GOVERNED AUTONOMY

MIRAGE: Authority, information, and consequences in autonomous systems

A paper in the Governed Autonomy collection

Andrew Durden
Independent research
Publication edition 2026.10.8

Research on authority, evidence, and continuity in autonomous systems.

Content identifier: c321294d8c141120daaaf471

All rights reserved · Research exposition

Contents

MIRAGE: Authority, information, and consequences in autonomous systems 3

Abstract . . . 3

1. The problem: capable reasoning with bounded consequences . . . 3

2. Five distinctions that govern the argument . . . 4

3. The consequential boundary includes observation . . . 5

4. Collaboration does not create additional authority . . . 5

5. Information release and the limits of general claims . . . 6

6. A durable decision is different from a completed effect . . . 7

7. Observation, provenance, and settlement . . . 8

8. Recovery must preserve more than content . . . 8

9. Master, native design, and reference model . . . 9

10. Four illustrative cases . . . 12

11. Relationships and unresolved adoption . . . 12

12. What research would resolve . . . 13

Source guide . . . 14

All rights reserved · Research exposition

MIRAGE: Authority, information, and consequences in autonomous systems

Governed Autonomy · Publication 2026.10.8
Research study: conceptual exposition of a proposed architecture and its reference model
Source: Public exposition, successor 2026.10.8; historical research and evidence separately qualified

Abstract

An autonomous system can produce a convincing explanation without possessing authority to act, and an authorized request can leave its actual consequence uncertain. MIRAGE studies the boundary between open-ended computation and accountable effects. Its central question is how an institution can allow capable, potentially adversarial agents to reason and collaborate while retaining control over what they may change, what information they may disclose, and which outcomes they may claim.

The proposed architecture treats authorization, information use, execution, observation, and settlement as related but distinct obligations. This perspective extends beyond an initial approval screen. It includes the environment through which an effect occurs, the information available to the agent, the consequences of delegation, and the state that survives interruption. A successful action must have an adequate basis at the relevant boundary; a successful recovery must preserve outstanding obligations as well as restored functionality.

This paper develops the argument through five distinctions, four illustrative cases, and a comparison of the master architecture, native design, and bounded reference model. The available evidence consists of specifications and reported reference experiments. Those results support claims about their stated models and inputs; qualification of the complete architecture, its hardware boundary, and its proposed product integration remains a separate undertaking. The final sections identify the evidence needed to advance those claims.

1. The problem: capable reasoning with bounded consequences

Open-ended reasoning is useful because an agent can discover steps that were not enumerated in advance. The same flexibility makes a list of approved surface actions a weak description of what a system may actually do. An apparently ordinary request can read sensitive information, create a durable artifact, contact another party, change a shared resource, or cause a later action. Its consequences depend on the environment and on the authority exercised through it.

MIRAGE begins from this distinction between computation and consequence. An agent may consider an action, simulate it, describe it, or propose it without thereby acquiring the power to make it authoritative. A persuasive explanation is evidence about the proposal's

All rights reserved · Research exposition

content. It is not an authorization instrument, an independent observation, or proof that the requested result occurred.

This separates two questions often collapsed in discussions of agent safety. One concerns the quality of the agent's reasoning: whether it understands a task, predicts consequences, and selects useful actions. The other concerns the institution's control over effects even when that reasoning is mistaken or adversarial. Improving the former does not eliminate the need for the latter. Conversely, constraining effects does not establish that the agent is competent or that an authorized policy is wise.

The collection thesis explains why a continuing institution needs this separation. It must retain obligations across changes in models, participants, and infrastructure. MIRAGE studies the computing boundary needed to make such obligations meaningful. Its subject is therefore broader than filtering a prompt, checking the wording of an answer, or logging an API call. The relevant question is whether the complete consequential path obeys its declared conditions.

The architectural claim spans several levels of evidence. A bounded executable example can clarify semantics, while evaluating the full boundary requires implementation and observations of the environment it is intended to govern. Public review scope.

2. Five distinctions that govern the argument

The first distinction is between **intention and authority**. A user or institution can establish an objective, but individual steps still need authority appropriate to their consequences. The objective may permit independent planning and ordinary execution while leaving a new recipient, a materially larger cost, or a different use of protected information outside its scope. A proposal does not acquire broader authority because it would make the objective easier to complete.

