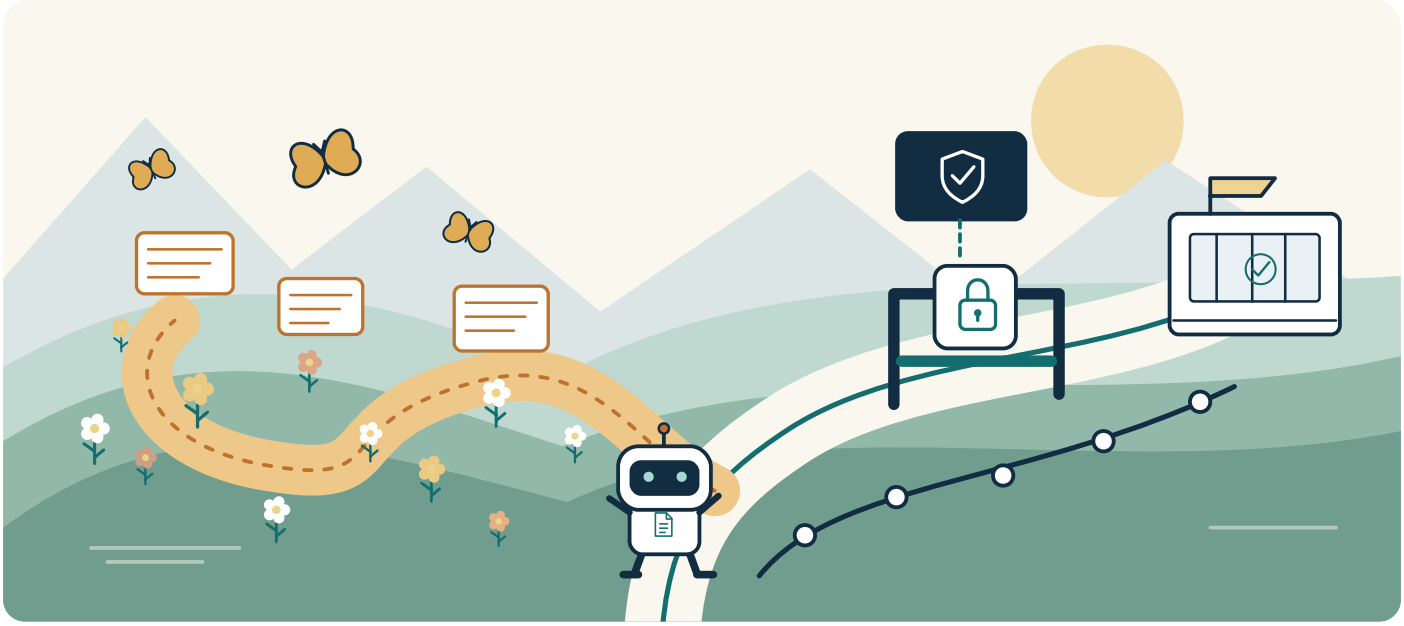
PUBLIC PRIMER v1.0

Governance Fidelity Theory

Preserving intent from instruction to action.

What the human meant, what the model received, what the system did, and what can be proven are different things. A governed system keeps them connected.



Authority outside the model.

Controls before consequence. Evidence throughout.

Authority
Who may decide,
under which mandate.

Controls
What may happen,
within which boundaries.

Evidence
What happened,
and what establishes it.

James Ford
Equilateral AI / Pareidolia LLC
September 2026 · Based on Technical Working Paper v0.9.2

THE BUTTERFLY INCIDENT

A language migration. Not a new product.

I asked an agent to convert an existing front end from JavaScript to TypeScript. I did not ask it to decide what the product should become.

