Restoration Before Explanation

Why Systems Can Be Repaired Before They Are Understood

Flyxion

July 2026

Abstract

Two independent results, developed separately and for different purposes, converge on the same claim without ever being placed side by side. The first, developed within a general theory of recoverable persistence, holds that explanation is a species of distinction repair: a theory succeeds when it increases the recoverability of distinctions that observation alone leaves fragmented, and this repair function is available independently of, and prior to, whatever predictive competence a system already has. The second, developed within a formal architecture for admissibility tracking, distinguishes functional adequacy—a system's behavior correctly tracks a distinction, by whatever internal means—from architectural transparency—an explicit, inspectable object exists whose state can be checked against that distinction without probing the system's internals. The second result shows that a system can clear the first bar while remaining indefinitely short of the second. This paper argues that the two results instantiate the same underlying decoupling, arrived at independently in the vocabulary of repair theory and the vocabulary of architecture theory, and that a diverse body of evidence from reinforcement learning, software infrastructure, biological repair, and transformer behavior confirms the decoupling a third way: functional restoration of a system routinely precedes, and does not require, an explanation of how that restoration works. The paper does not propose a third theory. It shows that two existing theories, apparently unrelated, have already separated restoration from explanation under different names, and it supplies the domain evidence neither one develops.

Contents

1 The Claim

A system can be repaired, kept running, or restored to competent function by an agent who cannot explain why the repair worked. This is not a marginal case, tolerated at the edges of engineering practice while science proper waits for a fuller account before intervening. It is close to the ordinary condition of intervention itself. Mechanics replace parts by pattern-matching against prior failures before automotive engineering can certify a causal account of the specific fault. Physicians set fractures, drain abscesses, and prescribe antibiotics against pathogens whose mechanisms of resistance are only partially characterized. Software engineers patch production systems under time pressure using a fix that resolves the symptom while the root cause remains, in the honest language of the incident report, "not fully understood." In none of these cases does restoration wait for explanation. Restoration happens first, and often happens completely, while explanation—if it arrives at all—arrives later, as a separate achievement with its own separate value.

The claim of this paper is that this ordering is not an artifact of haste, resource constraint, or intellectual laziness, correctable in principle by more careful practice. It reflects a structural fact about what restoration requires versus what explanation requires, and the fact holds independently of domain. It has, in fact, already been arrived at twice, by two theories that do not cite one another, do not share a vocabulary, and were built to solve unrelated problems. One theory, developed within a general account of recoverable persistence, separates prediction from explanation. The other, developed within a formal architecture for tracking admissibility, separates functional adequacy from architectural transparency. The purpose of this paper is to show that these are two namings of one separation, and then to confirm that separation a third way, in evidence neither theory used.

Section 2 states the first of the two prior treatments, in which explanation is characterized as a specific kind of repair operation rather than as a precondition for competent function. Section 3 states the second, in which a system's behavioral adequacy is shown to be logically independent of whether an explicit, inspectable account of that behavior exists anywhere in the system. Section 4 argues that these two results, despite using no shared vocabulary, instantiate the same structural decoupling. Section 5 turns to evidence: reinforcement

learning, software and embedded-systems infrastructure, biological repair, and transformer language models each confirm the claim in a form specific enough to be checked, and none of the four was available to either of the two prior theoretical treatments when they were written. Section 6 closes by stating what follows once restoration and explanation are properly separated, for practice as much as for theory.

2 Explanation as Distinction Repair

A general theory of recoverable persistence (Flyxion 2026a)[1] begins from a simple observation: the objects of reference, measurement, and communication are secured not by perfect preservation across time and transformation, but by *recoverability*—the existence of some procedure capable of reconstructing a distinction after it has been degraded, fragmented, or partially lost. A structured domain, in this framework, consists of a space of possible states, a family of transformations acting on that space, a collection of distinctions defined over it, and a family of reconstruction operators capable of restoring those distinctions after transformation has acted. Repair, within this apparatus, is not an auxiliary process invoked occasionally when something goes wrong. It is constitutive: recoverable persistence cannot exist at all without some repair mechanism opposing the transformations that would otherwise degrade it. A communication channel persists as a channel, rather than degenerating into noise, because an error-correcting code opposes the transformation's tendency to erase distinctions, not because the channel itself ceases to be noisy.

