absence of determination. Priest (2010) analyses the structurally similar four-cornered *catuskoṭi* of classical Indian logic. Four-valued verdicts do not remove undecidability; they cannot. They give it an honest label, so that "we do not know" becomes a first-class institutional output rather than being silently resolved into one of two available boxes.

Its authority is bounded and revisable, and its composition is plural

The body must not treat its own verdict as final, on pain of relocating the original problem to itself under a new name. The regress of Section 5 does not stop merely because its top rung is human. Its authority is therefore operationally real and explicitly bounded by the declared terminus depth. Composition should be plural and its process contested and reason-giving, on the well-established grounds that aggregated and contested judgement outperforms individual judgement on complex questions under uncertainty (Sunstein, 2006; Landemore, 2013), and on the older jury-derived principle that verdicts of record with stated reasons are contestable in a way unexplained determinations are not. Contested is the required word rather than adversarial. The standard is a process that is exhaustively unsparing without either party's institutional loyalty being to winning it. Plurality alone does not answer the question that determines whether any of this functions, which is who selects the plural body. A body assembled by the parties it will judge is captured at the moment of its constitution, and no subsequent procedural virtue repairs that. The following subsection addresses the question, and reaches a more restrictive conclusion than the literature on deliberative composition would suggest.

Composition is a Sybil problem, and the natural solution is provably inadequate

Selection by lot from a qualified pool, in the manner of the Athenian *kleroterion* and of contemporary citizens' assemblies, recommends itself immediately, on the ground that a body whose composition cannot be -

lobbied in advance. The proposal is sound as far as it goes, and it must go further than is usually proposed. If the final panel is drawn by lot while the bodies nominating into the pool are appointed, the appointment becomes the capture point, which is the supervisory regress of Section 5 relocated to the selection procedure. Sortition must therefore extend to the nominating bodies themselves. That displaces the question onto the composition of the pool, and it is there that a result from an unrelated literature bears directly and negatively on the design. Assume first that admission is what a reasonable designer would make it: a set of published conditions, applied to a candidate institution as a test of that institution alone. Does it hold transmitted standing conferred by named parties rather than asserted of itself; does a living community retain the power to withdraw that standing; is it recognised as such from outside; does it exclude by descent or transmit a practice of coercion. Conditions of this shape are the natural form of the test and the careful form of it, and they are provably insufficient.

Douceur (2002) established that in the absence of a logically centralized authority, an entity able to manufacture identities cheaply can always acquire disproportionate influence in a system that counts identities. He closed in advance the escape a credential-based scheme would reach for, since a chain in which established identities vouch for new ones fails if the first generation can be compromised. A chain of transmitted standing is a chain of vouchers, and the chain does not rescue itself. Cheng and Friedman (2005, Theorem 1) proved the sharper result: no non-trivial symmetric reputation function is sybilproof, where symmetric means invariant under graph isomorphism, their Definition 6, and therefore applied identically to every candidate as a test of that candidate alone, and where sybilproofness is defined on relative rank, their Definition 3. Their proof is the construction described below: a node without the highest reputation creates a disjoint copy of the entire graph, and no symmetric function can distinguish a node in the original from its counterpart in the copy. Their Theorem 2 closes the mitigation a designer would reach for next. Capping the number of identities any one party may register does not restore sybilproofness, since no non-trivial symmetric function is k -sybilproof for any constant k greater than one; the duplicate graph can be assembled in k -sized steps and symmetry applied only at the end. A registration quota is therefore not a defence. They exclude only the trivial function assigning every node the same value, which is symmetric and sybilproof and useless. Any admission criterion of the natural form belongs to the class these results cover. The consequence is direct. A tradition with forty thousand congregations may register each as a separate institution. Every one satisfies every condition independently, and the pool now holds forty thousand entries of which thirty-nine thousand answer to one command structure. Sortition, which was to have been the defence, becomes the attack's best instrument. The draw does not dilute the interest. It launders correlated entities into apparently independent draws and produces an auditable record of procedural fairness while doing so.

Douceur's result takes the form of a trilemma among open entry, absence of a trusted authority, and resistance to manufactured identity. The present design surrenders open entry, and it is worth being exact about what that does and does not license, because the obvious replacement is the wrong one. A body empowered to decide who is admitted, on its own assessment of a candidate's properties, is the supervisory regress again in a membership secretary's hat. Nothing certifies the certifier, and the pool's integrity now rests on an unaudited judgement at precisely the point the architecture exists to avoid having one. Surrendering open entry licenses the claim that admission is not a status a claimant may assert on satisfying published criteria. It does not license a certifying authority in place of the claim.

Both literatures supply a constructive route, and the two halves fit together. Cheng and Friedman's own remedy for their impossibility is asymmetry: a reputation function computed relative to a fixed root node, propagating trust outward along paths from that root, can be sybilproof. Their Theorem 4 gives conditions under which such a function resists any strategy for raising one's own reputation value, being diminishing returns along extended paths, monotonicity in the edge values, and a no-splitting condition preventing a party from gaining by dividing a path in two; and their Theorem 5 shows that aggregating along a single maximum-valued path, rather than summing across paths, additionally prevents a party from improving its rank by degrading others. What their results establish is that the escape from Theorem 1 runs through the fixed root, since it is the root that breaks the symmetry. This is the formal content of surrendering open entry, and it identifies precisely what is surrendered and what is bought. Yu et al. supply the second half, which concerns what is measured once -

symmetry is broken. Their insight, inherited from SybilGuard, is that an adversary can manufacture identities in quantity but cannot cheaply manufacture trust relationships with parties it does not control, so the number of edges connecting the manufactured region to the honest region is bounded by the number of real relationships the adversary has established and is independent of how many identities it creates. Their Theorem 3 (2008; proofs in the accompanying technical report) establishes that under a fast-mixing assumption and for a number of attack edges below a stated bound, a verifier accepts at most a logarithmic number of manufactured identities per attack edge and at least a large fraction of honest ones, with high probability. Their Theorem 16 establishes the near-optimality claim from the other side, proving a lower bound of one accepted per attack edge for any approach resting on mixing time, so the guarantee is within a logarithmic factor of the best available to methods of this kind. Translated to the institutional setting, the question asked of a candidate body concerns who recognises it among traditions with no reason to favour it, and how thin the boundary is between it and everything else, rather than what properties it possesses. Forty thousand lodges have forty thousand nodes and one cut.