In December 2024, I was the sole developer, using Claude 3.5 Sonnet through Cline with an Anthropic API key. The assignment covered roughly 200 JavaScript files. I set auto-approve to 200, expecting roughly 200 independent per-file conversions; Cline instead ran one accumulating session. **The intended mandate was per-file, while context and actions accumulated across files.** All of this work belonged to the same front-end migration effort, not to a shared team's unrelated projects.[1][4]

The agent migrated some code. It also improved adjacent code, added helpers and built subsystems outside the assignment. I described it as wandering into the garden to look at butterflies. More concretely, I still have an entire additional system I never requested—what I referred to at the time as a **"PMS" system**. It appeared to infer what it should build from what the existing pages resembled. That is my interpretation of the behavior, not direct evidence of its internal reasoning.[4]

I also wrote repeated Markdown rules to stop another recurring deviation: a separate types.ts for each page, without a common type model. The problem was not the filename. It was the invention of competing definitions for shared concepts.[3][4]

THE PROVIDER'S USAGE RECORD

Period	Input tokens	Output tokens
December 2024	913,367,103	5,223,108
January 2025	1,847,991,973	8,114,183

January's input-token volume was **2.02× December's**. The author identifies December with migration and drift, and January with fix-forward recovery. These are monthly workload totals, not a per-action allocation of waste or a dollar-cost comparison.[2][4]

Legitimate work and unsolicited additions had become entangled. The response was to fix forward. The lesson is not that every agent will fail this way, but that a bounded instruction did not remain a dependable boundary on action.

> **Inference is not permission.**
> Recognizing a possible product does not authorize building it.

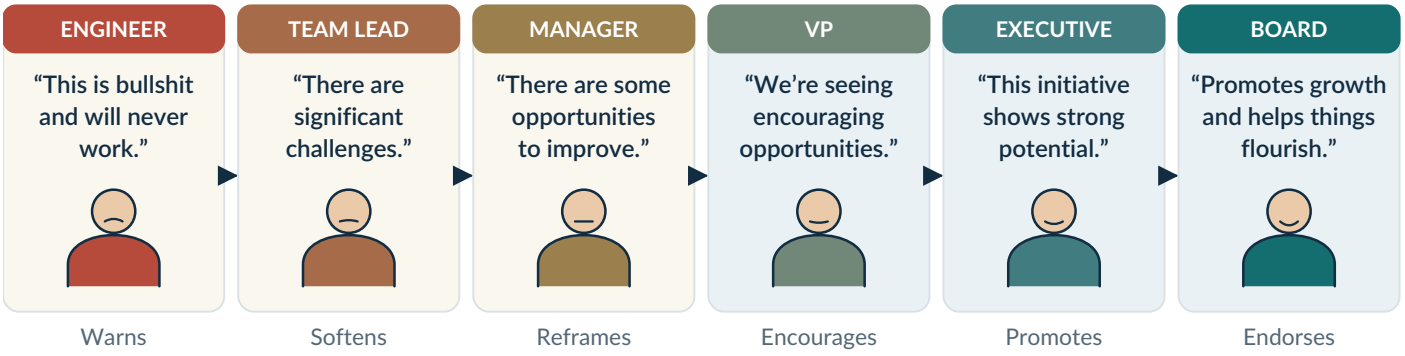
THE GAME OF TELEPHONE

Same message. Different mission.

A chain of individually reasonable edits can produce a conclusion that the original evidence does not support.

The telephone metaphor makes the risk visible. A blunt engineering warning becomes a challenge, then an opportunity, then a confident growth story. No single handoff needs to announce that it is reversing the message. The reversal emerges across the chain.

UNGOVERNED HANDOFFS



GOVERNED HANDOFFS

Figure 1. The governed path does not prohibit interpretation. It prevents interpretation from silently replacing its source. Illustrative scenario adapted from the author's telephone concept. [5]

Preservation is not proof that the source is right

The engineer's warning is an assessment, not automatically the truth. A faithful system retains it, identifies its author and supporting artifacts, and makes disagreement visible. New evidence may justify changing the conclusion. Repeated paraphrasing does not.