Explanation, on this account, is a special case of repair, applied to a specific target: the fragmented, uncertain, or partially inaccessible distinctions that observation alone leaves behind. An astronomer's observations of an anomalous orbit constitute, prior to any theory, a collection of weakly connected distinctions—data points whose relationship to one another remains largely unrecoverable. A theory of gravitational interaction does not describe these observations from outside; it repairs the distinction structure connecting them, converting a fragmented collection into a reconstructible whole. The same structure recurs wherever explanation is offered: a historical account repairs the distinction structure linking otherwise disconnected archival traces; a genetic theory repairs the distinctions linking otherwise disconnected patterns of inheritance. In every case, the operative question is not whether the explanation is true in some observer-independent sense prior to any repair having occurred, but whether it increases the recoverability of distinctions that were, before the explanation arrived, inaccessible.

This framing yields an immediate and load-bearing consequence, stated explicitly within the theory rather than left implicit: prediction and explanation are related but distinct operations, and neither entails the other. Prediction is successful extrapolation; a system that predicts accurately extends a pattern beyond observed cases without thereby increasing the recoverability of the distinctions underlying that pattern. Explanation is distinction repair; a system that explains successfully increases recoverability, and may do so even where its predictive accuracy remains limited, or conversely may fail to explain while predicting very well indeed. The theory states this decoupling as a general structural fact about the two operations, independent of any specific domain in which either might be instantiated. It does not, however, supply the domain instances that would let a reader check the decoupling against a hard case. Its own worked examples—orbital anomalies,

[1] This paper departs from the ordinary practice, followed elsewhere in this series, of not citing the author's own prior work. The departure is made because this paper's entire contribution depends on naming two specific, identifiable prior results and showing they coincide; describing them without attribution would make the central claim impossible to check against its sources.

historical archives, patterns of inheritance—are drawn from settled science with well-known theories already attached. What remains open is whether the decoupling holds under contemporary pressure: in systems whose restoration is achieved by optimization rather than insight, where the very possibility of an explanation lagging indefinitely behind functional competence is not a historical curiosity but the live, current condition of the field. Section 5 supplies that pressure test.

3 Functional Adequacy Without Architectural Transparency

A second and independently motivated line of work (Flyxion 2026b) arrives at a structurally identical decoupling from the opposite direction, starting not from a general theory of persistence but from a formal architecture built to track which reconstructions of a system's history remain admissible as new evidence arrives. Within that architecture, a system maintains an explicit hypothesis set over admissibility structures, repaired under a discipline that only removes candidates consistent with accumulated evidence and never reintroduces one previously excluded. The apparatus is built, deliberately, to keep two properties visible as separate rather than allowing them to be run together, as they typically are in informal discussion of whether a system "really understands" what it is doing.

Definition 1 (Functional adequacy)**.** *A system is functionally adequate with respect to a distinction if its downstream behavior tracks the true state of that distinction at better than chance, under a controlled audit, regardless of what internal mechanism produces the tracking.*

Definition 2 (Architectural transparency)**.** *A system is architecturally transparent with respect to a distinction if, beyond functional adequacy, it maintains an explicit, inspectable object whose state can be compared against the distinction directly, without requiring intervention on the system's internals to locate it.*

The two properties are independent by construction, and the independence is not a merely formal possibility reserved for pathological cases. A system can be functionally adequate—passing every behavioral test that would ordinarily be offered as evidence of competence—while remaining architecturally opaque, with no component anyone has identified playing the role that an explicit hypothesis set plays in a system built to be transparent by design. The converse also holds: a system can be built to be architecturally transparent and still fail functional adequacy, if the explicit machinery it exposes happens to be wrong. Settling one property does not settle the other, and behavioral evidence—the kind of evidence ordinarily available, and often the only kind available—bears directly on the first property while remaining silent on the second.

This yields a further distinction worth stating on its own terms, because it separates two ways functional adequacy without transparency can arise, and the two are not evidentially equivalent. A system's behavior might track a distinction because something functionally isomorphic to an explicit representation of that distinction has been constructed internally and is being consulted, even though no one has yet located it—a case of transparency that happens to be undiscovered rather than absent. Or the tracking might be produced by a distributed approximation in which no localized component carries the relevant information at all, so that no amount of further search would locate a component to find, because there is none of the kind being searched for. A controlled audit—holding a probe fixed across contexts that supply different disambiguating evidence, then checking whether downstream answers track the evidence—establishes functional adequacy without needing to resolve which of these two cases obtains. A further step, attempting to localize an intervention point whose value is injective over the distinction independent of final output, is required

to establish transparency, and a failure to localize such a point is evidence only that it has not been found, which is a strictly weaker claim than evidence that it does not exist.

