GOVERNED AUTONOMY

SEPION: What agreement among generated artifacts can establish

A paper in the Governed Autonomy collection

Andrew Durden
Independent research
Publication edition 2026.10.6

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

Content identifier: 058a680e5b3eee8e57a28486

All rights reserved · Research exposition

Contents

SEPION: What agreement among generated artifacts can establish **3**

Abstract . . . 3

1. The problem with a convincing consensus . . . 3

2. Defining the object of agreement . . . 4

3. Requirements, stability, and the limits of measurement . . . 6

4. Effects are part of behavior . . . 6

5. Evidence has conditions and a lifetime . . . 7

6. Worked conceptual cases . . . 8

7. Threat model and trust assumptions . . . 9

8. Evidence boundary and research questions . . . 10

Source guide . . . 10

All rights reserved · Research exposition

SEPION: What agreement among generated artifacts can establish

Governed Autonomy · Publication 2026.10.6
Research study: conceptual exposition of the SEPION specification
Source: SEPION ACP/1.1 Canonical Package, revision v2.1, with the 2026-09-22 evidence boundary

Abstract

Repeated generation can produce reassuring agreement without establishing that a result is correct, authorized, or robust to a change in circumstances. SEPION studies this gap for generated software artifacts. It proposes evaluating agreement in terms of declared behavior, separating that evaluation from compliance with requirements, and keeping both conclusions relative to the instruments and conditions that produced the evidence. A decision concerns a supported behavioral class; the particular artifact delivered to a consumer remains accountable to that decision.

This paper develops three implications of that approach: agreement needs a defensible behavioral subject, effects and their origins belong in the comparison, and changed evaluation conditions can invalidate an apparently stable result. Four hypothetical cases show how those distinctions affect use and recovery. SEPION remains a proposed protocol with a conformance program; implemented evaluation and empirical validation are needed to establish its practical value.

1. The problem with a convincing consensus

Suppose several generators produce a transformation for customer records. Their programs look different, yet they produce the same output on a familiar sample. A reviewer may reasonably see this as useful evidence. The difficulty begins when agreement becomes a substitute for answering what was compared, which inputs mattered, and whether the shared behavior was acceptable.

The generators could share a mistaken interpretation. They could all omit the same rare case. Their surrounding tools could supply the same hidden assumptions. Conversely, two programs that look very different could satisfy exactly the same stated behavior. Agreement in wording, structure, or visible output therefore has an uncertain relationship to the property a consumer actually needs.

SEPION places that relationship at the center of evaluation. Its subject is consistency among generated artifacts under an explicit behavioral contract. A contract describes the required meaning and the boundaries of the evaluation. A framing is the surrounding presentation through which the task reaches a generator. Changing the framing can test whether the generator is sensitive to irrelevant presentation, but a change in the requirements would be a different task. Keeping that distinction clear is necessary before interpreting stability. Source: scope and design laws.

All rights reserved · Research exposition

The research contribution is a disciplined account of what agreement can support. Behavioral similarity, compliance, confidence, permission to use a result, and continued validity are separate claims. Each needs its own basis. A large volume of agreement cannot compensate for a forbidden effect or an unmet requirement. Equally, an artifact that satisfies the requirements once does not establish reliable agreement across the variations relevant to a broader claim.

This connects SEPION to the collection's thesis: a capable generator supplies proposals, while consequential use depends on evidence and authority appropriate to the decision. The separation is especially valuable when the generator is persuasive, because a persuasive explanation does not reveal whether the underlying comparison was sound.

2. Defining the object of agreement

A behavioral class groups artifacts whose declared relationships support treating them alike under one common behavioral contract and evaluation scope. That common scope includes the evaluator and environment, relevant inputs and states, effects, attribution conditions, and any expressly excluded operational differences. Members are considered within the same fixed evaluation population. Comparing one pair under one contract and another under a different contract does not establish membership in a common class.

