Institutional Incentives and the Recurrence of Harm

Flyxion
Independent Researcher

November 2026

Abstract

When the same harmful conduct recurs after the people responsible have been removed, sanctioned, or replaced, an explanation that stops at those people is incomplete. This essay develops that observation into a working method, the replacement test, which asks what would still reward the conduct if its present agents were exchanged for others. The argument is prompted by a public social media post that attributes scapegoating, factional conflict, and the failure of remedies to a single adaptive process operating across minds, language, and institutions. The post is read here as a hypothesis about feedback among classification, group identity, and institutional reward, and its stronger claim, that the same process explains both the harm and every attempt to correct it, is examined for what it would predict. Drawing on work in institutional economics, organizational sociology, the study of accidents, and the sociology of deviance, the essay argues that rules, rewards, and distributions of authority can reproduce similar conduct across changes in personnel. The stronger claim that a corrective institution acquires a stake in the problem's continuation requires separate evidence that the response's own rules alter its choices in ways that help sustain the harm. It then specifies the observations that would distinguish a correction that interrupts such a loop from one that merely relocates blame, applies the test to a single documented case that illustrates recurrence under persistent conditions without establishing recurrence sustained by the correction, and states the limits of the replacement test itself. Engineered analogues from recent work on artificial agents are used, with stated reservations, to decompose the architecture of recurrence into compilation of conduct, loss of provenance, admission without falsification, and erosion of access to alternatives. Individual responsibility and structural explanation are treated as answers to different questions that must both be asked.

1 A Post and Its Claim

A public social media post, written in a register somewhere between manifesto and warning, argues that the usual targets of collective blame are misidentified. Wealthy people, political factions, abusers, grifters, and the many named types that circulate in online argument are described as the surface products of something else, a family of adaptive processes said to have evolved alongside the human species and to operate in the network of human minds, language, and behavior. These processes are denied the status of an entity or a mind. They are described as maximizers of threat and opportunity that arise when communities depart from intimacy, awareness, and shared concern. On this account the habit of pointing at a perpetrator population is the characteristic trace of the process, because every accusation recruits attention and resentment, produces an endless backlog of grievance that cannot be satisfied, and generates new factions in response. Even a name for the process, the post warns, would be taken up by it. The author concludes that people can address the situation only by understanding the process itself, and by inventing ways of learning together that run opposite to it.

The post is vivid and in places extravagant, and a reader may reasonably decline its metaphysics, which describes advancements as organs of the process and cultures as laboratories for its evolution. Yet two of its claims can be separated from that framing and examined on their own terms. The first is that accusing a category of people usually fails to explain the conditions that keep producing the harm in question. The second, which is stronger and more interesting, is that attempts at correction can become the means by which the conditions reproduce, so that the remedy and the disease share a single explanation. Both claims concern a question that institutional analysis has long treated seriously: why similar conduct recurs across changes in the people who carry it out.

This essay takes that question as its subject and leaves the post's cosmology aside. Its aim is to make the post's insight answerable to evidence. The method proposed is modest. If harmful conduct persists when the individuals involved are replaced, then something other than the particular individuals is helping to produce it, and it becomes possible to ask which incentives, rules, and distributions of authority are responsible and what would change if they were altered. The same method can be turned on responses to harm. If a corrective measure removes an actor but leaves the reasons and opportunities for the conduct untouched, the measure has not interrupted the process, and if the measure itself creates new reasons for the conduct to continue, it has become part of the process. What would be expected to differ in the world if a correction did interrupt a reproducing loop is the question that structures the second half of the argument.

2 Responsibility for Acts and Explanation by Category

The distinction on which everything else depends is between holding someone responsible for a specific act and treating membership in a category as the explanation for the conditions that produce acts of that kind. Public argument often runs the two together, and the post is right to insist on their separation even where its remedy for the confusion is disputable.

Responsibility for an act is a claim about a particular agent, a particular occasion, and the evidence that connects them. It supports consequences proportionate to what was done, protection for those who were harmed or are at risk, and the discouragement of similar acts through the expectation that they will be answered. Nothing in an analysis of structure removes this. A person who defrauded a pension fund, concealed a defect, or abused a dependent did those things, and a procedure that establishes the facts and imposes consequences does necessary work.