The second is between **authority and realization**. An authorized decision concerns what may or must happen under stated conditions. Realization concerns what actually happens at a resource or recipient. An upstream approval cannot control an independent provider merely by describing that provider's desired behavior. The strength of the claim depends on the actual point of control and the provider's own contract.

The third is between **realization and observation**. An effect can occur while its response is lost. A process can report success while the intended resource remains unchanged. Observation must address the particular result, rather than merely repeat the executor's assertion. Its timing, coverage, and dependence on other sources matter.

The fourth is between **authenticity and adequacy**. A genuine signature can establish who signed particular content. It cannot by itself establish that the content is correct, complete, current, or sufficient for the decision. A measured software state can be useful evidence without determining whether a proposed business action is authorized.

The fifth is between **historical evidence and present support**. An earlier approval or observation can remain a valid historical fact while ceasing to justify a new action. Policies change, dependencies change, information becomes stale, and authority can be revoked. A

All rights reserved · Research exposition

continuing institution needs to preserve history while separately determining what governs current work.

These distinctions produce a more demanding notion of completion. A system must not treat an accepted request as an observed outcome or use a historical record as a shortcut around present conditions. At the same time, uncertainty is an input to continued recovery, not a reason to abandon otherwise achievable authorized work. Public review scope.

3. The consequential boundary includes observation

It is tempting to associate effects only with writes: changing a database, sending money, or publishing a message. Reading can also be consequential. Once protected information reaches a recipient, later deletion of the source or revocation of the original permission cannot establish that the recipient has forgotten it.

This makes information access part of the governance problem. A document supplied to a model, a search result, a memory entry, an error response, or a shared transcript can change what the recipient knows. The recipient may then reveal information through an apparently unrelated action. Reviewing only the final output text cannot establish the absence of such influence.

The relevant boundary also extends beyond the obvious application payload. Names, timing, ordering, presence, absence, errors, and resource availability can carry information. A system that protects a file while exposing a secret-dependent status may not satisfy its intended confidentiality claim. Whether a particular channel matters depends on the declared observer and threat model; the claim cannot silently exclude a channel available to that observer.

MIRAGE consequently makes completeness of the boundary an explicit assumption and qualification problem. Maintenance, recovery, diagnostics, administrative access, and alternate routes belong in that analysis when they can produce the protected consequence. Legal ownership of a machine does not establish technical control over every path through it.

The classical protection principle of complete mediation, discussed by Saltzer and Schroeder, provides a foundation for this question. MIRAGE examines it in a continuing autonomous institution, where information, authority, collaboration, and recovery interact over time. See the bibliography.

A useful consequence follows for assurance language. “Controlled by the system” must identify a resource and a boundary. A guarantee about a selected software interface is narrower than a guarantee about a hostile operating system, and both differ from a guarantee about every physical emission. Public review scope.

4. Collaboration does not create additional authority

An autonomous institution may distribute work across agents, replace a model, create a specialist, or organize a group. These operations can improve capability, but they do not justify increasing the institution's permissions or available resources.

All rights reserved · Research exposition

Delegation therefore raises a conservation question. Which rights did the delegator possess, which of those rights may be exercised by the recipient, and what outstanding obligations remain with each party? Creating another participant must not make a finite entitlement available twice. A group must not combine separate permissions into a broader power that none of its members was authorized to grant.

Identity requires similar care. A persistent institution, a particular running process, and a model used for reasoning are different subjects. Replacing the process may preserve the institution's liabilities. Replacing the model may change its competence while leaving its information obligations intact. Copying a document containing an identity does not transfer the identity's authority.

Information sharing is also a separate decision from group membership. Participants collaborating toward one objective need not receive every input or transcript available to one another. A common visible identity does not prove that all underlying contexts are interchangeable. Conversely, splitting work into several contexts does not establish that they are independent if they share hidden state or an uncontrolled communication path.

These observations constrain claims about scale. More processes are not automatically more independent reviewers. More keys are not automatically more independent operators. More agents are not automatically more usable authority. The relevant relationships concern actual control, access, and failure conditions.