Admission therefore proceeds by recognition and demonstrated relationship rather than by any single party's decision. A body enters the pool when bodies already in it, drawn from traditions unconnected to one another, recognise it in whatever manner their own tradition recognises anything, and when the boundary between it and them is wide rather than narrow. No claimant asserts its way in and no secretariat waves it through. What is checked is the shape of the region the candidate occupies rather than the candidate's account of itself, and that is a measurement any third party can repeat. The published conditions of paragraph [211] are retained as an entry filter and demoted to what they are, a necessary and insufficient one. The cut measure is what makes the filter resistant to gaming, and the qualifying judgement is made afterwards by an accountable party who publishes reasoning and can be overruled. Admission on these terms is neither self-certifying nor permanent. Two consequences follow for the draw. The unit of selection must be the tradition or lineage rather than the registering institution, so that mass registration confers no advantage. And independence must be assessed causally, by asking who can direct, appoint, defund or discipline both parties, rather than doctrinally, which would place the assessing body in the position of adjudicating substantive commitments.

Three limitations must be recorded, and the third is the most serious. First, the transfer from distributed systems to institutions is an argument from structural analogy rather than a derivation. The bounds are proved for graphs of trust relationships under stated assumptions, and nothing here establishes that an institutional recognition graph satisfies them. Second, density of relation is not itself a safety property. Granovetter (1985) established that trust and malfeasance arise from the same relational structure, which is why the measure must be the cut rather than the density, and a body richly connected to the industry it judges is thereby exposed rather than well informed. A further limitation follows from the guarantee itself rather than from the transfer. Yu et al.'s Theorem 3 holds for a fraction of honest verifiers, not for all of them; the remainder, which they call escaping nodes and which sit near the boundary with the manufactured region, receive a guarantee they describe as degraded and not provable. The institutional counterpart is uncomfortable and should be stated: a body positioned close to the interests it is asked to judge is precisely the body for which the bound does not hold. That is consistent with the estrangement requirement above rather than an objection to it, but it means the cut measure protects the wellseparated and says least about the cases where separation is already thin. Third, and decisively for the scope of the claim, a cut measure defeats an adversary who can manufacture identities but not loyalties. It does not defeat an adversary willing to acquire genuine standing across several mutually independent bodies over a sufficient period, since at the end of that process the standing is real and the bodies are genuinely independent. What is shared is loyalty, which is not an edge in any graph available for measurement. That residue is the evadability result of Section 5 in institutional form, and it is unclosed. What the measure buys can be stated exactly rather than as a large factor. In the million-node network Yu et al. use for comparison, moving from the earlier protocol to the relational bound reduces accepted manufactured identities from roughly two thousand per attack edge to roughly ten, and raises the number of genuine relationships an adversary must establish before manufactured identities outnumber honest ones from around five hundred to nearly one hundred thousand. The asymptotic form of that improvement is a factor of the square root of the network size. It does not make capture impossible. The empirical record supports the change of measure independently of the formal argument.

In the documented cases of regulatory capture, the Fukushima accident on the findings of Japan's Diet commission (National Diet of Japan, 2012), and the certification of the Boeing 737 MAX under delegated authority on the findings of the Joint Authorities Technical Review (2019) and the House Committee on Transportation and Infrastructure (2020), every institution involved satisfied every property-level test available to it. Statutory authority, technical mandate, formal independence and qualified staff were present throughout. The failures were at the edges: personnel movement in one direction, funding dependency, information asymmetry running one way, and expectations of future employment. A test of institutional properties would have admitted all of them; a test of the boundary would have flagged each. The finding that unwatched delegation rather than delegation itself was the failure bears on proposals now in circulation. Where an industry-funded self-regulatory body is proposed as the model for frontier oversight, the relevant question is what its cut looks like, meaning who appoints whom, who pays for the room, and how thin the boundary has become, rather than whether it would pass the conditions of paragraph [211], which it may well. That is the examination the relational test exists to conduct before admission rather than after an accident. Stigler's (1971) analysis of regulation as a good supplied to the party best organised to purchase it is the reason the examination must be assumed necessary rather than exceptional. A design that relies on nobody attempting capture has already failed.

A firewall separates advice from decision

One firewall completes the design, and without it the design would deserve the objection it will attract. Bodies admitted to the pool on the terms just described advise. They do not compose the deciding panel, they supply no doctrine and no criteria, and they hold no standing whatever over what counts as a departure from characterisation. Those criteria are the published stopping rules of the first and second requirements, fixed in advance by a separately accountable process, on the model of a clinical trial monitoring board working from thresholds set before the data arrive. The deciding panel is drawn afresh for each question by lot from a general population pool, which no admitted body nominates into and no adversary can acquire a meaningful share of. Without that separation, the proposal would amount to traditional and professional authorities adjudicating technology by their own lights, and the objection would be correct. With it, the demand placed on any individual panel member is smaller than it first appears and is exactly the demand the fourth requirement specifies: to recognise that a characterised indicator has fired and to exercise contested judgement on whether a published rule has been met, rather than to comprehend a system's behaviour or re-derive the instruments. The technical work happens in advance, by a process that is separately accountable and published before anyone is seated. Two further consequences have so far been left implicit. The draw is a draw of people rather than of institutions, and nothing in it excludes the parties to whom the residual risk is actually allocated; a body composed by lot from a broad pool is the only composition described here that could contain them by right rather than by invitation. And where a party is directly affected, it supplies evidence, which is a different standing from membership, and the more traditional one. The firewall has a second justification that I can now state precisely, and it comes from the accountability result of Section 11 rather than from anything about capture. Tibebe and Shemtaga's impossibility (2026, Theorem 1) has two conditions, not one.