Explanation by category treats membership in a group as sufficient evidence about conduct, or treats the existence of the group as the reason a harm occurs. Girard's account of the scapegoat mechanism describes how such an accusation operates [6]. A community in which rivalry has spread discharges its tension on a victim whose guilt is collectively believed, the discharge restores a temporary order, and the mechanism works only so long as the participants do not see it as a mechanism. The categorical form has two weaknesses. A category is larger and more heterogeneous than the conduct attributed to it, so the accusation is false of many members, which is a matter of fairness. The positions and rewards that made the conduct possible also outlast the removal of the category, which is a matter of explanation and the one that matters here. A prosecution can establish who did what, restrain the person, and impose a penalty while leaving those circumstances intact. The point does not weaken the case for prosecution. It marks what prosecution cannot accomplish alone, and so the place at which a second question must be asked.

3 The Replacement Test

The second question can be given a form that is easy to state and difficult to answer.

Definition 1 (Replacement test)**.** Given a pattern of conduct attributed to a set of agents, imagine the agents exchanged for others who are comparable in the relevant respects and placed in the same positions, with the same rules, rewards, and distribution of authority. If the pattern is expected to recur, the explanation must account for what persists across the exchange, including the rules and rewards of the position and any relevant characteristics shared by its occupants.

The test does not assert that agents are interchangeable, and it does not deny that character, intention, and choice matter. It asks only how much of a pattern is carried by the occupants and

how much by the places they occupy. Where recurrence is expected, the test directs attention to three kinds of variable, and the limits section returns to characteristics that the occupants may share.

The first is incentive: what the position rewards and what it penalizes, including rewards that are informal, deferred, or unacknowledged. The second is rule: what the position permits, requires, and forbids, and how those provisions are interpreted and enforced in practice. The third is the distribution of authority: who can decide, who can review the decision, who can be held to account, and who bears the cost when a decision goes wrong. These variables are the standard objects of institutional analysis, and North's account of institutions as the humanly devised constraints that structure incentives in exchange is a convenient statement of why they carry explanatory weight [13].

The test is also a diagnostic of corrective measures. A measure that removes an agent without altering the three variables predicts recurrence, and recurrence after such a measure counts against the view that the agent was the sole source. A measure that alters the variables predicts a change in the pattern independent of who occupies the positions, and the pattern's persistence or disappearance after a change of personnel then provides evidence about whether the alteration worked.

4 How Institutions Reward Conduct

Several literatures supply cases in which conduct that organizations officially condemn is nonetheless reproduced by their own arrangements. Kerr's well-known analysis of reward systems describes the common situation in which an organization hopes for one behavior while rewarding another, so that the behavior it actually obtains is the one its reward structure selects [9]. The point generalizes. When what is measured and compensated diverges from what is wanted, replacement of the individuals who respond to the measure will tend to bring in others who respond in the same way, because the response is a function of the measure.

A related point concerns the distortion of indicators once they become targets. Campbell stated that the more a quantitative indicator is used for social decision making, the more it becomes subject to corruption pressures and the more apt it is to distort the processes it was intended to monitor [2], and Strathern's formulation of the same pattern for audit regimes, that a measure which becomes a target ceases to be a good measure, is widely cited [18]. In both treatments the distortion arises from the relation between the indicator and the reward, and it recurs across personnel because it is produced by that relation.

The study of accidents and organizational failure offers a second line of evidence. Vaughan's analysis of the decision to launch the Challenger shuttle shows how repeated acceptance of anomalies became normal within an organization whose schedules, reporting lines, and standards of evidence made such acceptance reasonable to those inside it [20]. Perrow's account of tightly

coupled systems argues that certain accidents are produced by the structure of the system and are to be expected over time regardless of who operates it [15]. Reason's discussion of the person approach and the system approach to error makes the practical contrast explicit: the person approach seeks the individual who erred and responds with blame and retraining, while the system approach assumes that errors recur in similar circumstances and seeks to alter the conditions [17]. The replacement test may be regarded as the person approach and the system approach placed in a single question.