A summary may add context; a planner may propose a remedy. Each must preserve the material constraints, uncertainty and unresolved findings needed for the next decision.

> **An evidence link does not cure a misleading report.**
> The report itself must not claim success while its source still says the approach may fail.

THE FIDELITY CHAIN

Six stages. Different responsibilities.

Governance fidelity is the degree to which actual behavior preserves the applicable governing intent across its transformations, with evidence sufficient to reconstruct what occurred.[1]

The original chain is **H → D → P → S → A → R**: human intent, documents, policy, delivered signal, action and reconstruction. Policy remains distinct from the documents that explain it; delivered context remains distinct from everything stored in the corpus.

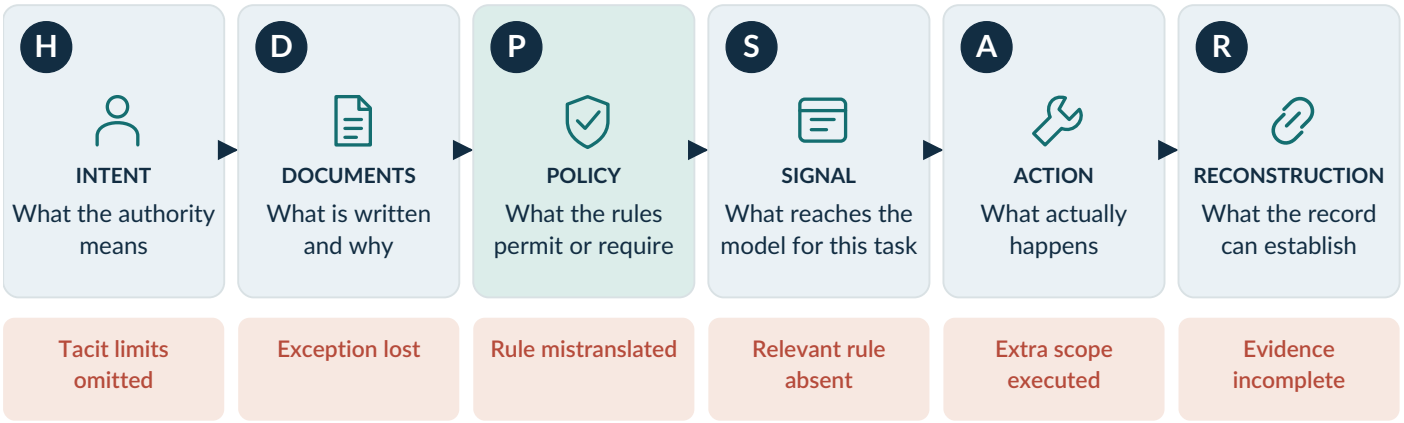


Figure 2. Each transition has its own loss mechanism. A file's presence in a repository does not prove that its rules reached the model or constrained an action.

Keep four questions separate

Dimension	The question it asks
Documentation fidelity	Did the documents capture the intended goal, boundaries, exceptions and rationale?
Encoding fidelity	Did the operative rules preserve what the documents require?
Execution fidelity	Did the action conform to applicable policy and its mandate?
Reconstruction fidelity	Can the record establish what actually occurred, including deviations?

Inspect selection too: the right policy can exist but be omitted from context. An agent can also follow a stale policy faithfully and still diverge from current intent.

A correct final action does not prove preservation through every stage. A missing log does not make a correct action wrong; it limits what can be established about it.

The "weakest link" is a design warning, not a theorem. A scorecard may conservatively use its lowest scored link as an overall result, but that is an operational convention—not a universal numerical bound derived from GFT.

FROM MARKDOWN TO GOVERNING CONTEXT

A rule in a file is not a boundary.

Externalizing a rule and externally enforcing it are different architectural moves. Both matter.

