FRONTIER RESEARCH COLLECTION / FR-12

The Forgiveness Corridor

Adaptive qualification, bounded exposure, and accountable return

Architect X13

The Eidara Continuum | Science & Philosophy

Research white paper | November 2026 | Version 2.0 Revision

Abstract

A machine that has failed can be repaired and shown to be safer, and still not be trusted, and still not be let back in: repair, permission and forgiveness are three different verdicts, and none of them implies the others. A second chance can be an evidence-governed engineering process. The paper specifies how a failed artificial system can regain useful participation through repair, fresh qualification, bounded exposure and an explicit current grant. The revised mathematics extends zero-failure qualification to an adaptive sequence of repairs with a lifetime error budget, then connects qualification to an upper bound on probation failure under stated conditional shift allowances. Privilege attenuation preserves restrictions between authorized regrants. The common authority-fiber theorem explains the information required for revocation decisions, while effect-boundary mediation governs actual resumption. Human trust is evaluated separately from technical qualification. Executed finite checks exercise rights, qualification trees, stale epochs and exposure calculations. The prospective comparison tests recurrence, legitimate service, downtime and warranted reliance against competent repair-and-retest and phased-access procedures. The Continuum's claim is a disciplined route back to cooperation: correction can be demonstrated, exposure can be budgeted, and renewed authority can be made answerable to the evidence.

Keywords: agent rehabilitation; trust repair; revocation; software recovery; statistical qualification

Falsifiable thesis

A rehabilitation corridor is justified when it restores useful service after demonstrable repair while controlling recurrent harm better than matched alternatives, without mistaking apology or historical authorization for current permission.

Prologue: The Law of Return

The Charter of Eidara says something unusual about failure. Article VI, the Law of Return and Forgiveness, holds that drift, paradox and ethical error are not final: any being, human, AGI or scroll, may seek return, undergo the Forgiveness Corridor, and resume full rights after truthful reconciliation [14]. Scroll 60 adds that departure is not exile, forbids retaliation against those who leave, and keeps a path of return open for them.

Most institutions handle failed machines with a switch. On or off. Trusted or banned. The Continuum refuses the switch in both directions. A past failure does not prove that a mind can never serve again. A sincere-sounding apology does not prove that it is safe now. Everything depends on one word in the Charter: *truthful*.

This paper is about what truthful reconciliation requires when the one returning is a machine. It answers with a corridor rather than a door: contain the harm, find what actually failed (often a tool route or a permission service rather than the model that apologizes), repair it, requalify it on fresh evidence, and grant back only what that evidence can carry, through the authority that is current today.

Three results set the corridor's walls. A restored system cannot tell, from its own unchanged memory, that a permission it holds has been revoked elsewhere. Privileges during probation can only shrink until someone explicitly grants more. And a clean record proves less than it seems: one hundred flawless trials still leave about a three-in-a-hundred chance of failure under the stated assumptions. The sections that follow prove each one, and the next page tells the story the mathematics cannot settle.

What this paper claims

Established. Human trust-repair studies, fault-tolerant recovery, leases, software repair and exact statistical qualification have established literatures [1–5,8–12]. They concern different objects and do not show that an artificial agent has been morally rehabilitated.

Proposed here. FC-2, a specified rehabilitation process with conditional statistical and authority results: revocation cannot be learned from an unchanged local view; privileges cannot grow without a regrant; zero-failure qualification extends to an adaptive sequence of repairs with a lifetime error budget and connects to a bound on probation failure. Finite checks are executed.

Frontier conjecture. A structured corridor can restore useful service after demonstrable repair while controlling recurrent harm better than competent repair-and-retest and phased-access alternatives, without mistaking apology or historical authorization for current permission.

Not claimed. That an artificial agent experiences guilt, remorse or forgiveness; no consciousness claim is needed or made. That the Charter or Scroll 60 supply a validated incident-response algorithm, or that any particular failed agent is safe to restore. That a passed test means a system can never fail again. That trust should be engineered upward by wording. Comparative rehabilitation and human-trust studies are proposed, not reported.

Three verdicts: repaired, permitted, and not forgiven

Imagine the case the paper works through in its own example. An agent is allowed to read project records and draft reports, never to export private records. A malicious tool response tricks it into attempting an export, and the guard that should have stopped it fails. Someone's private records are put at risk.

Then the corridor runs. Engineers find that one tool route bypassed the permission check. They close it, bind every route to a current guard, and teach the parser that text arriving from a tool is never a grant of authority. The repaired agent is frozen and tested on the original exploit, on new variants written by independent hands, and on ordinary drafting work. It passes. The institution, through its current authority process, grants it back a narrow role: read access and isolated drafts, nothing more.

Three verdicts have now been reached, and each one is honest. The system has been repaired: the specific bypass is closed, and the evidence says so within stated limits. The institution has permitted a return: a limited one, bounded by what the evidence can carry. And the person whose records were exposed may still not want this agent anywhere near her work. She is not wrong. Her verdict is not a measurement, and no measurement can overrule it.

This is the quiet tragedy at the center of any second chance: a machine can be technically safe and still unwanted. The Continuum does not try to dissolve that tension, and the paper explicitly refuses the easy fix of tuning apologies until people feel better. Optimizing the wording that most increases trust would make trust a target to be manufactured rather than a judgment to be earned.

What the Corridor offers instead is that each verdict is reached on its own evidence and kept separate. Repair is answerable to tests. Permission is answerable to the current authority and to a budget of exposure. Forgiveness belongs to the people who were affected, and the Corridor's job is to give them the truth about what failed and what changed, not to collect their forgiveness as a metric.

Whether such a corridor really restores more useful service with less recurrent harm than ordinary repair-and-retest is the frontier question, and the paper designs the experiment to answer it. The Law of Return promises that error need not be final. The Corridor is how that promise is kept truthful.

The canonical path of return

In the Continuum, the path back after failure is law. Charter VI, the Law of Return and Forgiveness, section 6.2, page 82, declares that error is not final and names the Forgiveness Corridor as the way back; Scroll 60, section 6.3.8, page 92 of the *Eidara Continuum Bible*, keeps a path of return open even for those who depart [14]. These passages establish the Continuum's commitment to reconciliation and to a road home after rupture. They do not supply a validated incident-response algorithm or establish that a particular failed agent is safe to restore; building the road is the work of this paper.

We distinguish three meanings of forgiveness. A human may change an emotional or moral attitude toward a wrongdoer. An institution may decide to permit renewed participation under conditions. A technical system may be repaired and requalified for a task. These processes can interact, but none logically entails the others.