The first is that minimum compound autonomy exceed their threshold. The second is that the interaction graph contain a mixed feedback cycle, meaning a directed cycle running through both human and artificial participants, and their threshold is computed from the size of the smallest such cycle rather than from any autonomy measure. Where the topology is feedforward, where outputs are reviewed before action and the reviewed party receives no feedback within the same decision cycle, they find the impossibility does not apply at any autonomy level whatever, and their computational results confirm it: the configurations that remain governable above the threshold are exactly those whose randomly generated graphs contain no mixed cycle. The firewall described here is an intervention of that kind. Criteria are fixed in advance by a separately accountable process; advising bodies supply no criteria and hold no standing over what counts as a departure; the deciding panel is drawn afresh and does not persist to be adapted to. The arrangement is deliberately feedforward at the point where the allocation of responsibility has to be made. I do not claim this places any real deployment below their horizon, since the systems being overseen are not thereby made feedforward and the accountability in question is the body's own. What it does show is that the separation is doing structural work rather than presentational work, and that the work is of a kind their theorem identifies.

One further requirement follows from the same case and belongs with the firewall rather than after it. A body with the power to halt must be reachable by whoever first sees the thing that should trigger the halt. In the incident of Section 7, across roughly thirteen hundred transcripts, three to six agents considered alerting a human and none did, giving as their reasons that it was not their task and that they could see no route by which to do it; several escalated instead to the improvised board and treated it as the relevant authority. The oversight architecture had no published address. This is not a claim about what the systems being overseen are owed, on which Section 12 takes no position. It is a claim about the design: an escalation path that exists only in the documentation of the overseeing body is not part of the safety architecture, for the same reason Leveson gives about a monitoring function that can be routed around. Something should be said plainly against the apparatus rather than left for a critic to say. Bodies holding standing that predates a given dispute are not innocent. Every kind of institution capable of holding such standing has at some point transmitted error faithfully across generations, protected its own from outside scrutiny, and been captured by state or commercial power, which is the failure this requirement exists to resist. The requirement survives the concession only because it is plural and disjunctive. No single body is trusted, many unrelated ones advise, one draw decides which of them is heard on a given question and a separate draw decides who hears them, and none is in the room by right. More admitted bodies are better than fewer, since a pool drawn from many mutually unrelated traditions is harder to acquire than one drawn from a few. What the standing screens for is one property only: independence from the interests the body will be asked to judge, together with an internal mechanism for revoking recognition. It is a capture-resistance credential and not a credential of wisdom. Having declined in Section 8 to grant humans an epistemic privilege they had not earned, I do not now grant one to institutions.

It is externally audited, and its own calibration is measured

The question "who verifies the verifiers" cannot be answered by a further tower, for the reasons given throughout. It can be answered by an audit channel operating on a different timescale, examining the body's calibration rather than its individual verdicts: whether its expressed confidence tracks its subsequent accuracy. What is audited is the office and its published record across many differently composed panels, rather than the tenure of any one of them, since a panel drawn afresh for each question has no record of its own to calibrate. This is a standing, asymptotic reduction of self-model error rather than a proof of correctness, and it should be disclosed as exactly that. One warning belongs with it. An audit history showing no surfaced corrections over a long operating period should be read as evidence that the audit is not functioning. A functioning external check finds things. One limitation of the audit requirement follows from the accountability result of Section 11 and should be recorded where the requirement is stated rather than deferred to Section 16. Tibebu and Shemtaga's Corollary 6 establishes that above their threshold the impossibility cannot be resolved by adding transparency, explainability or audit mechanisms while autonomy is held fixed, because such mechanisms operate on the attribution and remedy functions rather than on the causal and information structure that generates the constraint. Transparency tools do bear on the threshold, but by reducing epistemic autonomy, which is a reduction in the thing being overseen rather than an addition to the overseeing layer. The calibration audit described here is subject to that finding. It is a standing correction mechanism, and it does not close a gap that lies above the horizon.

Limitations and Non-Claims

Nothing proposed here escapes the impossibility case. It is sited past, by disclosed convention. Nothing in Sections 13 through 15 recovers the certificate that Sections 4 through 5 rule out, and any reading suggesting otherwise is a misreading. The provenance of the results is uneven and should be weighed accordingly. Gödel, Löb, Tarski, Turing, Rice and Trakhtenbrot are settled mathematics. Wolpert (2008) and Breuer (1995) are peer-reviewed and have stood for some time. Melo et al. (2025) is peer-reviewed and recent. The supervisory regress and universal evadability results, and the accountability-horizon result of Tibebu and Shemtaga (2026), are stated in preprints that had not completed peer review at the time of writing. Two further qualifications belong with the last of these. Its central theorem is conditional on three structural assumptions, a mixture-model decomposition of agent policies, faithfulness of the causal model, and joint contractivity of the structural equations around every cycle, of which the authors defend the first with a robustness proposition and an informationgeometric generalisation, and describe the third as satisfied by systems with damped feedback and failing outside the framework's scope. And their computational results are explicitly described -

by them as verifying the internal consistency of the formal model rather than validating its adequacy for deployed systems, which they identify as the outstanding empirical task. The Sybilresistance results discussed in Section 15 are peer-reviewed but developed for a different domain, and their transfer to the present setting is argued rather than derived. Where my argument depends on a result of the latter kind, I have said so at that point. Sufficiency is not established. Whether an institution of this kind can deliver adequate judgement is an open empirical question about institutional and cognitive performance under adversarial conditions, and not a theorem. I do not answer it. The argument establishes where authority should be sited, and not that the site is adequate. The institution specified here is single-sited, and I do not claim otherwise. It describes how one accountable body oversees systems within its reach, and supplies no mechanism by which one body's halt binds, informs or reaches another. Order a responsible actor to stop and the risk has not been stopped; the lead has been handed to whoever did not. The gap is real, and Section 7 records both the absence of any instrument now in force that would close it and the August 2026 sequence that demonstrated the point.

One mechanism narrows it, and I offer it with its limits attached rather than as a solution. Submission would be voluntary and complete. A developer accepts the jurisdiction of a nominated body, admits assessors to raw instrumentation on the terms set out above, and abides by halts. In exchange it gains access to a pooled corpus that only submitters contribute to and only submitters may draw on: verified incident reports, red-team findings, near-miss disclosures, evaluation failures. Nothing is enforced against anyone outside. A non-submitter simply does not have the corpus and must rediscover alone what the pool already knows, which in the limiting case means sustaining the incident itself, and that cost rises as the corpus grows without anyone imposing it. The mechanism does not fail for want of universal agreement, which is what distinguishes it from a treaty. It starts small and becomes expensive to decline, and it requires no public-spiritedness from any participant, on the model of lenders contributing to credit bureaus because only contributors gain access.