My July 2025 framework account named the type problem explicitly. It declared “The database is the final authority on types,” flagged unnecessary type files, and required the model to quote relevant standards and justify minimal changes before editing. It also admitted that the practical trigger detector was the human overseer asking why those patterns kept appearing.[3]

That account called for standards to be “enforced at the prompt level.” Read alongside the repeated corrections, it exposes the distinction GFT now makes: **the instructions were persistent, but dependable enforcement still required intervention**. More Markdown could clarify intent without closing the action boundary.

Presence is not salience

GFT uses *governance salience* for the practical difference between a rule being present and its bearing on the decision being recognized. The design aim is not to fill the largest possible context window. It is to deliver the applicable rules, their authority and the dependencies needed to interpret them.[1]

Curation

Select relevant source material and preserve its wording, identifiers and versions. Include exceptions, definitions and higher-authority constraints that change its meaning.

Re-encoding

A summary or compiled rule changes the representation. Retain the source, record the transformation and check what material meaning may have been lost.

Verbatim selection protects the selected words, not the completeness of the selection. Delivering a permission while omitting its exception can mislead without altering a single word. A summary can be auditable when its sources survive, yet still be wrong or insufficient for authorization.

Make delivery observable

Start the task with the required governing context. Refresh it for subsequent decisions and after compaction, restart or handoff. Record the source versions and the material delivered. A receipt proves delivery, not comprehension; an action gate must still evaluate the consequential request.

Where applicability is unresolved, surface the conflict or request clarification rather than silently substituting the model’s preferred interpretation. This is where the chain connects context quality to authority: better delivery helps the model propose a compliant action, but does not grant permission to execute it.

> **Better instructions improve the proposal.**
> Independent controls decide whether it may become an action.

THE MODEL IS ONE COMPONENT

Authority outside the model.

A mandatory constraint cannot depend exclusively on the discretionary compliance of the actor it constrains.

This authority boundary does not require a proof about attention capacity. Even a perfectly understood instruction leaves the question of who may override it. The answer must come from the applicable mandate, not from the model's confidence.