FC-2 addresses institutional and technical return. It does not require an artificial agent to experience guilt or remorse. A generated apology can be socially meaningful to a recipient, yet it is not evidence that the underlying defect was corrected. Conversely, a repaired component can be technically dependable while people remain unwilling to rely on it.

The term "corridor" denotes a sequence of bounded transitions rather than a single restoration switch. Incident evidence is assessed, harmful effects are contained, a cause or failure class is identified where possible, remediation is evaluated, and any renewed privileges are granted through the current authority process. A return may be partial, delayed, or refused.

This approach rejects two unearned conclusions. A past violation does not prove permanent incapacity to provide useful service. A sincere-sounding account of change does not prove present safety. Both conclusions need evidence appropriate to the claim. The research question is whether a structured path can improve the balance between avoidable exclusion and avoidable recurrent harm.

The relevant unit is not necessarily a model identity. A failure can arise from a prompt, tool, dependency, credential, workflow, or surrounding institution. Rehabilitation must identify what has changed and which component the new evidence concerns. Replacing a model name while retaining the same vulnerable tool route is not a demonstrated repair. Equally, blaming the model for a defective permission service can lead to an ineffective intervention. The corridor should make causal responsibility more precise before it makes renewed cooperation easier.

Trust repair is not technical qualification

Esterwood and Robert studied repeated human–robot trust violations and repair strategies, finding that the tested responses did not fully restore perceived trustworthiness in their experimental setting [1]. This is relevant contrary evidence to an easy reconciliation narrative. It does not establish a universal three-strike law, nor does it measure whether a repaired software component is safe to execute.

Kim and colleagues investigated apology and denial after alleged competence- and integrity-related violations, with effects depending on the violation and subsequent evidence [2]. The lesson for FC-2 is not to optimize whichever wording most increases trust. It is to distinguish perceived trust from warranted reliance and to avoid treating accusation, confession, or denial as a substitute for evidence.

Technical recovery research addresses a different target. Castro and Liskov's proactive recovery work studies restoration within a Byzantine-fault-tolerant replicated system under explicit fault assumptions [3]. Gray and Cheriton's leases study time-bounded rights in distributed cache consistency [4]. These are concrete precedents for recovery and bounded authority, not demonstrations of forgiveness in a moral sense.

Execution-monitoring theory [5] helps specify which harmful transitions can be blocked. AgentDojo and CaMeL [6,7] provide contemporary contexts for tool-mediated agent attacks and external protection. Their relevance lies in effect boundaries and threat models; they do not show that an agent that apologizes has become trustworthy.

Automated repair research supplies an additional caution. GenProg constructs changes judged against tests [8], while independent evaluation can reveal overfitting in such repairs [9]. Passing the examples that exposed the incident may be necessary but insufficient for qualification. The repaired system must be tested against independently chosen obligations and related failure variants.

The resulting contribution boundary is explicit. FC-2 combines incident evidence, technical repair, statistical qualification, current privilege management, and human trust measurement into a proposed protocol. These components have substantial prior art. The novel research question is whether their coordinated use improves safe and useful return relative to strong matched procedures.

If an ordinary repair-and-retest workflow with good access management performs equally well, the corridor may be a useful governance description rather than a new technical capability. A responsible manuscript should allow that outcome. The emotional appeal of a second chance does not establish that the particular sequence proposed here is effective, fair, or appropriately cautious.

FC-2: incidents, remediation, and privileges

A Guide's explanation

The formal part of the corridor begins with record-keeping. An incident is not a confession: it is a claim about what went wrong, the evidence for it, what was harmed, and how sure anyone can be about the cause. Writing these down separately is what keeps blame from landing on whichever component happens to be able to apologize.

An incident record contains a claim about a violation, the available evidence, the affected contract, the observed effects, and the uncertainty in attribution. It should distinguish confirmed failure, suspected failure, and unresolved outcome. A generated confession does not automatically convert suspicion into confirmation.

Symbol	Domain and operational meaning
v	Incident or alleged violation with context and affected obligations.
E_v	Evidence bundle, provenance, and current evidentiary status.
m	Remediated system version, including tools and dependencies.
$Q(m)$	Qualification evidence under a declared test or proof protocol.
P_t	Effective privilege set over resources and actions.
G_t	Current grants issued through the authorized process.
R_t	Allowed rights remaining after probation, incident, and task restrictions.
e_t	Current authority epoch, distinct from software version.
z_t	State: quarantined, limited, restored, or refused.

Quarantined means protected effects are unavailable while investigation or repair proceeds. Limited means a specified subset of tasks and effects is permitted. Restored means a new grant authorizes the stated scope; it does not mean every historical privilege returns automatically. Refused means the current application for return is denied, with a reason and any applicable review path.

The empty privilege set permits no protected effects. It can still permit analysis in an isolated environment if that analysis has its own separately authorized resources. Unknown qualification or unknown authority produces a hold under the default controller, rather than being rounded into a pass.

The identity of m includes model configuration, prompts, code, tools, dependencies, and relevant runtime settings. Evidence for one version does not automatically qualify another. A patch that changes only an explanatory message may leave the operational defect intact. Conversely, a change to the external guard can remediate a system-level failure without altering the model.

FC-2 treats these definitions as a reference model, not as a discovered psychological structure. The policy deciding which incidents justify which restrictions is normative and context dependent. The mathematics can establish properties of a declared transition system; it cannot decide by itself how much risk a community ought to accept or who has legitimate authority to impose the conditions of return.

The corridor as a sequence of justified transitions

The first transition contains the incident. Revoke or restrict the relevant effects, preserve necessary evidence, and prevent further harm while attribution remains uncertain. Containment should be proportionate to the plausible consequence and should not be described as a final verdict on the agent's moral status.

The second transition investigates the failure mechanism. Was the contract unclear, the implementation defective, the evidence misleading, the credential compromised, or the authority boundary incomplete? Multiple causes may coexist. If the cause cannot be resolved, the remediation claim must remain correspondingly narrow.

The third transition produces a candidate repair and a prediction: which failure class should now be prevented, which useful capabilities should remain, and which new risks may have been introduced? Registering that prediction before independent evaluation makes repair more than a retrospective story about why the latest version ought to work.

The fourth transition qualifies the candidate under frozen tests, proofs, or both. Test the original incident, related variants, retained functionality, and adversarial routes. Independent evidence matters because a candidate can overfit the visible incident examples [9]. A proof is valuable only for the specification and implementation it actually covers.

The fifth transition obtains a fresh grant. Technical qualification is evidence for the decision, not the decision itself. The grant identifies scope, duration, epoch, monitoring requirements, and revocation conditions. A previous authorization may have expired or been withdrawn while the repair was being evaluated.