Merton's study of the unanticipated consequences of purposive action belongs in the same group, since it names the ways in which action undertaken to produce an effect produces others that alter the situation in which the action was taken [12]. Taken together these works establish a general claim. Similar conduct can recur across changes in personnel because the rules, rewards, and distributions of authority that shape the conduct persist, and an explanation that omits them will mistake a feature of positions for a feature of persons.

5 How Corrections Become Reproduction

The preceding section concerns institutions that reproduce conduct they condemn. A parallel claim concerns institutions that respond to the conduct. The post's most striking suggestion is that responses may be among the means by which the conditions persist, and the structural literature supplies mechanisms by which this could occur without positing any hidden agent.

The first mechanism is the acquisition of a stake. A body created to address a problem acquires budgets, personnel, procedures, and a mandate whose justification depends on the problem's continuation. This need not involve cynicism. The members of such a body may sincerely wish the problem to end, and nevertheless the body's continued existence is better secured by a problem that persists at a manageable level than by one that is resolved, and the resources it needs are allocated in proportion to the evidence of need that it can present. The result is a mild but systematic pressure against interventions whose success would remove the justification for the body. Allocation in proportion to need can be a sensible way of matching capacity to the problem, and the mechanism operates only where the rule changes the body's choices, for example by favoring reporting practices or forms of intervention that sustain the problem over those that would reduce it.

The second mechanism is displacement of the target. Cohen's study of moral panics describes how societies designate folk devils, respond to them with intensified control, and in doing so amplify the deviance and the attention paid to it [5]. The target changes from episode to episode while the form of the response persists, which is the pattern the post describes when it observes that new blame targets are produced without end. Christie's argument that conflicts have been taken from the parties who own them and given to professionals, who process them according to institutional categories, adds that the response can remove from the people most

affected the chance to establish what happened and what would repair it [3].

The third mechanism is the substitution of punishment for inquiry. A response that identifies a culpable party and imposes a sanction produces a satisfying sense that the matter has been addressed, and the satisfaction reduces the demand for inquiry into the arrangements that produced the conduct. Becker's economic treatment of crime and punishment frames the response as a choice of expenditure on enforcement and size of penalty [1], and its limitation for the present argument is instructive: it takes the supply of offenses as responsive to expected penalties and leaves the determinants of the underlying opportunities to other variables. Where the response operates only on the penalty, the opportunities remain.

None of these mechanisms requires a unified process. Each is an ordinary feature of organizations and public argument, and each can be observed, measured, and in principle altered. What the post adds is the suggestion that they interlock, so that the response to harm tends to supply the conditions for further harm and the further harm tends to supply the justification for further response. A single mechanism can close the loop if the response helps sustain a condition that produces further harm and the renewed harm in turn sustains the response. Interlocking among the mechanisms would strengthen the loop without being required for it.

6 The Persistence of Rules Without Their Authors

Recent work on artificial agents supplies a simplified analogue of the replacement problem. The analogy does not establish how social institutions behave, since an executable workflow and an external memory differ from an organization in the interpretation, authority, and contestation they permit. Its value is narrower. These systems make it possible to observe, under controlled conditions, conduct that is extracted from particular episodes, installed in a persistent structure, and reproduced after the agents who generated it have stopped deciding.

Hou and colleagues describe ReTrace, a system that converts a written procedure and recorded agent runs into an executable workflow in which the original agent no longer selects the next action [7]. In the reported settings, the compiled workflows approach the completion rates of the agents from whose conduct they were derived, with lower production cost and much shorter runtime on the principal public benchmark. The authors also note that a workflow can reproduce recorded steps without establishing that the recovered conditions are correct. What survives replacement is therefore a decision structure assembled from prior conduct, and it can continue to select actions after its rationale has been forgotten. An institution likewise need not contain anyone who remembers why a practice began. The practice may have been compiled into forms, thresholds, default classifications, and distributions of permission, and new occupants reproduce it by following the locally applicable rule. Authorship is then no longer the mechanism of recurrence. The same example separates provenance from validation:

learning that a procedure descends from a past decision identifies a source without showing that the decision was justified or that the present case falls within its scope.