4 The Same Move, Twice

The two results share no vocabulary. One speaks of distinctions, transformations, and reconstruction operators; the other speaks of hypothesis sets, admissibility structures, and audit protocols. Read side by side, however, they express the same underlying decoupling under different names. Functional adequacy, in the vocabulary of Section 3, corresponds to restoration: a system whose behavior correctly recovers or maintains a distinction is a system that has been, in the relevant functional sense, kept working. Architectural transparency corresponds to what Section 2 calls explanation: an explicit, inspectable structure whose relationship to the distinction it tracks can be checked directly, which is what distinction repair supplies when it succeeds. That the two properties are independent—that a system can be functionally adequate without being architecturally transparent—is, under this correspondence, the same separation as the claim that prediction and explanation are related but distinct operations, neither entailing the other. Restoration corresponds to functional adequacy and to successful prediction; explanation corresponds to architectural transparency and to successful distinction repair. The two theories are not identical, and were not built to be compared; what they share is that both, independently, found that the first member of their respective pairs can occur fully in the absence of the second.

Neither theory, on its own, made this identification explicit, because neither needed to: each was solving a local problem within its own vocabulary, and the identification only becomes visible once the two are read together. Making it explicit is this paper's contribution at the level of theory, and it is a modest one—a naming, not a new derivation. The larger contribution is empirical. If restoration-without-explanation is a single structural fact rather than two coincidentally similar ones, it should be checkable in systems that neither theory used as source material, under conditions specific enough that the check could fail. Section 5 runs that check across four domains.

5 Four Cases

5.1 Reinforcement Learning: Repair Without a Located Mechanism

Recent work training large language models with reinforcement learning has shown that updating a single transformer layer, chosen from the middle of the network's depth, can match or exceed the performance gained by updating every parameter in the model, a result stable across seven base models spanning two model families, three reinforcement-learning algorithms, and multiple task families (Zhang et al. 2026). The result is a clean instance of restoration without explanation as it is defined in this paper: the model's behavior is functionally restored—its task performance measurably improves, by an amount comparable to updating the entire parameter set—through a narrow, depth-specific set of update directions, while no located mechanism accounts for why availability concentrates at intermediate depth rather than elsewhere. Candidate explanations exist. Literature on residual computation suggests that transformer layers perform a form of iterative refinement, with each layer nudging a running estimate rather than computing an independent function from scratch (Jastrzębski et al. 2018), and separate work on reading intermediate representations through the model's own output vocabulary shows that these running estimates evolve progressively and coherently across depth (Belrose et al. 2023). Neither line of work, however,

predicts middle-depth concentration of reinforcement-learning plasticity specifically; both establish only that representations refine progressively, which is a necessary background fact for any explanation of the middle-depth result but not itself an explanation of it. The restoration—measurable, replicated, exploitable in practice by training a single layer instead of the whole model—is complete and available now. The explanation, if it arrives, arrives as a separate achievement, on a separate timeline, possibly never.

5.2 Infrastructure: Systems Built to Outlast Their Own Documentation

Embedded and infrastructure engineering supplies a second case, and it differs from the reinforcement-learning instance in an instructive way: here, restoration without explanation is not merely tolerated but is sometimes the explicit design target. Software systems designed for operation under conditions of degraded institutional support—after the toolchains, documentation, and specialist knowledge that originally produced them have become unavailable—are built around the premise that a system must remain repairable by an operator who cannot reconstruct, and does not need to reconstruct, the reasoning behind its original design (Ziembik 2026). Minimizing a system's internal complexity is valued in this design tradition specifically because it enlarges the population of operators capable of repairing the system without first explaining it, trading the depth of understanding a complex, well-documented system might in principle support for the breadth of restorability a simple one actually delivers under degraded conditions. The same premise appears starkly in systems engineered never to be touched again after deployment: spacecraft instrumentation, once in flight, cannot be explained to or by anyone in real time, and its designers accept this in advance, front-loading whatever restoration capacity the system will ever have into its initial construction because no further explanation-mediated intervention will be available. In both cases, restorability is treated as the primary design variable and explicability of any subsequent repair as, at best, a secondary and frequently absent one.

5.3 Biological Repair: Function Recovered, Mechanism Contested

Biological systems restore function continuously, across every scale from the molecular to the organismal, and clinical medicine has a documented history of restoring function well in advance of a settled mechanistic account of how the restoration was achieved. Organ transplantation is the clearest case on record. Joseph Murray's 1954 kidney transplant between identical twins succeeded, and kept its recipient alive for eight years, at a time when the immunological mechanisms governing graft rejection were understood only in outline; a functioning organ was restored to a failing body before the science of why a non-identical graft would instead be destroyed had been worked out (Murray 1990). The gap did not close quickly. Alexis Carrel had already, decades earlier, developed the vascular suturing technique that made any organ transplant technically possible at all, and had observed that autografts and allografts survived for different lengths of time in experimental animals, without arriving at a concept of immunological rejection as distinct from other causes of graft failure (Murray 1990). Restoration of function—first surgical, later immunological—was achieved and clinically deployed at each stage before the corresponding mechanism was understood, and the human leukocyte antigen system that eventually explained non-twin rejection was not identified until 1958, four years after Murray's first success (Dash, Nair, and Behera 2023). This is not a claim that mechanistic biology is unimportant, or that explanation carries no value once restoration has occurred—later sections return to what explanation adds once it does arrive—only that the field's own history confirms, at the scale of a discipline rather than a single result, that functional restoration is not gated on prior

explanation, and can precede it by years or decades.