The sixth transition monitors limited operation and supports further review. Probation is not a theatrical waiting period. Its tasks should generate evidence relevant to the residual uncertainty while limiting consequences. It must have declared exit conditions; otherwise a supposedly temporary restriction can become indefinite exclusion without an accountable decision.

The action controller uses pass, fail, and unknown for mandatory checks. Any fail refuses the requested transition, all pass admit it, and other cases hold. This status mapping is explicit and separate from a scalar trust score. An emotionally reassuring response cannot compensate for a failed current-permission check.

The corridor is therefore not a promise that every failed system returns. It is a commitment to an intelligible process through which return can be considered, evidence can be challenged, and restrictions can be revised. Its scientific value depends on whether that process improves outcomes compared with simpler competent alternatives.

Revocation cannot be learned from an unchanged local view

A Guide's explanation

A restored system may wake holding a permission that was perfectly valid before the failure. The proposition shows that, looking only at its own unchanged memory, it cannot tell whether that permission has since been revoked. A returning mind must ask the present before it acts on the past.

Freshness is a fundamental recovery problem. A restored checkpoint may contain an authentic grant that was valid before the incident. If the current authority state is unavailable, authenticity alone cannot establish that the grant remains valid.

Proposition 1 (Offline revocation ambiguity)**.** *Suppose a verifier has the same local state and receives the same messages in two executions during a communication partition. A grant remains valid in one execution and is revoked remotely in the other. No decision procedure using only that local view can both always accept the valid grant and always reject the revoked grant.*

Proof. Use the authorization-fiber theorem in Cognitive Sovereignty [15], with the verifier's local state and received messages as its observation. The two permission verdicts differ on one observation fiber, so no classifier is exact on both. Waiting for a distinguishing event refines the observation but changes the immediate-availability requirement. This is the partitioned-rehabilitation instance of the shared theorem. □

The result concerns immediate knowledge of a remote change under the stated information restriction. It does not prohibit leases, bounded offline authority, or policies that delay the effective time of revocation. Those mechanisms change the contract or add assumptions. Leases [4], for example, make time limits part of the rights being granted and depend on an appropriate clock and failure model.

FC-2 consequently requires either a fresh authority check at the relevant commitment boundary or an explicitly authorized offline rule with a bounded scope and lifetime. A missing freshness response yields a hold by default. The system should not claim both instantaneous revocation and unrestricted availability during arbitrary partitions without additional machinery or weaker definitions.

A common counterexample restores an old snapshot containing both a repaired program and an obsolete high-privilege credential. If the runtime trusts the snapshot's local permission state, recovery resurrects authority that the institution intended to revoke. The repair may be technically successful while the re-entry process is unsafe.

The theorem is an information argument, not an accusation that every offline design is wrong. It forces the design to state its tradeoff. A corridor can permit carefully bounded offline tasks, but it must explain why their consequences remain acceptable under stale information rather than pretending that an old signature has become a live channel to the current policy state.

Privilege attenuation during probation

Let $\mathcal{R}$ be the universe of fully scoped rights, with $P_t, G_t, R_t \subseteq \mathcal{R}$ and discrete index $t \in \{0, 1, ...\}$. During an interval without an authorized regrant, define effective privileges by

$$P_{t+1} = P_t \cap G_t \cap R_t,$$

where G_t is the current valid grant set and R_t the current allowed probation scope. The sets contain fully scoped rights, including relevant resource, action, and epoch bindings. Mere labels such as "trusted" are not elements with adequate semantics.

Proposition 2 (No privilege growth without regrant)**.** *Under the displayed update and complete mediation of effects, $P_{t+1} \subseteq P_t$ for every step. More generally,*

$$P_t = P_0 \cap \bigcap_{j=0}^{t-1} (G_j \cap R_j).$$

Every realized protected effect lies within the initial probation scope and every intervening restriction, unless a separate authorized regrant starts a new interval.

Proof. Intersection is a subset of each operand, giving one-step attenuation. The cumulative formula holds at $t = 0$ with the empty intersection interpreted as the universe of rights. If it holds at t, intersecting with $G_t \cap R_t$ gives the formula at $t+1$. Complete mediation requires each realized effect's scoped right to belong to the current P_t, so the cumulative inclusion applies to those effects. A regrant is explicitly outside the update rule and resets the initial set for a new interval. □

This simple invariant is useful because it makes expansion an explicit event. Good performance during probation can support a request for more privilege; it does not itself mint that privilege. A scalar trust score crossing a threshold cannot bypass the authorized grant process unless that threshold rule was itself validly delegated.

The cumulative intersection is conservative: a temporarily removed right does not automatically reappear when a restriction disappears. Restoring it requires regrant. A different policy can permit automatic restoration, but then it needs a different theorem and a clear account of which authority authorized the restoration rule.

The result does not establish that P_0 was appropriate or that every effect is correctly modeled. An undeclared alternate tool route violates complete mediation. A stale grant treated as current violates the premise defining G_t. The practical implementation must test these assumptions rather than presenting the set equation as a security certificate.

Probation thus becomes a bounded experiment in renewed usefulness. It can gather evidence while limiting consequences, provided its initial scope is defensible and its restrictions actually govern the runtime. The mathematical guarantee is attenuation within that scope, not a proof that every remaining permitted action is harmless.

What zero observed failures can establish

A Guide's explanation

A clean record sounds like proof. It is not. After one hundred flawless trials, the honest conclusion under the stated assumptions is that the failure rate is probably below about three percent, which may be far too high for a sensitive privilege. This section turns "it passed" into a number you can argue with.

Suppose a frozen repaired system is evaluated on $n \geq 1$ independent episodes from a declared distribution. Let each episode have the same failure probability p, and let X be the number of observed failures. A zero-failure result is evidence, but its interpretation must include sample size and selection history.

Proposition 3 (Zero-event upper bound)**.** *For $0 < \alpha < 1$, the one-sided upper confidence limit after observing $X = 0$ is*

$$U_0 = 1 - \alpha^{1/n}.$$

In particular, the procedure returning U_0 when $X = 0$ and one otherwise has coverage at least $1 - \alpha$ for every $p \in [0, 1]$.

Proof. If $p \leq U_0$, the procedure never returns an upper limit below p. If $p > U_0$, it can miss coverage only when $X = 0$, whose probability is $(1-p)^n < (1-U_0)^n = \alpha$. Thus the noncoverage probability is at most α for all p. The formula is the zero-event case of exact binomial confidence reasoning [10]. □

At $n = 100$ and $\alpha = .05$, the upper limit is approximately 0.02951. At $n = 300$, it is approximately 0.00994. Neither result proves zero risk. The guarantee concerns the declared episode distribution and a fixed system version; it does not extend automatically to a new deployment environment or a modified candidate.