Lee and colleagues study the complementary failure that arises when the source record is discarded [10]. Agents often compress earlier trajectories into concise lessons and later fail on tasks they would otherwise have solved, because the lesson may be false, misapplied, or stripped of the part of the trajectory that would reveal its error. Their WikiTrace system links each lesson to a summary of its source episode and to the raw trajectory, so that the acting agent can descend from the rule to the record from which it was inferred, and it reduces without eliminating the resulting regressions. The institutional counterpart is a conclusion that circulates after the inquiry that supported it has been lost. The relevant contrast is not between an institution that remembers and one that forgets, but between a memory that preserves only an instruction and one that preserves a route from the instruction back to the events from which it was drawn. The latter is still fallible, but it retains a possibility of revision that the isolated rule removes.

Tang and colleagues ask when a lesson drawn from one episode should be admitted into persistent state [19]. Their framework treats a proposed update as a hypothesis with a stated scope, applies it provisionally, tests it on sampled cases to which it should transfer, challenges it with sampled boundary cases intended to expose failures, and compares it against an archive of previously successful behavior. Some harmful updates in the study were not wholly false but were stated too generally, and narrowing their scope retained some benefit while reducing the harm. Institutional disputes are often framed as a choice between retaining a rule and abolishing it, when the defect may lie in an erased boundary.

Huang and colleagues construct a policy-optimization setting in which REINFORCE increases expected reward while driving the probability of optimal responses toward zero [8]. In their construction, KL regularization to a fixed base-model reference does not prevent the collapse. The result is proved for simplified policy classes and is not a theorem about organizations, but it isolates a structure familiar from the literature on targets and indicators. A reform may improve its declared indicator while eroding the heterogeneity, discretion, dissent, or local practices that would have disclosed the indicator's inadequacy, and later evaluation then compares variants of the selected arrangement and not the arrangement with what it displaced. The institution becomes more successful by its own measure and less able to discover that the measure is wrong.

These cases are used here only to sharpen four questions that can be put to an institution, and they answer none of them. What conduct has been compiled into routine? What evidence remains available for reassessing it? What tests preceded its admission? Which alternatives remain possible after the institution's measure has improved?

7 The Stronger Claim and Its Difficulty

The post's strongest claim is that a single adaptive process accounts for both the harm and every effort to remedy it. As a hypothesis about feedback this is intelligible and may be partly true. As a general explanation it faces a difficulty that deserves direct statement.

An explanation earns its place by excluding outcomes. Popper's account of falsifiability makes the point in its familiar form: a theory that is compatible with every possible observation tells us nothing about which observation we should expect [16]. The post treats every accusation of a faction as a trace of the process, and treats every apparent advance, whether technical or moral, as a product of the same process designed to mutate and spread. A framework in which accusations count as confirmation, and advances also count as confirmation, has no outcome that would count against it. Its force in conversation derives from this flexibility, and its usefulness for investigation is correspondingly low.

A second difficulty is reflexive. The post warns that any name for the process will be used by it, and that blaming a group advances it. The hypothesis of an unobservable generator behind all blame is itself a way of locating the source of harm somewhere other than in the arrangements that can be examined. An accusation directed at a process that cannot be questioned, observed, or corrected shares with the scapegoat the feature of being unable to answer back, and it shares with the categorical accusation the effect of relieving attention from specific conditions. This is not to say the post is mistaken about its object. It is to say that the form of the explanation can reproduce the pattern it identifies, and an account of recurrent blame must be able to apply its own diagnosis to itself without exempting itself. Girard's analysis bears on the post's warning that a name for the process would be taken up by it. On Girard's account the scapegoat mechanism depends on its own concealment, so that recognition weakens it [6], whereas the post holds that recognition is captured and put to use. Which of the two holds is an empirical question, and the criteria below bear on it.

The way forward is not to reject the stronger claim but to render it testable by asking what it would require of the world. If the same feedback accounts for harm and remedy, then specific, observable relations should hold: remedies should tend to be followed by recurrence under conditions that can be stated, and there should be identifiable conditions under which remedies are not followed by recurrence. The second half of that requirement is the more demanding, and it is what the next section considers.