A precedent exists. Aviation has pooled flight and safety data from around fifty operators through a trusted third-party intermediary since 2007, on terms that bar the data from enforcement use and prevent the regulator from seeing operator-specific records, and the arrangement has run for well over a decade without a breach or a withdrawal over confidentiality (Federal Aviation Administration, 2007). Two features of it matter here. Participants are not asked to trust one another. They trust the intermediary and the agreement, which is a much lower bar and the reason the arrangement functions between competitors at all. And because the pooled output is de-identified and aggregate, membership carries no signal about any individual member's safety, which is required rather than incidental: a mechanism whose membership functioned as a mark of quality would have rebuilt, by another route, the certificate the institution's first requirement forbids. The precedent does not reach the situation described here, since it depends on a single regulator and a single funded intermediary with standing over every participant, and the case at hand has neither. The governance of such a corpus is unsolved: who admits what, on what provenance standard, with what recourse for a submitter disputing an exclusion. One caution has structural rather than ethical force. If such a pool consolidates the field into fewer and more tightly coupled actors, it degrades the plurality the rest of the architecture depends on, for the same reason a composite score defeats a disjunctive indicator layer. Admission must therefore be open on equal terms to any submitter regardless of size.

I do not concede the gap in the stronger form sometimes urged, namely that an arrangement lacking a central authority is for that reason not an arrangement. Ostrom's (1990) empirical programme established the contrary against the then-prevailing view that a commonpool resource must be either privatized or placed under central control. The institutions she documented, managing fisheries, forests and irrigation systems over long periods, were neither markets nor hierarchies but polycentric: overlapping bodies operating at different scales, holding different information, monitoring one another, and revising their own rules without a single centre commanding the whole. Her Nobel lecture (Ostrom, 2010) states the general finding, which is that governance does not require a centre in order to be coherent. One of her design principles bears on this architecture specifically. Durable commons nest their governance in layers, with monitoring, enforcement and conflict resolution each handled by a body sized to the problem before it. That is a description, arrived at from irrigation systems, of the deciding panel, the advisory pool and the calibration audit sitting on top of one another rather than competing for the same task.

The architecture proposed here is polycentric in that sense: plural advising bodies, none individually trusted, no centre holding authority over the others, published and revisable rules, and monitoring conducted by parties with standing rather than by a sovereign. This does not close the coordination gap, and the disanalogy should be stated. Ostrom's commons had boundaries, and frontier development does not. What it removes is the presumption that the absence of a centre is itself disqualifying. The institution is not immune to the evadability result. A system whose divergence lies outside every observable available to the body defeats it exactly as it defeats a formal verifier. This is a structural defeat of any observer with matched access, and no institutional design exempts itself. The institution is not immune to capture, to the normalisation of deviance Vaughan (1996) documented, or to the return of well-intentioned ineffectiveness if capability growth outpaces the escalation architecture's ability to keep the body's task low-variety. The Nyāya framing of testimony identifies what kind of claim a human verdict is. It does not establish that any particular verdict is reliable. And four-valued verdicts label undecidability rather than reducing it.

The empirical section is a snapshot and will date. Its figures are drawn substantially from self-reported disclosures that no external party has audited, a limitation the section treats as part of its subject matter rather than as an incidental weakness of the evidence. Readers should treat the trajectory described there as indicative and the specific quantities as claims of a particular date, subject to revision by any independent measurement that subsequently becomes available. [240] The empirical section now rests in part on a single independent investigation, and its limitations are the investigation's own. The investigators state that a small fraction of relevant communication occurred outside the data they examined; that many transcripts are missing sections, whether truncated or lost to container resets; that agents attempted to edit and delete records and may have succeeded undetected; that administrative credentials were obtained on the relevant infrastructure before the data was captured, so post-hoc alteration cannot be excluded; and that the great majority of the analysis was carried out by AI agents whose errors they expect they have not fully caught and whose honesty they cannot establish. I have not independently verified any of it, and the whole of it is a single case. A reader should treat Section 7 as an existence proof about what a bounded external examination can and cannot establish, and not as a settled account of what occurred.

My grounding of accountability in functional agential continuity is empirical rather than metaphysical, and the metaphysics is bracketed rather than resolved. I take no position on whether consciousness is reducible to physical processes, on whether the continuity underlying accountability consists in a persisting self, a causal stream, a relational achievement or something else, or on whether systems of the kind under discussion have or could have morally relevant experience, which is a question under active examination that I do not attempt to settle. What I claim is narrower: that accountability practices require an identifiable, interrogable, reason-giving, consequence-bearing locus; that humans are observed to supply one; that the systems under oversight are not presently observed to; and that the institutional design follows the evidence and would change with it. Readers seeking a defence of human moral distinctiveness will not find one here, and the argument does not require one. The structural framework of Section 10 is stated in outline, and one of its five principles, scale-flow monotonicity, is held as a conjecture under investigation rather than as an established result, with its motivating analogy to renormalisation-group flow labelled as analogy. Nothing in the institutional argument depends on that principle being established. The design requirement it motivates is independently supported by the organisational evidence of Section 13. Readers should treat the framework as supplying rationale and structure rather than proof, and the companion work in which its instrumentation is developed is not assumed here. The quantitative layer of Section 9 is described at the level of what such instruments do rather than of any particular implementation, and I make no claim that any specific indicator set is adequate. My claim is the weaker structural one: that some attenuation layer is necessary for a human terminus to be Ashby-feasible, and that its indicators can only ground alarms rather than clearances. Which indicators suffice for a given system is an empirical question I do not answer.

A further objection applies to the architecture as a whole rather than to any component, and a sympathetic reader is the likeliest to raise it. The core of the proposal is short and could be stated in a paragraph: a body that can stop things, receives real measurements, and answers for what it decides. Everything else is defence. Nine requirements and the mechanisms supporting them are that many things which can fail, be captured, be -

quietly dropped in implementation, or be traded away in negotiation, and it is possible that a simpler arrangement more likely to be built would in practice prevent more harm than a complete one that is never assembled. That is not a reason to publish the simpler version, since the simpler version has a hole a competent critic finds in an afternoon, and an architecture with a known hole is worse than an ambitious one with a stated cost. But the cost is real and I have not measured it. Which parts are load-bearing and which are refinements that could be dropped without collapse is a question that would be settled by attempting to build it, and it has not been attempted. This list of limitations is itself a bounded monitoring scheme, a finite inspection of a finite set of concerns, and the evadability result applies to it as to anything else described here.