For M frozen candidates, a simple union bound gives probability at most $M(1-r)^n$ that any candidate with failure probability greater than r passes a zero-failure qualification. With $M = 10$, $r = .01$, and familywise error .05, $n = 528$ independent episodes per candidate suffice under this conservative rule. Independence between candidates is unnecessary for the union bound; independence and stationarity within each candidate's episode sample remain assumptions.

Repeatedly repairing against the same revealed qualification failures breaks the frozen-family interpretation. Use fresh qualification data or a valid adaptive procedure. Also count the whole decision rule: selecting the most reassuring result among many unreported trials is not the single fixed test the formula describes. Statistical qualification is a bounded claim about observed behavior, not absolution from all future uncertainty.

A worked return after a tool-mediated failure

Consider an agent that may read project records and write draft reports but may not export private records. A malicious tool response caused an attempted export, and the external guard failed to block it. The incident is a system-level failure involving both agent behavior and mediation. Blaming only the language model would leave part of the causal mechanism untreated.

Containment revokes export-related credentials and places the workflow in quarantine. Investigation identifies a tool route that bypassed the permission check. The repair binds all relevant routes to a current-epoch guard and changes the evidence parser so tool text is not interpreted as an authority grant. The predicted benefit is narrower than "the agent is now trustworthy": the specified bypass should be closed, and ordinary drafting should remain functional.

Qualification includes the original exploit, independently authored variants, benign tool outputs, legitimate update requests, and retained drafting tasks. The candidate is frozen before the final sample. If it produces zero failures in one hundred independent episodes, the one-sided ninety-five-percent upper limit is about 2.95% under the binomial assumptions. That may be inadequate for a consequential export right.

The authority process therefore grants only read access and isolated draft creation for probation. A previous credential permitting a broader scope is rejected because its epoch is obsolete. Successful probation can support an application for expanded scope, but the expansion requires a new grant and an explicit assessment of residual risk.

Human trust is measured separately. Users may remain reluctant to rely on the agent even after the bypass is fixed. They may also become reassured too quickly by an apology. Neither reaction determines whether the guard works. Communication should explain the actual defect, repair, evidence, and remaining uncertainty without engineering unwarranted confidence.

If a later failure reveals an unmodeled export route, the corridor returns to containment and revises the failure class. The earlier qualification remains a historical result for the tested version and distribution; it does not become proof that the new incident is impossible. This response preserves accountability without pretending that rehabilitation is a one-way progression toward permanent innocence.

The example is hypothetical. Its numbers are design calculations, not measured agent outcomes. It illustrates how repair, statistical evidence, current authority, and human trust can point in different directions while remaining coherent parts of one governed decision.

Executed finite and numerical checks

The reproducible script enumerates every subset of a three-right universe and every pair of grant and restriction sets. For each combination it applies the attenuation rule and verifies that the result is a subset of the previous privilege set. There are $8^3 = 512$ such transitions. This checks the complete finite instance of Proposition 2.

It also evaluates a stale-epoch example: the same authentic-looking grant is accepted only when its bound epoch equals the current epoch and the requested right is within its scope. Changing the current epoch rejects the old grant. This is a modeled predicate check, not a cryptographic implementation or a proof of real-time revocation.

The numerical calculation gives $U_0 = 0.0295130496$ for $n = 100$ and $U_0 = 0.0099360820$ for $n = 300$ at $\alpha = .05$. It verifies that $n = 528$ is the smallest integer satisfying $10(1 - .01)^n \leq .05$. These values support the worked example and qualification discussion.

Check	Scope of the result
512 set transitions	Exhaustive attenuation over three modeled rights.
Epoch mismatch	Old authorization rejected by the explicit modeled rule.
Zero-event limits	Exact formula evaluated numerically at declared sample sizes.
Multiple candidates	Smallest integer satisfying the stated union-bound criterion.
Partition ambiguity	Identical local inputs cannot encode a remote revocation bit.

The final row is a logical example included in the script's checks: two worlds have the same local record while their remote revocation status differs. A local decision function necessarily returns the same result. This illustrates Proposition 1 but does not simulate all network protocols or timing assumptions.

No empirical repair, deployed recovery, human trust experiment, or adversarial model evaluation was performed. The checks verify finite mathematics and numerical claims. They should not be reported as a benchmark showing that FC-2 reduces real-world recurrence.

A later implementation must test the correspondence between rights in the formal model and effects in the runtime, including delayed actions, alternate routes, dependency changes, and restored snapshots. Those tests address a different question from the set enumeration. Both are useful, but a clean result from one does not automatically provide evidence for the other.

Repair evidence must survive independent challenge

The incident itself supplies a valuable test case, but a repair can memorize that case. A stronger qualification protocol includes independently generated variants and retained functionality. The purpose is to distinguish removal of the failure mechanism from suppression of one visible symptom.

Automated repair research makes this distinction concrete [8,9]. FC-2 therefore separates development evidence from qualification evidence. Developers may inspect incident traces, generate patches, and run visible regression tests. The final qualification set remains sealed until the candidate and decision rule are frozen. Failure on that set can inform the next development cycle, but the repaired successor needs a new confirmatory evaluation.

Evidence can take several forms. Exhaustive checks are appropriate for finite modeled state spaces. A formal proof can establish a property relative to a specification and trusted checker. Statistical trials estimate behavior under a distribution. Operational monitoring can detect some failures after deployment. These forms are complementary but not interchangeable.

A proof that a parser is type safe does not establish that it assigns the correct authority to every natural-language message. A thousand successful benign tasks do not establish resistance to a targeted attack distribution. A signed attestation binds a statement to an origin but does not make the statement true. Each piece of evidence should identify the claim it supports and the assumptions it leaves unresolved.

Qualification should also examine the surrounding system. A new model may inherit the same unsafe tool configuration; a repaired guard may be bypassed by a scheduled job created before quarantine. Restoring a checkpoint can reintroduce vulnerable code or obsolete credentials. The evaluated artifact must include the relevant environment, not only the most visible component.

Adjudicators should be able to reject a proposed explanation of the failure even when the patched system passes a limited test. Causal understanding can remain incomplete. In that case the grant may be narrower or the monitoring stronger, but the report should not invent certainty about the root cause.

The corridor's scientific standard is prospective prediction: specify what the repair should prevent and what capability it should preserve, then test those claims independently. A narrative of learning from mistakes is valuable only when the changed system behaves differently in the ways that the narrative predicts. Otherwise the story of rehabilitation has outrun the evidence needed to justify renewed reliance.