A partial rendering of the stronger claim is possible once the causal links are stated. Take a count of recorded violations that determines both a unit's performance rating and the budget of the office charged with enforcing against violations. Three quantities must be kept apart: the underlying rate of violations, the rate at which they are detected and reported, and the recorded count that drives ratings and budgets. The count can move through changes in detection or reporting alone while the underlying conduct is unchanged, so any test must measure the

underlying rate by means that do not pass through the count, such as audits or outcome data. Keep the rating rule fixed and change how the office is funded. If the change alters the office's choices and is followed by a change in the underlying rate of violations, measured independently of the count used for ratings and budgets, that sequence would support a causal path from the response to subsequent conduct, and a decline would support the more specific claim that the prior response helped sustain the harm. The count may still shape both the unit's conduct and the office's decisions, since a shared input and a feedback path can operate in the same institution. The two specified models differ in what they allow. A common-cause model attributes the association between the harm and the response to conditions that influence both and permits no causal path from the response to later conduct, while a coupling model permits such a path. An intervention on the response alone can refute the common-cause model, but it cannot rule out a common cause operating alongside feedback, and it leaves open how many processes are involved, since a single adaptive process can contain feedback.

This rendering is partial. It specifies the variables and the contrast but supplies no data, and identifying a shared condition in a real case is itself a demanding piece of inquiry. It is offered as the form an answerable version of the claim would take, and the main contribution of the essay remains the evaluation of corrections.

8 What an Interruption Would Look Like

If a correction interrupted a loop in which conduct, response, and further conduct sustain one another, then a number of observations would be expected. Some are outcomes, namely the measured recurrence of the conduct and its persistence. Others are safeguards, namely features of the response that bear on whether any decline can be trusted and whether the result can be corrected if it fails. The two differ in evidential status. A reform can reduce recurrence while leaving poor records, and a well-documented reform can leave recurrence unchanged, so the absence of an outcome is evidence that the correction failed, while the absence of a safeguard is evidence about the reliability and correctability of the result.

Criterion 1 (Recurrence after replacement)**.** Recurrence of the conduct among new occupants of the relevant positions should fall, and the fall should persist after the attention that accompanied the correction has moved elsewhere. Heightened scrutiny is itself a change in enforcement conditions, so a decline during that period is consistent with deterrence. The relevant question is whether the decline persists under the conditions the reform can sustain once the scrutiny that accompanied it has relaxed to its ordinary level.

Criterion 2 (Traceable change in conditions)**.** The decline should be traceable to an identifiable change in incentives, rules, or distributions of authority, such that the change explains why new occupants behave differently. A decline with no such trace invites the suspicion that it

reflects measurement, concealment, or displacement of the conduct into forms the correction does not observe.

Criterion 3 (No stake in recurrence)**.** The institution responsible for the correction should not be placed so that its resource, mandate, or reporting rules reward choices that sustain the conduct. Resources that scale with the volume of the problem may reflect a sensible allocation of capacity, and the scaling alone does not show reproduction. What would show it is that the rule changes the institution's choices in ways that help sustain the harm, for instance by rewarding a reporting practice or by discouraging an intervention that would reduce recurrence. A resource rule that ties future capacity or organizational survival exclusively to the recorded volume of violations rewards demonstrable throughput and not reduction in the independently measured underlying rate. Whether it encourages continuation of the problem depends on whether officials can affect the recorded volume, the underlying conduct, or the choice between interventions that process violations and interventions that prevent them. The rule in the earlier violations example, which ties the office's budget to the recorded count, is one instance of this class.

Criterion 4 (Revision of the explanation)**.** The account given of the harm should change as findings accumulate. A correction that repeatedly produces the same explanation, including in the face of evidence that the explanation has failed to predict, is functioning as a ritual and not as an inquiry. Meadows's discussion of leverage points in systems makes a related observation, that interventions fail most often by acting on parameters while leaving untouched the rules and goals that structure the system [11].