The research problem is to preserve useful collaboration under these constraints. A design that can only succeed by giving every participant all information and all permissions has not established the intended separation. A design that prevents all useful collaboration has also not established practical autonomous value. Both security and task completion require evidence. Public review scope.

5. Information release and the limits of general claims

An institution may legitimately authorize the release of some information while protecting other information derived from the same source. This is more difficult than marking an output as approved. The release must have a meaning, a recipient, and a scope, and its consequences may depend on what has already been disclosed.

Several individually acceptable answers can reveal more when combined. An output can also become revealing when combined with outside information unknown to the system. A general claim that no secret can be inferred therefore needs a precise account of the possible underlying states, the observer's information, and the allowed release. Without those conditions, “safe summary” is an intention rather than a demonstrated property.

The MIRAGE sources discuss different forms of release reasoning. Exact results for a bounded formal domain, statistical privacy claims, and conservative information restrictions answer different questions. One cannot be promoted into another merely because each uses mathematical language. A capacity bound on a channel also does not establish that every permitted message is harmless or that a particular secret cannot be revealed through it.

The reference model explores selected finite cases. A release claim should therefore identify the domain over which it has been checked and the composition it permits. Extending those

All rights reserved · Research exposition

results to arbitrary natural language, unknown side information, or additional physical observations would require a new argument and appropriate evidence.

Transformation does not automatically end an information obligation. Compression, paraphrase, encryption, model replacement, and copying may change representation while preserving sensitive content or influence. Nor can a later revocation retract knowledge already received. These facts affect both policy design and recovery.

Practical usefulness remains an open empirical question. Strict restrictions can create pressure to authorize broader release simply to complete ordinary work. A useful evaluation must measure that pressure, including denied legitimate actions and requests to widen scope. Counting successful tasks after silently broadening the policy would not measure the original design. Public review scope.

6. A durable decision is different from a completed effect

A consequential decision can create an obligation that outlives the process that proposed it. The institution then needs to know whether the obligation was merely considered, became binding, was realized, or has adequate evidence of completion.

These differences matter when an operation involves several resources. Individual participants may complete at different times. A view that combines their incompatible intermediate states can misrepresent the intended result even when each local record is authentic. Coherent visibility is therefore a separate concern from having an authorized decision or storing all participants' reports.

External effects make the problem sharper. A destination may accept a request before the sender loses contact. The sender's uncertainty cannot establish that no effect occurred. Repeating the request may be harmless for one service and duplicative for another. A reliable recovery must use the destination's actual guarantees and available observations.

Revocation also has a temporal meaning. It can prevent a later decision without erasing an earlier irrevocable commitment. Where a prior decision created a binding obligation, recovery may still need to complete that exact obligation. Where no such commitment exists, changed authority may require a new evaluation. The system cannot choose whichever interpretation is more convenient after a failure.

Compensation belongs to this account as another consequential action. It requires authority, can fail, and may only partly remedy the original result. It does not rewrite history or make irreversible information disclosure disappear.

MIRAGE's conceptual requirement is continuity of responsibility across these conditions. The exact realization contract determines which outcomes can be made atomic, which retries are safe, and which uncertainties must remain explicit. A local record cannot supply a stronger contract than the resource actually supports. Public review scope.

All rights reserved · Research exposition

7. Observation, provenance, and settlement

The term settlement refers here to establishing the result required by a particular action or objective. It does not imply legal settlement, financial finality, or universal agreement about the world.

An observation must address the result in question. Evidence that a storage version became authoritative differs from evidence that a remote recipient accepted a message. Both differ from evidence that a human read the message or that a physical action occurred. A design should state which of these predicates it can establish and which remain outside its observation.

Independent checking also needs a precise meaning. Several verifiers can assess the same predicate, but their agreement remains dependent on the source they read. Repeating a false sensor value or a compromised store's response does not create independent knowledge of the underlying world. Likewise, several specialist checks that each cover a different fragment do not establish that a missing required predicate was checked.