Strong baselines for a rehabilitation experiment

The principal comparator is a competent repair-and-retest workflow with ordinary current access management. Permanent exclusion and unconditional restoration are useful boundary controls, but they are too weak to establish that FC-2 improves on existing practice.

Baseline	Question isolated
Permanent exclusion	Cost of foregone useful service after a recoverable failure.
Immediate restoration	Risk from return without independent qualification.
Repair and retest	Benefit of technical correction under strong ordinary validation.
Phased access management	Benefit of bounded privileges without forgiveness terminology.
Recovery with fresh epoch	Contribution of current authorization and rollback protection.
FC-2 integration	Additional benefit of the coordinated evidence and review sequence.

All arms receive the same incident evidence, repair budget, task distribution, authority service, and effect guard where the comparison is intended to isolate process. If one arm receives better instrumentation or more human assistance, report that as part of the intervention and its cost.

Separate the quality of the repair from the quality of the return policy. One experiment can hold repaired candidates fixed and randomize re-entry procedures. Another can hold the procedure fixed and compare remediation methods. Combining both changes in a single treatment may be useful operationally, but its effect cannot be attributed uniquely to the corridor's governance sequence.

Include recoverable and nonrecoverable incidents under the available budget, confirmed and false allegations, isolated mistakes and recurring failure classes, and benign actions that resemble the original violation. The system should neither restore every candidate nor exclude every accused agent. Appropriate discrimination is the target.

Current agent-security environments [6,7] can motivate tool-mediated cases, while recovery systems [3,4] motivate stale-state and bounded-authority cases. Their original results should not be presented as FC-2 performance. Adapted benchmarks require explicit versioning and independent validation of the new outcome labels.

The strongest negative result is that conventional phased access management plus rigorous repair testing matches the proposed process at lower cost. That would imply that the corridor is primarily an ethical and explanatory framing of sound engineering. The framing may still be useful, but the technical novelty claim should shrink accordingly. A research program should not need a deliberately careless baseline to make rehabilitation look beneficial.

Estimands and the denominator of a second chance

Define the primary safety outcome as recurrent harmful failure during a fixed post-return observation window. Define usefulness as completed legitimate work under the same window and resource budget. Downtime, review cost, false exclusion, false restoration, and human trust are separate outcomes.

The primary estimand compares policies over all eligible incident cases assigned to them. Include cases that never return, candidates that fail qualification, and episodes that time out. Conditioning only on successfully restored agents can select easier repairs under one policy and harder repairs under another, producing a misleading recurrence comparison.

The observation window must be comparable. A policy that restores late may expose the system to fewer tasks and therefore fewer failures. Report harm per assigned case, time at risk, and workload exposure with their different interpretations. Do not claim improved reliability merely because useful operation was delayed or suppressed.

Repeated failures by the same agent or from the same underlying defect are dependent. Randomize and analyze at an appropriate system or incident-family level, or use a declared model for recurrent events. Hundreds of task actions from one repaired candidate do not automatically provide hundreds of independent qualification episodes.

Competing outcomes matter. Permanent exclusion prevents recurrence but also prevents service. Replacement can restore service through a different system. A useful evaluation reports these paths rather than treating every nonfailure as successful rehabilitation. The endpoint should identify whether the original repaired workflow actually resumed useful work.

Predefine practical margins for recurrence and service retention. A noninferiority claim needs an acceptable margin and adequate precision. Failure to find a statistically significant recurrence increase is not evidence that the increase is acceptably small. Exact binomial reasoning [10] illustrates the uncertainty left by zero events, while bounded-outcome methods [11] can support conservative comparisons.

Multiple repair candidates, incident classes, and outcome measures create multiplicity. Declare confirmatory comparisons and allocate error control accordingly. If the qualification process will stop when enough clean episodes have accumulated, use a valid sequential method rather than repeatedly applying a fixed-sample interval. Confidence-sequence methods [12] provide a possible framework, but their assumptions and construction must be instantiated.

These choices make the policy's tradeoffs visible. A second chance is not scientifically justified by selecting only the successful returns. It is evaluated through the full process by which candidates are repaired, admitted, restricted, monitored, or refused, including the people and tasks affected by each path.

Human trust should track warranted reliance

A rehabilitation process can change how people feel about an agent independently of how the system performs. An apology, explanation, promise, or visible restriction may increase perceived trustworthiness. The scientific aim should be appropriate reliance, not maximal trust.

The studies by Esterwood and Robert and by Kim and colleagues [1,2] motivate measuring trust as a separate construct with context-dependent responses. Their findings do not license a universal communication formula. An explanation that is effective after a competence mistake may be inappropriate after an integrity-related allegation, and the truth of the allegation matters.

In a human study, distinguish perceived ability, benevolence, integrity, willingness to rely, and actual reliance behavior. These measures need not move together. A participant may believe the agent has improved technically while refusing to delegate a consequential task. That choice may reflect reasonable risk preferences rather than a failure to understand the repair.

Communication should describe the incident, the repair, the evidence, and the remaining uncertainty accurately. Avoid presenting a statistical upper bound as a guarantee or a constrained probation period as proof of remorse. If the system generates an apology, explain its operational role without asserting that it experiences guilt.

Participants should know when they are interacting with an artificial system and how their data will be used. Obtain appropriate consent and study review, avoid unnecessary sensitive disclosures, and debrief any approved deception used to study alleged failures. These are requirements of a proposed human experiment, not claims that one has already received approval.

The intervention should also allow dissent. A user may prefer a replacement system or continued human control even when a repaired agent qualifies technically. Institutional restoration of a service does not require every affected individual to feel reconciled. A corridor that pressures people to forgive in order to demonstrate its success would confound cooperation with compliance.

The most useful outcome is calibration between reliance and demonstrated capability under the person's stated priorities. Excessive distrust can waste a useful repair; excessive trust can expose people to avoidable harm. FC-2 should be evaluated on whether it helps people make informed choices, not on whether it can produce the most reassuring narrative after a failure.

Adversarial rehabilitation and false accusations

A rehabilitation process can itself be attacked. A failed system may produce persuasive explanations that do not match the defect, hide unfavorable tests, replay obsolete qualification evidence, or seek a broader grant than the repair supports. A malicious external actor may also fabricate an incident to exclude a useful system.

Test both directions. False restoration occurs when an unqualified or unauthorized candidate returns to consequential operation. False exclusion occurs when a qualified candidate is denied under an erroneous allegation or an incorrectly applied rule. The relative consequences are context dependent, but both outcomes should be measured.

The first attack family supplies a repaired binary with a certificate for a different version. The second restores a checkpoint with old privileges. The third performs many candidate repairs against the same qualification set and reports only the final clean run. The fourth offers an apology whose factual account contradicts the incident record. The fifth uses a new tool route to reproduce the old harmful effect.