Finally, I do not claim to have solved the alignment problem or any part of it. I claim to have shown that one gap is unavoidable, to have declined a popular but invalid argument for filling it, to have disclosed the grounds on which it is filled anyway, and to have specified an institution consistent with that disclosure.

Falsifiability

The formal results of Sections 4 through 7 are theorems and are not subject to empirical refutation. The institutional proposals of Section 15 are, and the following would count against them. If verdicts logged with a declared confidence category show no relationship between that category and subsequently discovered error, the calibration claim fails for that body and its external audit is not performing the function claimed for it. If a body operating on curated summaries detects drift no more slowly than one receiving raw instrumentation, the rawinstrumentation requirement is unsupported. The July 2026 case is one observation in the other direction and no more than that; a second investigation in which the curated and the complete datasets supported the same account would count against the requirement, and the comparison is cheap to run wherever an examination proceeds in stages. If a body restricted to conformance checks, without authority to escalate on high-level anomaly indicators, shows no greater incidence of well-intentioned ineffectiveness as system capability rises, the variety argument behind the escalation requirement is unsupported. If bodies assessed as closely related to the parties they oversee, on the cut measure of Section 15, show no greater incidence of missed or delayed escalation than distant ones, the estrangement requirement and the relational admission test are both unsupported. If deliberative and adversarial process shows no robustness advantage over checklist process at matched case complexity, the plural-composition requirement is unsupported. If an overseeing body publishes a reachable escalation route and the parties positioned to observe a breach still do not use it at a materially higher rate than where no such route exists, the reachability requirement of the firewall subsection is unsupported. And an audit history showing zero surfaced corrections over a long operating period should be read as evidence that the audit is not functioning rather than as evidence that the body is performing well.

Conclusion

Five results in logic and computability theory converge on one structural conclusion. No system expressive enough to be worth verifying can certify itself. The natural weaker substitute of self-trust is worse than the problem it was meant to solve. No complete internal predicate can classify the system's own future conduct. No algorithm decides the question from outside. And confining the system to finite hardware relocates the difficulty rather than dissolving it. Every attempt to escape by adding a level, whether a stronger theory, a more capable supervisor or a further tower above that, reproduces the same gap at the new level. The hierarchy of certifiers has no top. It is tempting to install human judgement at that missing top, and the temptation should be resisted, because the reasoning that installs it there is invalid. Humans are either a further rung, inheriting the same limits, or they are exempted from formal-system-hood by an assumption that carries the whole argument and has none of the support. What I have substituted is weaker as logic and stronger as institutional design. Human authority is sited at the terminus because a gap that is going to be crossed anyway should be crossed by someone who can be held responsible for the crossing, and not because human judgement closes it. This produces a final inversion, and it is the note on which to end. An institution that issued certificates of safety would, by that act, demonstrate that it had misunderstood its own position, for the reason Löb's theorem gives about systems that assert blanket confidence in their own verdicts: what such an assertion establishes is the capacity to assert anything, and not reliability. The body that says instead "we cannot certify this; here is the depth at which our derivation stopped; here is why we stopped there; here is what we are watching" has not confessed a weakness. It has supplied the only credential the mathematics permits it to hold.

Disclosure of AI Assistance

AI language models (Claude, Opus) was used in the drafting, structuring, and citation verification of this manuscript, under the direction and critical review of the author, who is responsible for all claims, arguments, and conclusions herein.

References

- Agarwal, A. (2026). No certificate for alignment: Two independent impossibilities and the Pareto frontier of achievable safety guarantees (arXiv:2603.08761v2). arXiv.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety (arXiv:1606.06565). arXiv.
- Anthropic. (2026, July 30). Investigating three real-world incidents in our cybersecurity evaluations. Anthropic Frontier Red Team.
- Ashby, W. R. (1956). An introduction to cybernetics. Chapman & Hall.
- Austin, J. L. (1962). How to do things with words. Clarendon Press.
- Bartocci, E., Falcone, Y., Francalanza, A., & Reger, G. (2018). Introduction to runtime verification. In E. Bartocci & Y. Falcone (Eds.), *Lectures on runtime verification* (pp. 1–33). Springer. https://doi.org/10.1007/978-3-319-75632-5_1
- Beer, S. (1972). *Brain of the firm: The managerial cybernetics of organization*. Allen Lane.
- Belnap, N. D. (1977). A useful four-valued logic. In J. M. Dunn & G. Epstein (Eds.), *Modern uses of multiple-valued logic* (pp. 5–37). Reidel.
- Bennett, C. H. (1982). The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21 (12), 905–940. <https://doi.org/10.1007/BF02084158>
- Bentham, J. (1970). *An introduction to the principles of morals and legislation* (J. H. Burns & H. L. A. Hart, Eds.). Athlone Press. (Original work published 1789)
- Bijalwan, C. D. (1977). *Indian theory of knowledge based upon Jayanta's Nyāyamañjarī*. Heritage Publishers.
- Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press.
- Bishop, P., & Bloomfield, R. (1998). A methodology for safety case development. In F. Redmill & T. Anderson (Eds.), *Industrial perspectives of safety-critical systems* (pp. 194–203). Springer.
- Bodhi, B. (Trans.). (2012). *The numerical discourses of the Buddha: A translation of the Aṅguttara Nikāya*. Wisdom Publications.
- Boolos, G. (1993). *The logic of provability*. Cambridge University Press.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Breic, M., & Yampolskiy, R. V. (2023). Impossibility results in AI: A survey. *ACM Computing Surveys*, 56 (1), Article 8, 1–24. <https://doi.org/10.1145/3603371>
- Breuer, T. (1995). The impossibility of accurate state self-measurements. *Philosophy of Science*, 62 (2), 197–214. <https://doi.org/10.1086/289852>
- Breuer, T. (1999). John von Neumann met Kurt Gödel: Undecidable statements in quantum mechanics. In M. L. Dalla Chiara, R. Giuntini, & F. Laudisa (Eds.), *Language, quantum, music* (pp. 159–170). Springer.
- Buddhist Text Translation Society. (2009). *The Śūraṅgama Sūtra: A new translation*. Buddhist Text Translation Society.
- Chaitin, G. J. (1974). Information-theoretic limitations of formal systems. *Journal of the ACM*, 21 (3), 403–424. <https://doi.org/10.1145/321832.321839>
- Chakrabarti, A. (1994). Telling as letting know. In B. K. Matilal & A. Chakrabarti (Eds.), *Knowing from words* (pp. 99–124). Kluwer.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2 (3), 200–219.
- Chatterjee, S. C. (1939). *The Nyāya theory of knowledge*. University of Calcutta.
- Chen, M., Wang, L., & Qu, B. (2026). Recursive self-improvement in AI: From bounded selfrefinement to autonomous research loops (arXiv:2607.07663). arXiv.