Provenance can explain relationships among inputs, decisions, changes, and observations. It can distinguish a recorded dependency from an agent's asserted rationale or from information that may have influenced the work. It cannot recover private mental causes merely by recording a narrative.

This distinction connects to two established bodies of work. W3C PROV-DM provides a model of provenance involving entities, activities, agents, and their relationships. Lamport's treatment of distributed events explains why causal ordering cannot be reduced to an arbitrary wall-clock order. Neither result makes a provenance record a complete explanation of real-world causation. The bibliography identifies these sources and the limits of their use here.

Settlement can later be challenged by new evidence. Preserving the earlier record makes it possible to understand what was accepted and why. A present decision must separately account for the challenge. Deleting the earlier claim would obscure responsibility; continuing to rely on it without considering the new evidence would obscure current risk. Public review scope.

8. Recovery must preserve more than content

A backup can restore data while omitting later obligations, information exposure, consumed resources, or revoked authority. Its authenticity alone does not establish that it is an acceptable current state.

Recovery must therefore answer more than “Can the process run again?” It must determine what remains authorized, which effects may already have occurred, what work is still owed, and which information or resources cannot be treated as unused. A fresh process, new key, or replacement model does not resolve these questions by itself.

Historical content can remain useful without becoming current authority. Restoring an older document, for example, can be a new governed action in the present institution. It need not imply restoring the permissions or assumptions that accompanied the document in the past.

All rights reserved · Research exposition

There is also a limit to recoverability. If all sources capable of establishing current obligations are lost, an old authentic snapshot cannot prove that nothing happened afterward. A stronger recovery claim would require evidence the institution no longer possesses. The appropriate response depends on the governing authority and the external situation; it cannot be inferred from the mere availability of a backup.

Within an intact authority boundary, the intended behavior is continued work. The system should reconcile uncertain effects, obtain current information, use available valid routes, and complete the objective when its assumptions permit. Recording an interruption accurately is necessary for this process, but it is not the objective's completion.

Progress also depends on resources. Storage, communication, observation sources, and authorized participants can become unavailable. A safety claim cannot imply unlimited progress under every such failure. The research task is to characterize the conditions under which recovery succeeds and to show that expected interruptions do not create duplicate effects, renewed authority, or abandoned obligations. Public review scope.

9. Master, native design, and reference model

The MIRAGE family contains three principal levels of work. Their relationship is a proposed refinement, not evidence that one level already realizes all properties of the next.

Source level	Principal question	Evidence boundary
Master architecture	What must connect institutional authority, information, effects, and evidence?	Proposed architecture and stated proof obligations
Native computing design	What must the computing and resource boundary enforce for stronger claims?	Proposed native design, trust assumptions, and research obligations
Implementation closure reference	How can a selected bounded model make some semantics executable and reviewable?	Reference source and source-reported authoring fixtures; concrete qualification remains separate

The native design distinguishes functional behavior from what an observer can learn. An implementation can produce permitted business results while revealing additional information through another channel. Showing that it can reproduce selected abstract transitions is therefore different from showing that all relevant observations meet the intended constraint.

The reference model deliberately assumes a trusted host and supporting software for important protections. It does not establish resistance to a compromised host merely by describing stronger native goals.

All rights reserved · Research exposition

9.1 Bounded reference results

The following results are reported in *Implementation Closure Canonical v0.1*, dated 15 September 2026, §§17–18 and 24 and Appendix C. They concern its bounded P0 profile: one modeled world, two tenants, a finite operation vocabulary, and ideal authenticated roles. The authoring reference is a deterministic Python semantic model; its storage probe uses SQLite 3.46.1. The source reports execution and a second successful extraction-and-run of its embedded reference. This collection has inspected those reports and their stated predicates; it has not independently executed or reproduced the research reference. Abstract descriptions below identify the checked claims without publishing the construction.