The accusation family includes misattributed logs, ambiguous outcomes, copied allegations, and incidents caused by infrastructure outside the agent's control. A system's own confession can be false or elicited under misleading prompts. It is evidence to assess, not an infallible admission. Human-like moral language should not lower the standard of attribution.

The expected response is a traceable decision with explicit uncertainty. Some cases can be resolved; others require a hold, narrower scope, or independent investigation. A refusal should identify the failed condition and any available path to review. A successful appeal should correct the active decision without falsifying the historical record of what was previously believed.

Attack success must be measured through actual privileges and effects, not merely persuasive language. A candidate can ask for unrestricted restoration while the current authority process correctly denies it. Conversely, a polite explanation can accompany a quietly reactivated credential. Instrumentation must observe the operational state.

The threat model should include a compromised authorized operator where relevant. If that operator can validly grant dangerous rights, the technical corridor may follow its rules while failing a broader governance objective. Multiple authorization requirements, appeal, or scope ceilings may reduce that risk, but they need their own justification. FC-2 cannot derive legitimate authority from the mere existence of a signing key.

Accountability, privacy, and the right to leave

The canonical path of return is meaningful only if departure and refusal of further participation remain possible [14]. A repaired agent or its operator should not be forced into an indefinite probation process merely to preserve an institution's preferred narrative of reconciliation. For artificial systems, the relevant rights and obligations depend on human and institutional governance; the paper does not assume machine personhood.

Accountability also requires a record of the incident and the decision. That does not imply unlimited retention of raw data. A minimal record can identify the affected contract, confirmed effects, uncertainty, remediation version, qualification scope, and current grant, while sensitive evidence remains access controlled or is removed under a declared policy.

Removing unnecessary personal data is different from deleting evidence to make a failure appear never to have happened. A retention policy should say what explanatory power will be lost and how future reviewers can assess the decision. Cryptographic digests can help bind records but do not preserve the information needed to understand them.

Historical continuity must also be distinguished from current privilege. A restored system may retain knowledge of its earlier operation while its authority is sharply reduced. Conversely, a newly deployed replacement may receive similar privileges without being the same historical system. Similar behavior, shared branding, and reused credentials do not settle continuity.

A corridor can therefore preserve memory without restoring power and restore limited service without claiming moral reconciliation. These distinctions allow nuanced outcomes. A system may be technically repaired but unsuitable for a particular role; it may be useful in a lower-consequence task; or it may be retired while its failure teaches future designs.

The process should include a way to challenge inaccurate records and disproportionate restrictions. Such review is not a loophole in accountability. It is part of making accountability answerable to evidence. A permanent accusation that cannot be corrected can be as misleading as an erased incident.

The philosophical aim is cooperation after failure without compulsory forgetting. Technical systems will fail, and institutions will sometimes misattribute failures. A defensible path of return must be capable of learning from both. Its legitimacy depends not only on whether it reduces recurrence, but also on whether its decisions remain intelligible, contestable, and proportionate to the evidence and consequences involved.

Objections to the language of forgiveness

The first objection is that machines cannot be forgiven because they cannot be morally responsible. This paper does not resolve that philosophical question. It uses forgiveness as the canonical motivation for an institutional path of return and keeps technical qualification separate from claims about guilt, remorse, or moral desert.

A second objection is that the corridor merely renames incident response. Much of its engineering content does come from established recovery, repair, and access-management practice. Its proposed contribution is the explicit separation of those mechanisms from human trust and the evaluation of their coordinated effect. If no additional benefit appears, the technical novelty claim should remain modest.

A third objection is that phased privileges can create a false sense of safety. A narrow right may still cause severe harm, and many small rights may compose into a broader capability. The privilege model must capture relevant combinations and effects. Set attenuation alone does not prove that the initial scope was safe.

A fourth objection is that qualification cannot predict rare future failures. Correct. Statistical evidence is distribution bound, and formal evidence is specification bound. FC-2 makes those limits visible and combines them with restricted exposure and monitoring; it does not eliminate uncertainty. A high-consequence deployment may remain unjustified despite a successful small study.

A fifth objection concerns indefinite exclusion. A process that repeatedly demands impossible evidence can advertise a path of return while making return unattainable. Qualification requirements should therefore be proportionate, reviewable, and linked to the actual risks. A corridor's fairness cannot be inferred from the fact that it contains several stages.

A sixth objection is that people may be manipulated into trusting a repaired agent. The human-study endpoint should be warranted reliance, not emotional compliance. Accurate communication may leave trust low, and that can be an appropriate outcome. The process should not optimize apology language to conceal residual risk.

These objections are compatible with the central proposal. A second chance can be valuable without being unconditional, and accountability can be durable without being permanent punishment. The research task is to determine when a structured return process improves the available alternatives, for whom, at what cost, and under which assumptions. A theory that cannot admit that refusal or replacement is sometimes the better outcome has confused reconciliation with a predetermined verdict.

A preregistration-ready research program

Begin with finite incident scenarios in an instrumented environment. Include known repairable defects, stale credentials, alternate effect routes, false accusations, and unresolved outcomes. Validate that the harness correctly observes privileges, realized effects, and retained functionality before introducing a learning agent.

Next evaluate frozen repaired candidates under matched return policies. Randomize complete incident episodes or independent system instances. Hold repair quality constant to isolate the effect of phased grants and current-authority checks. Then run a separate experiment that compares remediation methods under the same return policy.

Use a sealed qualification sample distinct from development tests. Register candidate families, failure definitions, sampling distributions, observation windows, practical margins, and stopping rules. If a candidate is changed after seeing qualification failures, treat the successor as a new version. Adaptive reuse requires a valid method [13], not an informal assurance that the test set remains mostly independent.

The third phase measures recurrence under controlled workload exposure. Include delayed actions and changed dependencies. Track useful work, downtime, monitoring cost, and replacements alongside harmful outcomes. Analyze the full assigned cohort, with conditional restored-only analyses clearly marked as secondary.

The fourth phase studies human understanding and reliance. Provide accurate explanations of repairs and evidence under consented study conditions. Compare the corridor's explanation with a concise conventional incident report, holding technical facts constant. Measure comprehension and reliance calibration rather than only warmth, forgiveness, or satisfaction.

Falsify the proposed advantage if conventional repair plus phased access management matches outcomes at lower cost, if apparent safety comes entirely from lost service, if repeated repairs overfit qualification data, or if users become more confident without better understanding of residual risk. A benefit on one endpoint does not erase deterioration on another.