The positive criterion has two evidentiary forms. In a proof-supported class, checked arguments establish the required equivalence relation under the common scope. Valid transitivity of that relation can justify a relationship that was not checked as a separate direct pair. In a class supported partly by unsuccessful attempts to distinguish artifacts, all required direct-pair distinguishing obligations must be completed before grouping is considered. The complete group relationship must then be justified by the required pair evidence or an independently supported group bound under the declared method. A group argument does not excuse unfinished required obligations. The group claim remains attached to that evidence and its coverage. An observed distinguishing case within the declared scope excludes the affected grouping. Missing or inconclusive evidence leaves artifacts separate for that claim; it does not count as agreement.

An equivalence claim backed by a checked proof has a different status from one for which a bounded search found no distinguishing case. The latter can be useful, but the absence of a discovered difference is not proof that a difference is impossible. SEPION retains that distinction in its account of evidence. The strength of a group cannot exceed what is supported for the relationships on which it depends. Source: formal model and primitive definitions.

There is a second difficulty: local agreements need not justify a larger grouping. If two artifacts agree in one comparison, and one of them agrees with a third in another comparison, the evidence may still fail to establish that the whole group can be treated alike. The source requires support for the complete group relationship before it can govern a decision. Evidence from bounded distinguishing work retains its weaker status even when all required relationships are covered. Combining it with proof evidence does not turn the resulting group into a wholly proved class.

All rights reserved · Research exposition

The following hypothetical matrix considers three artifacts, A, B, and C, under the same contract and instruments. “Bounded agreement” means that the declared distinguishing evaluation found no counterexample with adequate stated coverage; “unknown” means that the required relationship has not been established. This is an example of evidence interpretation, not a new grouping algorithm.

Comparison	A	B	C
A	Same artifact	Bounded agreement	Unknown
B	Bounded agreement	Same artifact	Bounded agreement
C	Unknown	Bounded agreement	Same artifact

This matrix supports the recorded A–B and B–C comparisons at their bounded scope. It does not support treating A, B, and C as one class merely because B links the other two. Any required direct A–C distinguishing work must be completed, and the whole-group claim must have adequate evidence under the declared method. A separately justified group bound remains subject to those obligations. If A–C yields a distinguishing case, the three-member grouping is excluded. If the required work supports bounded agreement and the group justification is adequate, a group can be considered at that lower evidence tier, with its limits retained. If instead A–B and B–C are sound checked proofs of the same equivalence relation, the A–C relationship follows from that relation's transitivity. The word “unknown” must therefore be interpreted in relation to the proof obligations: absence of a direct comparison differs from absence of a justified relationship.

Agreement on a common finite suite is a third, narrower statement. If A, B, and C have equal observed outputs on every shared case, they agree on that suite. Such equality can be transitive for those recorded cases without establishing equivalence over the full input, state, or effect scope. A finite-suite cluster is eligible to support only the claim its observations warrant. The source does not permit ordinary pass/fail tests, majority agreement, or surface similarity to stand in for its behavioral class criterion.

This point matters even when no adversary is present. Different tests reveal different differences. A parser might agree with another parser on ordinary numbers and with a third on ordinary dates while differing on ambiguous values. It would be misleading to erase those distinctions through a label such as “same result.” The evaluation needs a defensible account of the relationship that label denotes.

SEPION also distinguishes the behavioral decision from the artifact eventually delivered. The consumer receives a concrete artifact, but the reason for its acceptance comes from an evaluated relationship to requirements and evidence. Substitution requires the replacement to belong to the supported class at the applicable tier and to retain the separately required compliance and current-use conditions. A new artifact, changed dependency, or expanded consumer cannot inherit that basis from resemblance alone. Otherwise a well-supported evaluation could become an empty wrapper around a different result.

The term *class* is thus an accountability device. It asks a reviewer to identify the unit about which the evidence speaks and to keep later consumption tied to that unit.

All rights reserved · Research exposition

3. Requirements, stability, and the limits of measurement

Compliance means satisfying the declared requirements under the relevant evaluation. Stability concerns the persistence of a supported behavioral result across the tested task presentations and conditions. These questions are related, but their answers can differ.