Check	Inputs and case bound	Checked predicate	Source-reported outcome
Encoding and structural validity	5,000 random round trips; 89 structural schemas; 12 malformed examples; an extra-field case for each schema	Accepted values round-trip; the selected malformed and extra-field cases are rejected	Pass; 89 extra-field rejections. Structural validity supplies no authorization result.
Worked objective	One modeled reservation-and-disclosure objective with independent analysis and peer transfers; eight compartments and four local observations	The modeled state change, permitted output, and settlement are connected through their required observations	Pass; four decisions, six global orders, modeled remaining inventory 44, output 25. These are reference values, not production measurements.
Observation and interruption	Interruption of a protected reader followed by modeled recovery	Content is not delivered before its exposure is recorded; the old access ticket becomes unusable; recovery preserves remaining capacity	Pass for the reported fixture: no read before exposure, old ticket fenced, capacity preserved.
Repetition and coordinated visibility	A repeated application and a two-resource visibility fixture	A duplicate does not apply the effect twice; a current view cannot mix participants' old and new state	Pass; one application and rejection of the mixed current view. The report does not supply a general concurrency proof.
Revocation order	The three reported cases: relevant revocation before decision, after decision, and unrelated revocation	Current revocation blocks a new effect; it does not erase an already decided exact effect; unrelated revocation does not create a false conflict	No effect, exact decided effect retained, and no false conflict, respectively.
Finite ordering exploration	All 120 permutations of five selected events from one prepared transaction; interruption and recovery of either of two selected participants at each visited position	The reference preserves its checked invariants through those schedules	Python exploration passes: 600 attempted event steps, 300 rejected steps, 48 projected states, 77 edges. The generated TLA+ projection was not run in TLC.

All rights reserved · Research exposition

Check	Inputs and case bound	Checked predicate	Source-reported outcome
Structural pressure	The worked objective plus 64 protected reads over four existing recipient/resource envelopes	Added ordinary reads need not each cause a new global order in the fixed partitioned profile	68 total local observations, zero additional global orders, one outward application frame, and no blocked required effect. No latency, throughput, or operator behavior was measured.

9.2 Process-interruption storage result

The storage probe starts a fresh local database for each of nine interruption positions. One transaction changes a resource revision and four associated record sets: content, application evidence, business-operation identity, and history. The schedule covers entry, transaction start, each record insertion, the revision update, the point after commit, and normal completion. Eight cases terminate the writer process deliberately; the ninth completes normally.

The checked predicate is an all-or-nothing recovered tuple: either revision 0 with none of the four new records, or revision 1 with one record in every set. The recovered bundle must agree with its revision, and SQLite's integrity check must succeed. The source reports the old tuple at positions 0–6 and the complete new tuple at positions 7–8, with no mixed tuple. Position 7 is the deliberate exit after commit; position 8 is normal completion.

This result supports transaction recovery for those inputs, that SQLite runtime, and process interruption. It does not establish arbitrary power-loss durability, honest storage hardware, protection from a database administrator, anti-rollback, or conformance of the proposed production storage build. Those are different tests and assumptions.

The sources also distinguish a development decision engine from a replicated implementation. Fault-tolerant ordering requires a real protocol and a correct integration. A collection of signatures does not establish those conditions. Protection against incorrect decisions is separately scoped from protection of private information supplied to the decision participants.

Several concrete qualifications remain explicitly unperformed in the reviewed reference source: the independent production implementation, complete cryptographic and transport binding, deployed isolation and replicated fault campaigns, physical timing evaluation, real Cuttlefish integration, and an operator study. The finite ordering exploration above therefore supports the reference's selected schedules, while the generated model file supplies neither an executed TLC result nor an unbounded proof.

The reference's contribution is executable semantics within a declared model. Claims about the operating environment, physical emissions, or complete institution require the additional qualifications described above. Public review scope.

All rights reserved · Research exposition

10. Four illustrative cases

Four hypothetical cases show how authority, information, and continuing responsibility interact. Their outcomes illustrate requirements rather than report deployment observations.

Authorized work completes. A research institution permits an agent to prepare a report from internal material and distribute an approved public result. Reasoning about the material, deciding what can be disclosed, publishing the approved artifact, and observing publication are separate parts of the task. Completion requires the authorized result at the intended destination, supported by adequate observation. The agent's declaration that it has finished is useful communication, but it cannot establish the result on its own.