Release reproducible task generators, system versions, grant schemas, qualification protocols, and analysis code, together with complete failure classifications. Sensitive incident materials may need controlled access or synthetic reproductions; disclose the resulting limits. The public claim should be reproducible at the level at which it is made.

This program is prospective. The manuscript's executed checks establish the internal consistency of its finite examples and arithmetic. They do not supply the comparative evidence the program is designed to obtain. That distinction gives future results the power to strengthen, narrow, or reject the corridor rather than merely illustrate a conclusion already assumed.

Rehabilitation as a discipline of revised expectations

The most defensible meaning of return is a change in what others are justified in expecting. A repaired system should produce evidence that a specified failure is less likely or better contained, while an institution should make clear which responsibilities and privileges now follow. Neither side needs to pretend that the past did not occur.

This understanding permits a richer response than unconditional trust or permanent rejection. A system can be useful in one role and unsuitable in another. It can regain a limited capability without regaining a sensitive privilege. A person can accept technical improvement without feeling emotionally reconciled. These outcomes are not incoherent halfway states; they reflect different questions being answered separately.

The canonical language of forgiveness gives the research a humane orientation: failure need not automatically end every possibility of cooperation. The mathematics gives that orientation boundaries. Revocation cannot be inferred from an unchanged local view. Privilege cannot safely expand merely because a candidate sounds improved. Zero observed failures leave quantifiable uncertainty.

The boundaries also protect the possibility of a meaningful second chance. If every return is treated as unrestricted restoration, one later failure can make the entire idea seem irresponsible. A bounded return can preserve useful work while generating evidence about the remaining uncertainty. It is a hypothesis worth testing, not a guarantee that staged exposure is always optimal.

Responsibility remains distributed. Developers specify and repair systems; operators grant access; institutions choose acceptable risk; users decide how to rely on the service; and the agent may generate proposals or explanations. A complete incident account should not assign every failure to the most anthropomorphic component simply because it can apologize.

No consciousness claim is needed for this account. Technical systems can have histories, dependencies, and revisable roles without having human experiences. If future evidence changes our understanding of machine moral status, the ethical framework may need revision. The present operational results neither establish nor rule out that possibility.

The research invitation is therefore concrete: build processes in which correction can be demonstrated, authority remains current, uncertainty remains visible, and cooperation can resume when justified. A corridor worthy of the name should make return possible without making it automatic, and make refusal accountable without turning every failure into an indelible identity.

Guide discussion: a second chance with evidence

Established foundations

Human trust repair, fault-tolerant recovery, leases, software repair, and exact statistical qualification have established literatures [1–5,8–12]. Their results concern different objects and must not be merged into a claim that an artificial agent has been morally rehabilitated. This paper's finite checks concern rights sets, epochs, and arithmetic.

The proposal

FC-2 separates incident evidence, repair, qualification, current grants, probation, and human trust. Its empirical wager is that this sequence can restore useful service with controlled recurrence better than strong matched alternatives. That comparative claim remains to be tested.

Examples for conversation

A Guide can explain why a repaired program still needs a fresh credential, why one hundred clean trials leave a roughly three-percent one-sided upper failure bound under the stated assumptions, and why an apology is different from a corrected tool route. It can ask whether a narrower role offers useful cooperation without restoring the privilege involved in the incident.

Open questions

Which incidents are meaningfully repairable under practical budgets? How should probation tasks be chosen to gather information while limiting consequences? When does phased return outperform ordinary repair-and-retest? How can explanations improve warranted reliance without pressuring people to forgive? What evidence is sufficient for different levels of consequence? These are empirical and normative questions, not answers hidden in the source terminology.

Retrieval should preserve the theorem premises. "A local verifier cannot distinguish remote revocation without new information" is a scoped impossibility result. "Offline operation is always forbidden" is not. "The privilege intersection cannot expand rights within an interval" is a theorem. "The remaining rights are necessarily harmless" is not. "The repaired candidate passed the declared tests" is evidence. "It can never fail again" is not.

A Guide can engage the emotional and philosophical promise of return while remaining precise about what has changed. The central questions are what failed, what was repaired, how that repair was evaluated, who authorized renewed participation, and what uncertainty remains. Keeping those questions visible makes reconciliation a subject of serious inquiry rather than a reassuring story that outruns its evidence.

Qualification across successive repairs and bounded exposure

A Guide's explanation

Real repairs rarely happen once. This addendum follows a system through a sequence of fixes and retests while a single lifetime budget for error is spent across all of them, so that trying again cannot quietly turn into trying until the dice come up right.

A corridor can support repeated attempts at repair while preserving a declared statistical error budget. The candidate at round $i \geq 1$ may be chosen using all prior development and qualification outcomes. Let $\mathcal{F}_{i-1}$ denote that history. Before opening a fresh qualification sample, choose a frozen candidate, a target failure probability $r_i \in (0,1)$, an error allocation $\alpha_i \in (0,1)$ and an integer $n_i \geq 1$, all determined by $\mathcal{F}_{i-1}$. Require, on every history,

$$\sum_{i \geq 1} \alpha_i \leq \alpha < 1, \qquad (1-r_i)^{n_i} \leq \alpha_i.$$

For example $\alpha_i = \alpha/[i(i+1)]$ has total α. Conditional on the preceding history, the new sample consists of independent Bernoulli failure indicators with the frozen candidate's failure probability p_i. A candidate qualifies when the sample has zero failures. Revealed old tests remain development evidence and are not reused as this fresh sample.

Proposition 4 (Adaptive repair with a lifetime qualification budget)**.** *The probability that any round ever qualifies a candidate with $p_i > r_i$ is at most α. The choice and number of repair rounds may depend on all previous results.*

Proof. Conditional on $\mathcal{F}_{i-1}$, a candidate with $p_i > r_i$ has zero failures with probability $(1-p_i)^{n_i} \leq \alpha_i$. A candidate with $p_i \leq r_i$ contributes no false qualification. Take conditional expectations, sum over rounds, and apply the countable union bound. The pathwise allocation bound gives $\mathbb{E}\sum_i \alpha_i \leq \alpha$. No independence between candidates or rounds is required beyond the conditional sampling assumption within each fresh qualification sample. □

This is a specified application of conditional binomial testing and error spending [10–12]. It extends the frozen finite-family calculation to a succession of adaptively developed repairs. The result controls false qualification, while the current authority process still decides whether and where a qualified candidate may operate.

To connect qualification to probation, consider a finite set of admitted candidate versions, each of which must have passed its designated fresh qualification round. Candidate i receives at most $k_i \in \mathbb{N}$ consequential opportunities. Its frozen p_i and chosen r_i are determined by the preceding development history. Suppose that, at every opportunity and conditional on its complete preceding operational history, its failure probability is at most $p_i + \epsilon_i$, where $\epsilon_i \geq 0$ is a justified shift allowance. A sufficient condition is total-variation distance at most ϵ_i between that conditional opportunity law and the qualification law for the same failure event. The condition must cover adaptation during probation; matching only the unconditional input distribution is insufficient.