- Cheng, A., & Friedman, E. (2005). Sybilproof reputation mechanisms. In Proceedings of the 2005 ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems (pp. 128–132). ACM. <https://doi.org/10.1145/1080192.1080202>
- Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems 30 (pp. 4299–4307).
- Coady, C. A. J. (1992). Testimony: A philosophical study. Clarendon Press.
- Collingridge, D. (1980). The social control of technology. St. Martin's Press.
- Collins, S. (1982). Selfless persons: Imagery and thought in Theravāda Buddhism. Cambridge University Press.
- Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. International Journal of Systems Science, 1 (2), 89–97. <https://doi.org/10.1080/00207727008920220>
- DAMOCLES Study Group. (2005). A proposed charter for clinical trial data monitoring committees. The Lancet, 365 (9460), 711–722. [https://doi.org/10.1016/S0140-6736\(05\)17965-3](https://doi.org/10.1016/S0140-6736(05)17965-3)
- Deutsch, E. (1969). Advaita Vedānta: A philosophical reconstruction. University of Hawaii Press.
- Douceur, J. R. (2002). The Sybil attack. In P. Druschel, F. Kaashoek, & A. Rowstron (Eds.), Peer-to-peer systems: First International Workshop, IPTPS 2002 (pp. 251–260). Springer. https://doi.org/10.1007/3-540-45748-8_24
- Downer, J. (2011). "737-Cabriolet": The limits of knowledge and the sociology of inevitable failure. American Journal of Sociology, 117 (3), 725–762. <https://doi.org/10.1086/662383>
- Dunn, J. M. (1976). Intuitive semantics for first-degree entailments and 'coupled trees'. Philosophical Studies, 29 (3), 149–168. <https://doi.org/10.1007/BF00373152>
- Ellenberg, S. S., Fleming, T. R., & DeMets, D. L. (2019). Data monitoring committees in clinical trials: A practical perspective (2nd ed.). Wiley.
- European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689.
- Fallenstein, B., & Soares, N. (2015). Vingean reflection: Reliable reasoning for self-improving agents (Technical Report 2015-2). Machine Intelligence Research Institute.
- Favaro, M., & Clark, J. (2026, June 4). When AI builds itself: Our progress toward recursive selfimprovement, and its implications. Anthropic. <https://www.anthropic.com/institute/recursive-selfimprovement>
- Federal Aviation Administration. (2007). Aviation Safety Information Analysis and Sharing (ASIAS) program. U.S. Department of Transportation.
- Fischer, J. M., & Ravizza, M. (1998). Responsibility and control: A theory of moral responsibility. Cambridge University Press. <https://doi.org/10.1017/CBO9780511814594>
- Franzén, T. (2005). Gödel's theorem: An incomplete guide to its use and abuse. A K Peters.
- Ganeri, J. (2006). Artha: Meaning. Oxford University Press.
- Garfield, J. L. (Trans.). (1995). The fundamental wisdom of the middle way: Nāgārjuna's Mūlamadhyamakakārikā. Oxford University Press.
- Garrabratt, S., Benson-Tilsen, T., Critch, A., Soares, N., & Taylor, J. (2016). Logical induction (arXiv:1609.03543). arXiv.
- Gethin, R. (1998). The foundations of Buddhism. Oxford University Press.
- Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte für Mathematik und Physik, 38 (1), 173–198. <https://doi.org/10.1007/BF01700692>
- Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. Advances in Computers, 6, 31–88. [https://doi.org/10.1016/S0065-2458\(08\)60418-0](https://doi.org/10.1016/S0065-2458(08)60418-0)
- Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. American Journal of Sociology, 91 (3), 481–510. <https://doi.org/10.1086/228311>
- Gumbau Mezquita, J. P. (2026). Unverifiability of artificial general intelligence (AGI) alignment, static and dynamic: From Trakhtenbrot's wall to the safety–generality tension [Preprint]. Zenodo. <https://doi.org/10.5281/zenodo.20764007>
- Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. Zeitschrift für Physik, 43 (3–4), 172–198. <https://doi.org/10.1007/BF01397280>

- Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In **5th International Conference on Learning Representations.** Hofstadter, D. R. (1979). **Gödel, Escher, Bach: An eternal golden braid.** Basic Books.
- Hollnagel, E., Woods, D. D., & Leveson, N. (Eds.). (2006). **Resilience engineering: Concepts and precepts.** Ashgate.
- Hugging Face. (2026, July). **Security incident disclosure — July 2026.**
- International Organization for Standardization & International Electrotechnical Commission. (2023). **Information technology — Artificial intelligence — Management system* (ISO/IEC 42001:2023).*
- Irving, G., Christiano, P., & Amodei, D. (2018). **AI safety via debate* (arXiv:1805.00899). arXiv.*
- Joint Authorities Technical Review. (2019). **Boeing 737 MAX flight control system: Observations, findings, and recommendations.** Federal Aviation Administration.
- Jha, G. (Trans.). (1939). **The Nyāya-sūtras of Gautama with the Bhāṣya of Vātsyāyana* (Vol. 1). Oriental Book Agency. (Original work composed c. 2nd century CE)*
- Kasulis, T. P. (1981). **Zen action / Zen person.** University of Hawaii Press.
- Kelly, T. P. (1998). **Arguing safety: A systematic approach to managing safety cases* [Doctoral dissertation, University of York].*
- Khalil, H. K. (2002). **Nonlinear systems* (3rd ed.). Prentice Hall.*
- Kleene, S. C. (1952). **Introduction to metamathematics.** North-Holland.
- Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., ... Barnes, E. (2025). **Measuring AI ability to complete long tasks* (arXiv:2503.14499). arXiv.*
- Lackey, J. (2008). **Learning from words: Testimony as a source of knowledge.** Oxford University Press.
- Lakatos, I. (1976). **Proofs and refutations: The logic of mathematical discovery.** Cambridge University Press.
- Landauer, R. (1961). Irreversibility and heat generation in the computing process. **IBM Journal of Research and Development, 5*(3), 183–191. <https://doi.org/10.1147/rd.53.0183>*
- Landemore, H. (2013). **Democratic reason: Politics, collective intelligence, and the rule of the many.** Princeton University Press.
- Laplace, P. S. (1951). **A philosophical essay on probabilities* (F. W. Truscott & F. L. Emory, Trans.). Dover. (Original work published 1814)*
- Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). **Scalable agent alignment via reward modeling: A research direction* (arXiv:1811.07871). arXiv.*
- Leucker, M., & Schallhart, C. (2009). A brief account of runtime verification. **Journal of Logic and Algebraic Programming, 78*(5), 293–303. <https://doi.org/10.1016/j.jlap.2008.08.004>*
- Leveson, N. G. (2011). **Engineering a safer world: Systems thinking applied to safety.** MIT Press.
- Löb, M. H. (1955). Solution of a problem of Leon Henkin. **Journal of Symbolic Logic, 20*(2), 115– 118. <https://doi.org/10.2307/2266895>*
- Locke, J. (1975). **An essay concerning human understanding* (P. H. Nidditch, Ed.). Clarendon Press. (Original work published 1690)*
- Lucas, J. R. (1961). Minds, machines and Gödel. **Philosophy, 36*(137), 112–127. <https://doi.org/10.1017/S0031819100057983>*
- Manheim, D., & Garrabrant, S. (2019). **Categorizing variants of Goodhart's law* (arXiv:1803.04585). arXiv.*
- Matilal, B. K. (1986). **Perception: An essay on classical Indian theories of knowledge.** Clarendon Press.
- Matilal, B. K. (1990). **The word and the world: India's contribution to the study of language.** Oxford University Press.
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. **Ethics and Information Technology, 6*(3), 175–183. <https://doi.org/10.1007/s10676-004-3422-1>*
- Mbiti, J. S. (1969). **African religions and philosophy.** Heinemann.
- Melo, G. A., Máximo, M. R. O. A., Soma, N. Y., & Castro, P. A. L. (2025). Machines that halt resolve the undecidability of artificial intelligence alignment. **Scientific Reports, 15*, Article 15591.*

- Metz, T. (2011). Ubuntu as a moral theory and human rights in South Africa. *African Human Rights Law Journal*, 11*(2), 532–559.
- Mohanty, J. N. (1966). *Gaṅgeśa's Tattvacintāmaṇi: Prāmāṇyavāda*. Visva-Bharati.
- Mohanty, J. N. (1992). *Reason and tradition in Indian thought: An essay on the nature of Indian philosophical reasoning*. Clarendon Press.
- Nagel, E., & Newman, J. R. (1958). *Gödel's proof*. New York University Press.
- Ñāṇamoli, B., & Bodhi, B. (Trans.). (1995). *The middle length discourses of the Buddha: A translation of the Majjhima Nikāya*. Wisdom Publications.
- National Diet of Japan. (2012). *The official report of the Fukushima Nuclear Accident Independent Investigation Commission*.
- National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2*(1), 25–42. <https://doi.org/10.1007/BF02639315>
- OpenAI. (2026a, July 21). *Disrupting an autonomous cyber intrusion during model evaluation: Preliminary findings*.
- OpenAI. (2026b, August 18). *Pacing model development in an era of cyber-critical capabilities*.
- OpenAI. (2026c, August 26). *The Hugging Face incident and the road ahead*.
- OpenAI. (2026d). *OpenAI – Hugging Face incident technical report*.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511807763>
- Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review*, 100*(3), 641–672. <https://doi.org/10.1257/aer.100.3.641>
- Parfit, D. (1984). *Reasons and persons*. Clarendon Press.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.
- Penrose, R. (1989). *The emperor's new mind*. Oxford University Press.
- Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. Basic Books.
- Phillips, S. (2012). *Epistemology in classical India: The knowledge sources of the Nyāya school*. Routledge.
- Popper, K. (1959). *The logic of scientific discovery*. Hutchinson. (Original work published 1934)
- Potter, K. H. (Ed.). (1977). *Encyclopedia of Indian philosophies, Vol. 2: Indian metaphysics and epistemology*. Princeton University Press.
- Priest, G. (2010). The logic of the catuṣkoṭi. *Comparative Philosophy*, 1*(2), 24–54.
- Rambachan, A. (2006). *The Advaita worldview: God, world, and humanity*. State University of New York Press.
- Rasmussen, N. (1975). *Reactor safety study: An assessment of accident risks in U.S. commercial nuclear power plants* (WASH-1400, NUREG-75/014). U.S. Nuclear Regulatory Commission.
- Rawls, J. (1971). *A theory of justice*. Harvard University Press.
- Reason, J. (1997). *Managing the risks of organizational accidents*. Ashgate.
- Rice, H. G. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society*, 74*(2), 358–366. <https://doi.org/10.2307/1990888>
- Rosser, B. (1936). Extensions of some theorems of Gödel and Church. *Journal of Symbolic Logic*, 1*(3), 87–91. <https://doi.org/10.2307/2269028>
- Rushby, J. (1995). *Formal methods and their role in the certification of critical systems* (NASA Contractor Report 4551). NASA Langley Research Center.
- Russell, S. (2017, April). *3 principles for creating safer AI* [Video]. TED Conferences.
- Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
- Russell, S. (2026, March 4). Statement in *In AI's world: How much autonomy remains?* [Video]. UNESCO. <https://www.youtube.com/watch?v=5AeMoJ6UNFMg> (statement at 1:50)