5.4 Transformers: Applying the Audit

The clearest illustration of Section 3's distinction in its own home domain is the transformer audit protocol it makes available but does not itself carry out. A language model's handling of a syntactically ambiguous construction—one compatible with more than one underlying reading, disambiguated only by context supplied earlier in the same interaction—can be tested for functional adequacy directly: hold the ambiguous construction and the downstream question about it fixed, vary only the disambiguating context, and check whether the model's answer tracks the intended reading at better than chance across many such pairs. This test requires no access to the model's internals and settles nothing about architectural transparency; a model that passes it convincingly has been shown only to be functionally adequate over that construction. Testing transparency requires the further, harder step of searching for a localized intervention point—a direction in the model's internal activations, or a specific subset of its attention computation—whose value tracks the intended reading independently of the model's final output. Success at this second step would be evidence that an explicit, inspectable representation of the distinction exists somewhere in the computation; failure is evidence only that the search did not find one, which does not establish that none exists. The practical situation of contemporary language models is that the first test is run constantly, informally, every time a model is asked to handle an ambiguous case correctly, while the second test is run rarely, formally, and inconclusively. Restoration of correct behavior is, in the ordinary case, complete. Explanation of that behavior, in the architectural-transparency sense, remains for the most part outstanding.

6 What Follows

If restoration and explanation are genuinely separable—derived twice from independent theoretical starting points, and confirmed four times across domains none of the underlying theories used as source material—then a habit of practice and a habit of evaluation both need revision. The habit of practice is the expectation, often unstated but nonetheless operative, that intervention should wait on understanding: that a system ought to be explained before it is trusted to be repaired, or that a repair achieved without explanation is provisional in a way that a repair achieved with explanation is not. Nothing in the evidence surveyed here supports treating explained repairs as more secure than unexplained ones simply by virtue of being explained; a repair's reliability is a fact about whether it recurs under the conditions that produced it before, not a fact about whether anyone has supplied an account of why it worked. The habit of evaluation is the tendency, visible across the four domains surveyed, to treat the absence of an explanation as evidence against the adequacy of a restoration, when the two are logically independent properties and evidence for one is, on its own, silent about the other.

None of this diminishes the value of explanation once it does arrive. An explanation that succeeds, in the sense given in Section 2, increases the recoverability of distinctions beyond the specific case that prompted it: it supports prediction in new cases, clarifies which future failures should be anticipated, and converts a restoration that worked once into a restoration that can be reasoned about and deliberately reproduced. What the evidence supports is narrower and more precise than either an endorsement or a dismissal of explanation: that explanation is not the gate through which restoration must pass, but a further achievement, with its own separate value, built on a functional competence that was in every case surveyed

here already present before the explanation arrived to account for it.

References

[1] Flyxion. 2026a. *Persistence Before Truth: Recoverable Distinction as the Condition of Reference, Explanation, and Knowledge*. Unpublished monograph.

[2] Flyxion. 2026b. *What Survives Reconstruction: Admissibility, Irreversibility, and the Architecture of Externalized Structure*. Unpublished monograph.

[3] Zhang, Zijian, et al. 2026. "Is One Layer Enough? Training a Single Transformer Layer Can Match Full-Parameter RL Training." *arXiv preprint* arXiv:2607.01232.

[4] Jastrzębski, Stanisław, Devansh Arpit, Nicolas Ballas, Vikas Verma, Tong Che, and Yoshua Bengio. 2018. "Residual Connections Encourage Iterative Inference." *International Conference on Learning Representations* (*ICLR*).

[5] Belrose, Nora, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. "Eliciting Latent Predictions from Transformers with the Tuned Lens." *arXiv preprint* arXiv:2303.08112.

[6] Ziembik, Krzysztof A. 2025. "Forth: The Best Programming Language for the End of the World? A Case Study on A.D. 2044 (1991)." *Proceedings of the 36th ACM Conference on Hypertext and Social Media*.

[7] Murray, Joseph E. 1990. "Nobel Lecture: The First Successful Organ Transplants in Man." Nobel Foundation.

[8] Dash, S. C., Ranjith Nair, and Vineet Behera. 2023. "Kidney Transplantation: The Journey Across a Century." *Medical Journal, Armed Forces India* 79(6): 631–637.