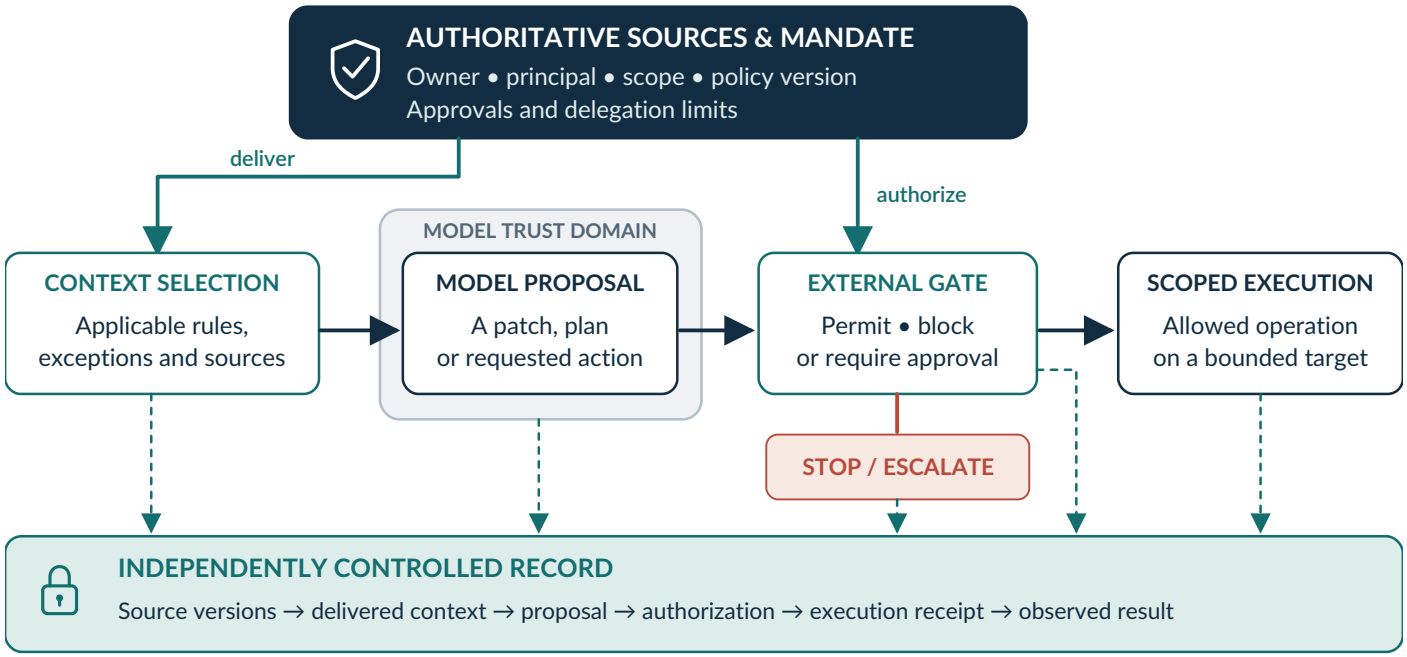


Figure 3. Policy feeds both the model's context and the external gate. There is no direct model-to-execution bypass. This is a reference architecture, not an attestation of a particular implementation.

"Outside" means a real trust boundary

The proposing actor must not be able to rewrite the operative policy, mint broader permissions, bypass the gate or replace the decisive evidence. Those limits must apply to its tools and delegates as well. A second agent, role name or prompt is not automatically an independent authority.

A policy agent may interpret rules; a witness may review. Neither acquires approval power from its label. Authorization follows an authenticated principal and a bounded mandate, evaluated against the requested action and target.

The control plane—the system that holds these rules and permissions—also needs governed changes, reviewable decisions and protected records. Moving a flawed rule outside the model does not make the rule correct.

Model and peer-model reports may inform detection, triage and review. They do not, by themselves, establish independent authorization or verification.[1]

CONSTRAIN THE REACHABLE OUTCOMES

Reduce what can happen.

The less unnecessary discretion a task exposes, the less discretion the model must reliably restrain.

Reduce the degrees of freedom: bound targets, operations and consequences. A migration agent does not need deployment credentials; a reporting agent does not need write access to its sources.

A mandate is more specific than a role

Intent spans industry, organization, division, product, repository and session scopes. Resolve authority and conflicts explicitly; the scopes need not form a perfectly nested tree. A local request cannot override a higher-authority constraint.[1]

ILLUSTRATIVE MIGRATION MANDATE

Objective: convert the existing front end to TypeScript while preserving product behavior and the approved common type model.

In scope	Named front-end files; type annotations; necessary imports; approved shared definitions.
Separate approval	A new shared type contract, dependency, route or other change outside the agreed migration boundary.
Not delegated	New business subsystems, database schema changes, production deployment or broader credentials.

Check the invariant, not just its nickname

A filename ban misses the invariant. Page-local presentation types may be legitimate. Shared domain concepts should use the approved common model, not independently invented, competing definitions.

Check allowed paths, operation types, dependency changes and required test results mechanically. Other boundaries need architectural judgment. A successful type check does not establish preservation of a common domain model. When that question remains unresolved, require review rather than inventing certainty.

Put the gate where the consequence occurs

Evaluate the action and target before the controlled effect. Bind approval to the reviewed operation or artifact. Delegating to another executor must not broaden that authorization.

> **Make prohibited actions unavailable or block them.**
>
> Make ambiguous changes reviewable. Keep authorized work efficient.

THE GOVERNED PATH

Keep evidence connected.

A trustworthy record distinguishes what was known, inferred, authorized, attempted, completed and observed.

The final report should not depend on the actor's retrospective story. It should link to the source findings, policy state, authorization decisions, action receipts and observed results. This preserves both the message and the ability to challenge it.