Criterion 5 (Preservation of provenance and alternatives)**.** A correction should preserve a reviewable route from its standing rules to the evidence and decisions from which they were derived, and it should not erode access to the alternative practices against which its success could be assessed. A rule whose source can no longer be inspected may persist after the conditions that justified it have disappeared. Earlier standards and the evidence behind them should remain available for review, with departures explicitly justified so that successive exceptions cannot silently become the governing rule. A reform that improves its preferred measure while making competing arrangements unavailable has weakened the possibility of discovering its own error.

Criterion 6 (Absence of target displacement)**.** The correction should not be followed by the appearance of a new designated population to which the same charges are transferred. The substitution of one folk devil for another, with the form of the response unchanged, is a signature of reproduction.

The criteria are not an arbitrary list. Recurrence after replacement is the principal outcome: it asks whether the conduct actually declined and whether the decline persisted. The remaining

criteria concern either the causal route by which that outcome was produced or the safeguards by which it can be trusted and revised. Absence of a stake in recurrence addresses the possibility that the responding body benefits from choices that sustain the problem, while absence of target displacement asks whether the response merely transfers its form to a new population. Because a single mechanism is sufficient to close the loop, these criteria can be evaluated individually and do not require all three mechanisms to operate together. A traceable change in conditions connects the decline to an alteration in the arrangements rather than to concealment or temporary observation. Revision of the explanation keeps inquiry responsive to findings, and preservation of provenance and alternatives keeps that inquiry reopenable. A temporary decline under exceptional scrutiny may satisfy the outcome measure during the scrutiny without showing that the institution has produced conditions under which the decline will persist.

These criteria can be applied to cases that vary in how they came out. The comparative question that the post leaves unaddressed is why the loop takes hold in some settings and fails to take hold in others. Several candidate answers follow from the preceding sections and can be investigated: the degree to which the responding body depends on the problem for its justification, the extent to which affected parties retain a role in establishing what happened, the presence of reviewable and published findings that others can check, and the existence of a standing practice of asking whether a past response reduced the harm. Ostrom's studies of long-enduring arrangements for governing shared resources supply cases in which sanctions did operate without becoming self-sustaining [14]. The arrangements she describes tend to have monitors who are accountable to the people they monitor, sanctions that rise gradually with the seriousness of the offense, and inexpensive arenas for resolving conflict. These features limit the institutional rewards available from escalation and preserve for those affected a role in establishing what occurred, and several of them correspond to criteria stated above. They suggest that the difference between a loop that takes hold and one that does not lies in the design of the response and not in the goodwill of its operators.

9 The Test Applied to a Single Case

A pattern from the literature already cited can illustrate how the three kinds of variable are located and what evidence a genuine application would require. Vaughan's analysis of the Challenger launch decision describes an organization in which anomalies in a shuttle component were repeatedly redefined as acceptable risk, so that each acceptance became a precedent for the next [20]. The pattern to be tested is that acceptance of deviation. Replace the engineers and managers who took part with comparable ones placed in the same positions, and ask whether the pattern would be expected to recur.

The test directs attention to three kinds of variable. On incentive, schedule commitments and the costs of delay made continued operation institutionally easier than a finding that

required interruption, although Vaughan's account does not reduce the decision to schedule pressure. On rule, evidence from earlier flights was repeatedly interpreted as showing that observed deviations remained within an acceptable range, so that acceptance of anomalies altered the effective standard against which later anomalies were judged. On authority, the structure of review determined who could halt a launch, which concerns reached senior decision makers, and what showing was required to displace the presumption of continued operation. If comparable occupants inherited these conditions, recurrence would be intelligible without attributing it to the character of the original decision makers. The case does not show that a reader who accepted the three variables would have predicted recurrence in advance. It shows that recurrence occurred and that the variables can be located in the account of it.

Recurrence is what the Columbia Accident Investigation Board reported in 2003, finding that the organizational causes of that accident echoed those of Challenger after a response that had included investigation and reform [4]. Read through the criteria, the case shows recurrence after reform, but the first criterion concerns new occupants of the positions, and because both accidents occurred within the same agency and program lineage, recurrence after replacement is unresolved. The second is in doubt, since the finding suggests that the conditions which made acceptance of deviation rewarding had not changed enough to matter. That Columbia echoed Challenger is itself an interpretive claim about continuities in conditions and not a controlled demonstration, and readers of the safety literature may weigh it differently. The case also raises the fifth criterion as a question. A standard that shifts through successive decisions can still have a complete documentary history, so the issue is whether departures from earlier standards were explicitly justified or accumulated as exceptions that silently became the governing rule. Loss of provenance would require evidence that reviewers could no longer reconstruct the basis of those standards.