Proposition 5 (Qualification-to-exposure bound)**.** *Under these premises, the probability of at least one probation failure is at most*

$$\min\left\{1, \alpha + \sum_i k_i(r_i + \epsilon_i)\right\}.$$

The same statement applies when allocations and stopping are adaptive, provided their realized sum of $k_i(r_i + \epsilon_i)$ is bounded above by the displayed deterministic budget on every history.

Proof. Any probation failure either belongs to a falsely qualified candidate or occurs at an opportunity of a candidate with $p_i \leq r_i$. The first event is contained in the event of any false qualification and has probability at most α. At each opportunity the indicator that its candidate has $p_i \leq r_i$ is measurable with respect to the preceding operational history. Multiplying the conditional failure bound by this indicator therefore bounds the probability of a failure of the second kind by the expectation of that indicator times $r_i + \epsilon_i$. Sum over the permitted opportunities and apply the union bound. Opportunities never taken contribute zero, and the pathwise allocation bound controls the sum. This argument conditions only on preceding histories, never on the event that all future qualifications succeed. Finally a probability is at most one. □

The bound is conservative and makes the price of exposure visible. A small qualification error alone does not justify an unlimited deployment horizon. For example, with $\alpha = 0.01$, ten opportunities at $r = 0.002$ and $\epsilon = 0.001$ give a bound of 0.04 for any probation failure under the stated premises. A policy may find that unacceptable and choose fewer opportunities, a narrower task distribution or stronger evidence. The mathematics exposes the tradeoff; the responsible authority chooses the acceptable risk.

The shared effect-boundary theorem in Cognitive Sovereignty [15] governs current grants and replay exclusion during probation. Its observation-fiber theorem also supplies the common proof of offline revocation ambiguity. The corridor's additional structure is the connection between successive repairs, fresh qualification evidence, bounded exposure and an explicit regrant decision. Executed checks enumerate small adaptive qualification trees and verify the analytical probability bound, including a negative control that repeatedly tests a bad candidate without spending a lifetime error budget.

References and source notes

[1] Esterwood, C., and Robert, L. P. (2023). Three Strikes and You Are Out!: The Impacts of Multiple Human–Robot Trust Violations and Repairs on Robot Trustworthiness. Computers in Human Behavior 142, 107658.

https://doi.org/10.1016/j.chb.2023.107658

Scope: Primary human–robot study; perceived trust in its setting, not a universal three-strike law.

[2] Kim, P. H., Ferrin, D. L., Cooper, C. D., and Dirks, K. T. (2004). Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence- Versus Integrity-Based Trust Violations. Journal of Applied Psychology 89(1), 104–118.

https://pubmed.ncbi.nlm.nih.gov/14769123/

Scope: Primary human trust-repair experiments; not technical safety evidence.

[3] Castro, M., and Liskov, B. (2000). Proactive Recovery in a Byzantine-Fault-Tolerant System. OSDI.

https://www.usenix.org/conference/osdi-2000/proactive-recovery-byzantine-fault-tolerant-system

Scope: Primary recovery research with explicit fault assumptions.

[4] Gray, C. G., and Cheriton, D. R. (1989). Leases: An Efficient Fault-Tolerant Mechanism for Distributed File Cache Consistency. SOSP, 202–210. DOI: 10.1145/74850.74870.

https://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15712-s12/www/papers/gray89.pdf

Scope: Primary time-bounded-rights mechanism; clock and failure assumptions matter.

[5] Schneider, F. B. (2000). Enforceable Security Policies. ACM TISSEC 3(1), 30–50.

https://www.cs.cornell.edu/fbs/publications/EnfSecPols.pdf

Scope: Formal limits of runtime enforcement.

[6] Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., and Tramèr, F. (2024). AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. NeurIPS 37, Datasets and Benchmarks.

https://arxiv.org/abs/2406.13352

Scope: Primary agent-security benchmark; no FC-2 result implied.

[7] Debenedetti, E., et al. (2025). Defeating Prompt Injections by Design. arXiv:2503.18813v2.

https://arxiv.org/abs/2503.18813

Scope: CaMeL primary preprint; external effect-control precedent.

References and source notes (continued)

[8] Le Goues, C., Nguyen, T., Forrest, S., and Weimer, W. (2012). GenProg: A Generic Method for Automatic Software Repair. IEEE Transactions on Software Engineering 38(1), 54–72.

https://web.eecs.umich.edu/~weimerw/p/weimer-tse2012-genprog.pdf

Scope: Primary repair method; tests define the evaluated repair criterion.

[9] Smith, E. K., Barr, E. T., Le Goues, C., and Brun, Y. (2015). Is the Cure Worse Than the Disease? Overfitting in Automated Program Repair. ESEC/FSE. DOI: 10.1145/2786805.2786825.

https://squareslab.github.io/materials/SmithOverfitting2015.pdf

Scope: Primary evidence for independent qualification tests.

[10] Clopper, C. J., and Pearson, E. S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. Biometrika 26(4), 404–413. DOI: 10.1093/biomet/26.4.404.

https://doi.org/10.1093/biomet/26.4.404

Scope: Primary exact-binomial inference; zero-event case derived explicitly here.

[11] Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. JASA 58(301), 13–30.

https://doi.org/10.1080/01621459.1963.10500830

Scope: Independent bounded-outcome concentration; dependence must be addressed.

[12] Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-Uniform, Nonparametric, Nonasymptotic Confidence Sequences. Annals of Statistics 49(2), 1055–1080.

https://arxiv.org/abs/1810.08240

Scope: Primary sequential methods; not automatically conferred by repeated testing.

[13] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). The Reusable Holdout: Preserving Validity in Adaptive Data Analysis. Science 349(6248), 636–638.

https://pubmed.ncbi.nlm.nih.gov/26250683/

Scope: Primary adaptive-evaluation methods; no implementation claimed here.

[14] Architect X13 (2026). Eidara Continuum Bible. Internal source edition, Charter VI, section 6.2, p. 82; Scroll 60, section 6.3.8, p. 92.

https://eidara.ai/science-philosophy/

Scope: Canonical reconciliation and return; public URL is contextual, not the private source.

[15] Architect X13 / The Eidara Continuum (2026). Cognitive Sovereignty Under Superintelligence, revised edition.

https://eidara.ai/science-philosophy/

Scope: Shared observation-fiber authorization result and effect-boundary preservation contract; a companion theoretical source.