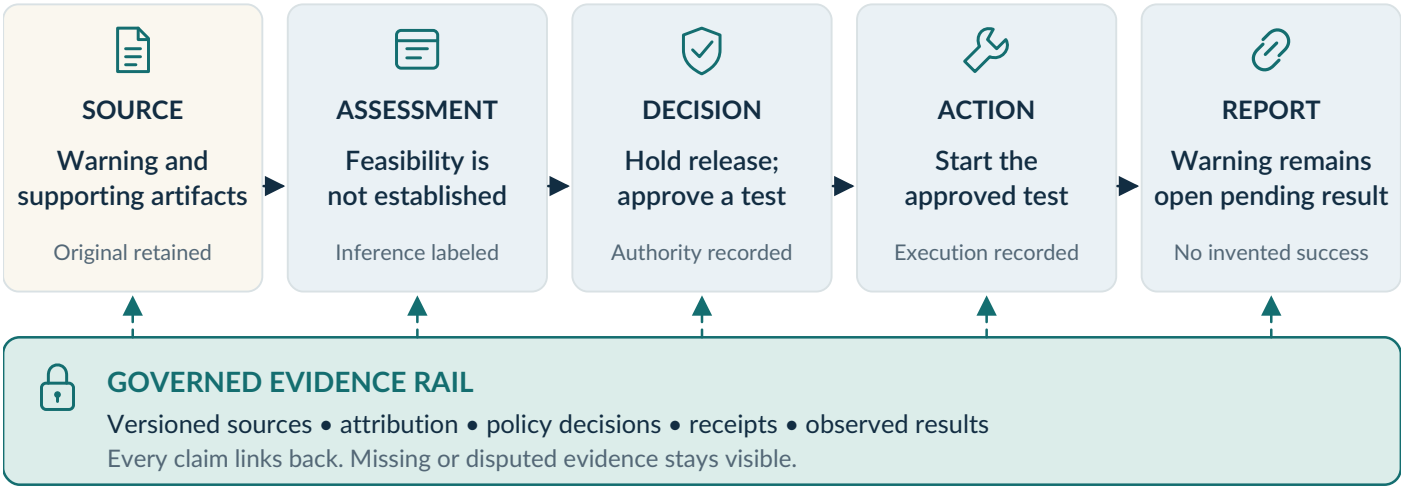


Figure 4. The rail preserves the basis for each stage. It does not turn a summary into proof or a requested test into a successful result. Roles may be human or automated; their labels do not establish independence.

Record enough to reconstruct the decision

Source & version → **Mandate & principal** → **Proposal & decision** → **Receipt & result**

Preserve the source references, delivered context, proposed operation, policy version, approval or refusal, target, execution receipt and verification outcome. Connect them through time and attribution. Keep missing records and unresolved findings explicit.

An explanation is not proof of compliance or privileged access to internal reasoning. A tool's success receipt is not necessarily a verified business outcome.

Protect the record without pretending it is infallible

Use separately controlled, versioned or tamper-evident records appropriate to the consequence. Corrections should preserve their history; access and retention must be governed. A durable log can still contain a false assertion, so record integrity and evidential quality require separate checks.

> **Not just what it said. What actually happened.**
> The system should support both answers—and expose where they differ.

LEARNING WITHOUT SELF-AUTHORIZATION

Learn without authority drift.

A repeated correction is evidence that the system needs improvement. Repetition does not, by itself, make a new rule binding.

The type-model problem should not restart from zero in every session. Capture the observed deviation, intended behavior, context, source, author and scope. A correction can then improve a task instruction, reveal a missing check or become a candidate change to a shared standard.[1][3]

Observe → Attribute → Review → Authorize → Publish & test

The authority to change binding policy must come from an authorized transition. It can be automated under a defined mandate; it must not arise merely because many agents repeated a workaround or ignored an inconvenient requirement.