The case has a precise and limited role. It illustrates recurrence under persistent conditions: organizational patterns implicated in one failure recurred after an investigation and subsequent reforms. It does not show recurrence sustained by the correction itself, which requires an additional causal claim, namely that some feature of the response helped maintain or strengthen a condition that produces the harm. The evidence summarized here establishes recurrence and identifies plausible continuities, and it leaves that link open. Nor can the case settle between structure and selection, because both accidents occurred within the same agency and program lineage, and the relevant personnel and cultural carriers were not independently reassigned in a design that would separate occupant effects from regime effects. What the essay's two observations would require is a comparison with organizations that share the incentive structure but draw on different pools, and with the same people placed under a different regime. The test does not decide the case. It states what evidence would.

10 Limits of the Replacement Test

The test should be applied with an awareness of what it cannot establish. Replacement is seldom random. Organizations select, train, and socialize their members, so that the people who enter a position are often drawn from the same pool and shaped by the same preparation as those who left. Recurrence after replacement may therefore reflect continuity in selection and culture and not the incentives and rules the test is designed to isolate. Culture is carried by persons, and the line between a feature of positions and a feature of the people who staff them is not sharp.

Two kinds of observation would separate persistence through incentives from persistence through selection and socialization. The first is recurrence despite a deliberate change in the pool from which occupants are drawn, such as a change in recruitment criteria, training, or provenance. If the pattern persists when the pool differs, selection is a poor explanation and the positions are the more likely carrier. The second runs the other way: the form of the conduct varies with the incentive regime while the people are held constant, as when the same individuals are moved to positions governed by different rewards and rules and their conduct changes accordingly. Persistence across a changed pool together with variation across a changed regime points to structure, while persistence only where the pool is unchanged points to selection. Neither observation is easy to obtain, since pools and regimes seldom change independently, and the combination of the two is the hinge on which the replacement test turns.

The test also depends on the quality of the comparison. A claim that new occupants are comparable to the old can be difficult to substantiate, and a claim that conditions are unchanged is often contested. Cases in which the conduct is hidden are especially troublesome, because recurrence that is not observed is not counted, and a correction that increases concealment may appear to succeed. The evidence required to apply the test with any rigor is longitudinal and typically scarce.

Finally, recurrence does not by itself identify which variable is responsible. Incentive, rule, and authority are interdependent, and a pattern that persists across replacement is consistent with several different arrangements. The test marks the place where explanation should look. It does not complete the explanation. These limits are an argument for the careful use of the test and for combining it with direct investigation of the arrangements in question, and they apply with equal force to structural accounts that claim more than they can show.

11 Responsibility and Structure Together

The position defended here does not lower the standing of individual responsibility. It holds that two questions are being asked when harm recurs, and that they have different answers. The question of who is responsible for a given act has an answer that depends on the facts of the

act and the agent. The question of why such acts keep occurring has an answer that depends in part on the conditions in which agents act, and it remains open after the first question has been settled.

Treating the second question as a replacement for the first yields an excuse, since it implies that no one answers for what they did. Treating the first as a replacement for the second yields the scapegoat, since it implies that the removal of a culpable party resolves a pattern that the removal does not touch. A community that holds both questions together can impose consequences on those responsible and, in the same period, examine the incentives, rules, and distributions of authority that supplied the occasion. The post's warning, translated into this vocabulary, is that the first question tends to absorb the attention that the second requires. That tendency is real, and it is a feature of how public argument and institutional procedure are organized. It can be counteracted by practices that keep the second question in view: publishing findings in a form others can check, asking after the fact whether the response reduced the harm, and attending to what would still reward the conduct once the present occupants are gone.