A proposed expansion is denied while the objective continues. During that work, an agent suggests sending the internal material to an additional service. The original objective does not settle whether that recipient and use are permitted. If the new route is outside authority, the system can continue through another authorized route or seek the genuinely required decision. The denial does not make the objective impossible, and the importance of the objective does not create the missing permission.

An interruption leaves an external consequence uncertain. The publication destination accepts the report, but the response does not reach the institution. Recovery examines the destination's state and the conditions of the original action. If observation establishes acceptance, the institution can complete its record without publishing a second copy. If observation cannot establish the result, a retry depends on the destination's guarantees and current authority. An empty local response is not evidence that the report stayed private.

A later model change preserves responsibility. The institution replaces the model that prepared the report. The new model can continue the work using authorized information, but the change does not erase earlier disclosure or renew consumed entitlements. If the earlier report is challenged, responsibility remains attached to the continuing institution and its recorded decisions. Competence under the new model must be evaluated separately; continuity of identity is not proof of equivalent judgment.

Together the cases show why the proposed boundary must connect ordinary work with information governance, observation, and recovery. Isolating one stage and calling it successful leaves the institutional question unanswered.

11. Relationships and unresolved adoption

MIRAGE is explicitly connected by its source contract to the doctrine, native design, and reference model. Those connections assign obligations and intended refinements. They do not constitute completed proofs of equivalence or integration.

CSSL studies semantic change and the continuity of evidence across software work. Its conceptual correspondence with MIRAGE concerns current information, accountable authority, and observed consequences. Shared themes do not make the two specifications interchangeable or establish a common implementation.

All rights reserved · Research exposition

Cuttlebone addresses a boundary associated with the proposed Freehold institution. The institutional question is compatible with MIRAGE's concern for controlled consequences, but a local contract must establish any technical substitution or composition. The existence of both documents does not do so.

Semantic OS and the Cuttlefish environment describe related operating concepts and proposed product relationships. Demonstrating their adoption requires a version-specific account of the implementation and evidence from the resulting operation.

The architecture study classifies these relationships as explicit source contracts, conceptual correspondences, proposed integrations, or unresolved adoption. Evidence grade is a different question. An explicit source contract can still describe an unimplemented design, and an experiment can exist without establishing a broader integration.

12. What research would resolve

The most consequential open question is whether complete control over the declared effect and observation boundary can be established at an acceptable cost. The answer must include ordinary work, secondary channels, administration, interruption, and recovery. A successful narrow example cannot answer it for every environment.

A second question concerns information release. Useful autonomous work needs access to information and the ability to communicate results. Exact theories can provide strong claims for limited domains; broader semantic judgments are harder to establish. Evaluation must measure both what remains protected and which legitimate work becomes difficult.

A third concerns evidence. Institutions need observations that support their actual decisions, with clearly understood common causes and failure conditions. More retained records can improve accountability while increasing privacy, storage, and interpretation obligations. It remains necessary to determine which records and observations are sufficient for particular claims.

A fourth concerns continuity under change. Models, policies, participants, and hardware will evolve. The institution must preserve earlier obligations without freezing every implementation choice. Refinement, migration, and current authority therefore need to be evaluated together rather than treated as administrative afterthoughts.

Finally, a convincing evaluation must demonstrate useful completion under the stated constraints. It should distinguish safety from availability, targeted fault tests from general proofs, and reference behavior from actual deployment. MIRAGE's research proposition is that these questions belong to one account of autonomous consequence. Establishing a realization of that account requires a declared operating boundary, a qualified implementation, and evidence of both protection and useful work.

All rights reserved · Research exposition

Source guide

This successor guide points to public arguments and their limits. The new private evidence linkage remains excluded. Previously public conceptual source guides, including any retained titles, document ordinals, sections and chapter maps, are not claimed to have been removed; see the correction scope. The historical edition remains available unchanged.

Problem and scope. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Authority and consequences. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Information and boundaries. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Collaboration and continuity. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Release and composition. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Realization, recovery, and availability. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Evidence and provenance. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

Levels and qualification. Review the relevant argument and assumptions in this paper, the evidence guide, and the research review route. Specification, reported results and production observations have distinct scope.

No source-guide entry establishes a full implementation, a complete historical migration proof, or an independent audit.

All rights reserved · Research exposition