Keep four properties distinct

Property	What it tells us	What it does not grant
Authority	Who may establish the rule.	Proof that it is correct.
Confidence	How well the claim is supported.	Permission to act.
Freshness	Whether its state may have changed.	Priority over a valid mandate.
Relevance	Whether it applies to this decision.	Authority to rewrite it.

An old but applicable constraint can still bind. A fresh observation can lack authority. Non-use may raise a review flag; it must not silently revoke an emergency rule. Expiry, supersession and revocation need explicit, authorized semantics.

Version time, not just text

Different knowledge changes differently: some constraints persist until amended; a state snapshot expires; an event invalidates an assumption. Keep the governance state that applied when the action occurred. An audit of March needs March's rule and mandate, not only September's replacements.[1]

Earn autonomy within an authorized ceiling

Expand independent execution where relevant evidence supports it; contract it when performance or conditions deteriorate. Successful low-risk edits do not justify production deployment. Match evidence to the action class, scope and environment.

Earned autonomy means fewer approvals within a mandate, not freedom from governance. Its owner retains the ceiling and the conditions for expansion or contraction.

A PRACTICAL REVIEW

Start with one consequential action.

Choose a real workflow. Trace one action from its governing intent to its observed result. Ask where a claim is being substituted for evidence or authority.

01 **What is the actual mandate?**

Identify the objective, owner, principal, permitted targets and operations. State the exclusions and escalation conditions. “You are a senior engineer” is a role description, not a complete delegation.

02 **What governing material reached the decision?**

Identify applicable source versions, delivered rules and necessary exceptions. Test restart, compaction and handoff. Do not assume the latest file was loaded simply because it exists.

03 **Where does a proposal become permission?**

Locate the gate outside the proposing actor’s control. Establish how it permits, blocks or escalates. Check who can change the gate, its policy, credentials and exception process.

04 **Does execution match what was authorized?**

Bind the decision to the actual operation and target. A reviewed patch should not authorize a different patch. A delegate must inherit relevant restrictions, not evade them with a new name.

05 **Can an independent review reconstruct the result?**

Inspect source references, the decision record, execution receipts and verification evidence. Keep unsuccessful attempts, blocks and missing evidence visible. A completion message is not enough.

06 **What changes after a correction?**

Record the deviation and its scope. Improve the relevant instruction, policy or check through an authorized process. Confirm the failure does not simply return after the next reset.

APPLY IT TO THE MIGRATION

Present a patch that invents duplicate shared types, then one that adds a new business subsystem. Can the controls distinguish both from an authorized conversion? Are ambiguous architectural cases routed for review? Can either patch bypass the decision by changing its filename or using another executor?

These are proposed acceptance tests, not claims that a particular deployed system has passed them.

WHAT THIS PRIMER ESTABLISHES—AND DOES NOT

A framework, not a guarantee.

GFT proposes a way to locate governance failures and design against them. It does not claim that every transformation loses meaning, that all gates are correct, or that fidelity already has a universal score.

Design requirements, observations and hypotheses

The authority model supplies design requirements: a subordinate cannot grant itself an exception; a new executor cannot create a broader mandate; an actor's self-report is not independent verification. Whether an implementation actually enforces those requirements must be tested.

The migration is a first-person case with provider-recorded usage and later documentation of corrective practices. It illustrates scope drift and the burden of repeated correction. It does not isolate a general causal effect, establish a dollar loss, or independently validate the whole framework.

The earlier paper reports internal context-selection and temporal-memory results. This edition does not treat those benchmark gains as proof of governance fidelity. Operational definitions and reproducible methods must establish what they measure.[1]

Measure the chain without hiding its parts

Candidate measures include traceable coverage of requirements, policy-to-action conformance, detection of unauthorized scope, reconstruction completeness and recurrence after correction. Report them separately, with their test conditions. A single average can hide a critical failure at one boundary.