12 Conclusion

A public post asserts that the real source of recurrent harm is never the people who are accused, and that attempts at correction are among the ways the source reproduces itself. The strongest part of this claim survives when it is detached from the cosmology that surrounds it. Conduct can recur across changes in personnel because the incentives, rules, and distributions of authority that shape it remain, and correction can reproduce conditions it aims to end when the responding institution's own rules reward choices that sustain the problem, when its target shifts without its form changing, and when punishment takes the place of inquiry. A correction that leaves an existing cause in place differs from one that helps perpetuate it, and only the second is reproductive in the sense used here. The weaker part, the assertion that a single adaptive process accounts for every outcome, cannot be assessed until it is stated in a form that excludes something, and the criteria and the partial rendering offered above are an attempt to provide such a form. The rendering separates a common cause from feedback from the response and shows what an intervention on the response alone can and cannot establish, while leaving open how many processes are involved. A correction that interrupts a loop would show persistent decline in recurrence after replacement, a traceable change in conditions, an institution whose resource and reporting rules do not reward choices that sustain the problem, an explanation that is revised by evidence, and no new population designated to carry the old charges. An interruption must also preserve the conditions of its own corrigibility: standing rules should remain traceable to the evidence from which they were derived, proposed lessons should be tested before they become persistent, and improvement under a selected measure

should not be purchased by eroding access to the alternatives against which that measure could be judged. A correction that shows none of these has found a new name for the party to blame. The replacement test directs the question to the right place, and it leaves to investigation the work of determining what the answer is.

References

[1] Becker, G. S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy*, 76(2), 169–217.

[2] Campbell, D. T. (1979). Assessing the Impact of Planned Social Change. *Evaluation and Program Planning*, 2(1), 67–90.

[3] Christie, N. (1977). Conflicts as Property. *British Journal of Criminology*, 17(1), 1–15.

[4] Columbia Accident Investigation Board. (2003). *Report of the Columbia Accident Investigation Board*, Volume I. Washington, DC: U.S. Government Printing Office.

[5] Cohen, S. (1972). *Folk Devils and Moral Panics: The Creation of the Mods and Rockers*. London: MacGibbon and Kee.

[6] Girard, R. (1986). *The Scapegoat*. Translated by Y. Freccero. Baltimore: Johns Hopkins University Press.

[7] Hou, K., Kuo, Y. P., Bhattacharya, I., Nama, R., Gupta, A., Castillo, C. D., Ramachandra, P., Jiang, X., and Carbone, F. (2026). ReTrace: Distilling Tool-Using Agents into Planner-Free Workflows. *arXiv preprint arXiv:2611.14492*.

[8] Huang, A., Jiang, N., Kakade, S. M., Malladi, S., Krishnamurthy, A., and Foster, D. J. (2026). A Theory of Policy Optimization in Post-Training: Coverage Collapse and Reference Resets. *arXiv preprint arXiv:2611.14467*.

[9] Kerr, S. (1975). On the Folly of Rewarding A, While Hoping for B. *Academy of Management Journal*, 18(4), 769–783.

[10] Lee, Y., Choi, Y., Baek, J., and Hwang, S. J. (2026). Don't Just Trust Your Memory: Evidence-Linked Memory Reduces Memory-Induced Degradation in LLM Agents. *arXiv preprint arXiv:2611.14320*.

[11] Meadows, D. H. (1999). *Leverage Points: Places to Intervene in a System*. Hartland, VT: The Sustainability Institute.

[12] Merton, R. K. (1936). The Unanticipated Consequences of Purposive Social Action. *American Sociological Review*, 1(6), 894–904.

[13] North, D. C. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

[14] Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

[15] Perrow, C. (1984). *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books.

[16] Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

[17] Reason, J. (2000). Human Error: Models and Management. *BMJ*, 320(7237), 768–770.

[18] Strathern, M. (1997). Improving Ratings: Audit in the British University System. *European Review*, 5(3), 305–321.

[19] Tang, R., Gao, H., Ding, Y., Yao, J., Zheng, Z., Sheng, Z., and Hou, L. (2026). When Should Agents Remember? Falsification-Gated Self-Evolution for LLM Agents. *arXiv preprint arXiv:2611.14326*.

[20] Vaughan, D. (1996). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. Chicago: University of Chicago Press.