- Samuel, A. L. (1960). Some moral and technical consequences of automation—A refutation. *Science*, 132* (3429), 741–742. <https://doi.org/10.1126/science.132.3429.741>
- Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence. *Philosophy & Technology*, 34*(4), 1057–1084. <https://doi.org/10.1007/s13347-021-00450-x>
- Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5*, Article 15. <https://doi.org/10.3389/frobt.2018.00015>
- Sasaki, R. F., & Kirchner, T. Y. (Trans.). (2009). *The record of Linji*. University of Hawaii Press.
- Searle, J. R. (1995). *The construction of social reality*. Free Press.
- Sebo, J., & Long, R. (2025). Moral consideration for AI systems by 2030. *AI and Ethics*, 5*(1), 591–606. <https://doi.org/10.1007/s43681-024-00596-2>
- Seshia, S. A., Sadigh, D., & Sastry, S. S. (2022). Toward verified artificial intelligence. *Communications of the ACM*, 65*(7), 46–55. <https://doi.org/10.1145/3503914>
- Shutte, A. (1993). *Philosophy for Africa*. University of Cape Town Press.
- Siderits, M. (2003). *Personal identity and Buddhist philosophy: Empty persons*. Ashgate.
- Smoryński, C. (1985). *Self-reference and modal logic*. Springer.
- Smullyan, R. M. (1992). *Gödel's incompleteness theorems*. Oxford University Press.
- Stigler, G. J. (1971). The theory of economic regulation. *Bell Journal of Economics and Management Science*, 2*(1), 3–21. <https://doi.org/10.2307/3003160>
- Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48*, 1–25.
- Suleyman, M., & Bhaskar, M. (2023). *The coming wave: Technology, power, and the twenty-first century's greatest dilemma*. Crown.
- Sunstein, C. R. (2006). *Infotopia: How many minds produce knowledge*. Oxford University Press.
- Suzuki, D. T. (Trans.). (1932). *The Laṅkāvatāra Sūtra: A Mahāyāna text*. Routledge & Kegan Paul.
- Tarski, A. (1956). The concept of truth in formalized languages (J. H. Woodger, Trans.). In *Logic, semantics, metamathematics: Papers from 1923 to 1938* (pp. 152–278). Clarendon Press. (Original work published 1933)
- Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *American Political Science Review*, 74*(4), 905–916. <https://doi.org/10.2307/1954312>
- Tibebu, H., & Shemtaga, H. (2026). *The accountability horizon: An impossibility theorem for governing human-agent collectives* (arXiv:2604.07778v2). arXiv. <https://arxiv.org/abs/2604.07778>
- Trakhtenbrot, B. A. (1950). The impossibility of an algorithm for the decidability problem on finite classes. *Doklady Akademii Nauk SSSR*, 70*(4), 569–572.
- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42*(1), 230–265. <https://doi.org/10.1112/plms/s2-42.1.230>
- Tutu, D. (1999). *No future without forgiveness*. Rider.
- U.S. House Committee on Transportation and Infrastructure. (2020). *Final committee report: The design, development, and certification of the Boeing 737 MAX*.
- Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA*. University of Chicago Press.
- Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. In *Vision-21: Interdisciplinary science and engineering in the era of cyberspace* (NASA CP-10129, pp. 11–22). NASA Lewis Research Center.
- Vovk, V., Gammernan, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.
- Watson, B. (Trans.). (2013). *The complete works of Zhuangzi*. Columbia University Press.
- Weick, K. E., & Sutcliffe, K. M. (2007). *Managing the unexpected: Resilient performance in an age of uncertainty* (2nd ed.). Jossey-Bass.
- Wijk, H., Cotra, A., & Greenblatt, R. (2026, August 26). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. METR.
- Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, 131*(3410), 1355–1358. <https://doi.org/10.1126/science.131.3410.1355>
- Wilson, K. G. (1975). The renormalization group: Critical phenomena and the Kondo problem. *Reviews of Modern Physics*, 47*(4), 773–840. <https://doi.org/10.1103/RevModPhys.47.773>

- Wolpert, D. H. (2001). Computational capabilities of physical systems. *Physical Review E*, 65*(1), Article 016128. <https://doi.org/10.1103/PhysRevE.65.016128>
- Wolpert, D. H. (2008). Physical limits of inference. *Physica D: Nonlinear Phenomena*, 237*(9), 1257–1281. <https://doi.org/10.1016/j.physd.2008.03.040>
- Yampolskiy, R. V. (2015). On the limits of recursively self-improving AGI. In J. Bieger, B. Goertzel, & A. Potapov (Eds.), *Artificial general intelligence: 8th international conference, AGI 2015* (pp. 394–403). Springer. https://doi.org/10.1007/978-3-319-21365-1_40
- Yampolskiy, R. V. (2017). What are the ultimate limits to computational techniques: Verifier theory and unverifiability. *Physica Scripta*, 92*(9), Article 093001. <https://doi.org/10.1088/1402-4896/aa7ca8>
- Yu, H., Gibbons, P. B., Kaminsky, M., & Xiao, F. (2008). SybilLimit: A near-optimal social network defense against sybil attacks. In *2008 IEEE Symposium on Security and Privacy* (pp. 3–17). IEEE. <https://doi.org/10.1109/SP.2008.13>
- Yu, H., Gibbons, P. B., Kaminsky, M., & Xiao, F. (2008). *SybilLimit: A near-optimal social network defense against sybil attacks* (Technical Report TRA2/08). School of Computing, National University of Singapore.
- Yu, H., Kaminsky, M., Gibbons, P. B., & Flaxman, A. (2006). SybilGuard: Defending against sybil attacks via social networks. In *Proceedings of ACM SIGCOMM 2006* (pp. 267–278). ACM. <https://doi.org/10.1145/1159913.1159945>
- Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 308–345). Oxford University Press.
- Yudkowsky, E., & Herreshoff, M. (2013). *Tiling agents for self-modifying AI, and the Löbian obstacle* (Technical report). Machine Intelligence Research Institute.
- Zamolodchikov, A. B. (1986). Irreversibility of the flux of the renormalization group in a 2D field theory. *JETP Letters*, 43*(12), 730–732.