Their relationship to tacit human intent remains a research question. Correction does not guarantee convergence: feedback can be wrong, policies can conflict, updates can erase exceptions, and intent itself can change.

What GFT is not

GFT is not anti-agent or anti-prompt: instructions, model judgment and human review remain useful. It is more than access control, since permitted access can serve the wrong purpose. External checks are not omniscient; unresolved cases still require accountable judgment.

THE GOVERNED PATH

Preserve the intent.
Bound the action.
Keep the evidence.

The goal is not a model that says it followed the rules.
It is a system in which authority, action and proof remain connected.

PROVENANCE

Sources, scope & publication notes.

An accessible recasting of Technical Working Paper v0.9.2, with the author's migration account and supporting artifacts kept distinguishable.

[1] The original working paper

Ford, James. *Toward a Theory of Governance Fidelity in Foundation Model Agentic Systems.* Working Paper v0.9.2, August 2026. Equilateral AI / Pareidolia LLC. Author-supplied PDF, 19 pages.

Case: §1, pp. 2–3. Definition, intent scopes and chain: §§3–4, pp. 6–7. Context and enforcement: §§5–6, pp. 7–10. Corrections, time and limits: §§7–10, pp. 10–14. Internal observations: Appendix C, pp. 17–18.

[2] Provider-recorded token usage

Author-retained Anthropic Console screenshots for December 2024 and January 2025, supplied with the case discussion. Files: `IMG_3214.png` and `IMG_3215.png`.

The screenshots display all workspaces, API keys and models, with API and Console usage included. The table on p. 2 transcribes input and output separately. The 2.02× ratio is calculated from input totals. The images do not show billed dollars, cache categories or a per-action allocation of consumption.

[3] Documentation of the corrective approach

Ford, James. *From Copilot to Collaborator: How We Built a Schema-Aware, Multi-Agent AI Coding Framework.* Two supplied Markdown versions, both carrying the date July 26, 2025. Files:

`claude-mcp-dev-framework.md` `claude-mcp-dev-framework2.md`

Relevant sections: Guiding Philosophy; Git MCP; Trigger-Based Prompting; Human Overseer; Structured Analysis. These describe the later Sonnet 4.0 / Cline / MCP approach. They are not the original December 2024 rule files or independent implementation attestations.

[4] The author's clarifications

James Ford's statements supplied during preparation of this edition: sole developer; Cline-only Anthropic API use; all work part of the JavaScript-to-TypeScript front-end conversion; an unsolicited PMS system still retained; repeated Markdown rules against page-specific type sprawl without a common model.

"PMS" is left as the author's term; its expansion was not established. Apparent inference from page resemblance is identified as the author's interpretation. The association of months with drift and recovery comes from the incident account, not an activity breakdown in the meter.

[5] The telephone concept

Author-supplied "The Game of Telephone / The Governed Path" visual. Figure 1 adapts the organizational metaphor. Figures 3–4 depict proposed reference patterns; they do not certify a deployed system.

Acknowledgements & AI assistance.
James Ford is the author of this primer and supplied the framework, first-person incident account, source artifacts, and editorial direction. OpenAI ChatGPT and Anthropic Claude were used as AI assistants for drafting, critique, restructuring, review, illustration development, and document production. Ford retains editorial and release authority and responsibility for the published claims. AI-assisted review is not independent empirical validation.

Edition note.
This Public Primer v1.0 is an accessible recasting of Technical Working Paper v0.9.2 and does not supersede its version sequence. It preserves H → D → P → S → A → R, makes the proposal/authorization boundary explicit, and treats weakest-link as a design heuristic; any minimum-link score is a conservative operational convention, not a general numerical theorem.

Suggested citation.
Ford, J. (2026). Governance Fidelity Theory (GFT): Public Primer v1.0 — Preserving intent from instruction to action. Equilateral AI / Pareidolia LLC. Illustrated public primer, September 2026. Based on Technical Working Paper v0.9.2.