A population of generators could reliably choose the wrong unit of measurement. Its behavior would be stable while violating the task. A correct artifact might appear rarely, making its compliance evidence useful without establishing that the generation process reliably reaches it. Treating either observation as a complete verdict hides information that a decision maker needs.

The quality of the contract is another dependency. Mechanical evaluation can check a stated property while leaving the human objective incompletely represented. If a contract asks only for syntactically valid records, satisfying it says little about whether the records preserve the required business meaning. SEPION treats adequacy of the declared requirements as an explicit concern, while acknowledging that the original intent gap remains. A formal statement cannot prove that its author wanted the right thing. Source: compliance, convergence, and known limits.

Statistical confidence adds a further layer of conditions. Looking repeatedly at accumulating samples and stopping when a desirable result appears can change the meaning of a confidence statement. Discovering additional behavioral alternatives can also change what must be ruled out. A published confidence claim needs to remain valid under the way evidence was collected and the way the decision was made.

Similarly, several model names do not establish independent evidence. Shared training, prompts, scaffolding, or tools can produce correlated mistakes. SEPION treats independence claims as matters for measurement, not as consequences of branding. Agreement across a fleet has value only to the extent that the fleet and the claim have been characterized.

The direction of variation also matters. A result may be stable under harmless rewording and fragile under a different presentation channel. An average can conceal that concentration of weakness. Refusals, failed generations, and unresolved evaluations belong in the observed population; removing them can make performance appear more dependable than the actual experience.

These considerations lead to a restrained interpretation of a result: the evaluated artifacts behaved with a stated degree of consistency under specified conditions, and the supported behavioral outcome separately met the stated requirements. Extending that claim to new conditions needs further evidence.

4. Effects are part of behavior

Output comparison becomes especially misleading when artifacts act on their surroundings. Two programs can return identical text while making different persistent changes, contacting different destinations, or reading information they were never authorized to use. If these differences matter to the contract, they belong in the semantic comparison.

All rights reserved · Research exposition

The source distinguishes meaningful artifact effects from certain declared operational differences and from the behavior of the evaluation environment itself. This is a scope distinction, not a general permission to ignore inconvenient observations. An effect cannot be treated as harmless simply because it resembles a temporary file, a log, or a compiler activity. Its role and origin matter. Source: observation and effect boundaries.

Consider a diagnostic log. If nothing in the declared task consumes its contents and the relevant operational constraints are satisfied, a difference in that log might fall outside the behavioral equality being claimed. If a downstream process reads the log to decide whether to charge a customer, the same information has become consequential. Calling it “diagnostic” cannot remove it from the task’s meaning.

Origin matters for a related reason. A familiar tool can execute instructions supplied by a generated artifact. Recognizing the tool’s name does not establish that every resulting effect belongs to the trusted environment. A supported attribution needs to distinguish the environment’s own work from the artifact’s influence through it. Unresolved attribution limits the conclusion that can be drawn.

The resulting claim is conditional: some operational differences may fall outside a declared equivalence scope while remaining subject to observation, control, and appropriate labeling. The comparison establishes equality only over the properties it actually covers.

The boundary has residual limits. An unmodeled outside channel can escape the comparison. A downstream consumer absent from the contract may turn an apparently irrelevant difference into a meaningful one. Complex execution can make origins difficult to distinguish. SEPION records these as assumptions and research obligations, rather than offering an unrestricted guarantee about arbitrary software.

5. Evidence has conditions and a lifetime

An evaluation result depends on the instruments that produced it. A changed evaluator, environment, set of requirements, or policy can change the meaning of the result. Reusing an old decision is justified only when the necessary relationship to the new conditions remains supported.

More demanding evaluation does not always make an old score easier to interpret. A stronger test can reveal distinctions that a weaker test missed. An apparent increase in consistency might instead result from a changed definition of what counts as the same behavior. Comparing the two scores without addressing that change would mix different measurements. Source: instrument scope and drift.

SEPION studies drift as movement in supported behavioral relationships. A superficial source difference might leave the relevant behavior unchanged; a small change could alter an important boundary. The question is what changed in the relationship among behavior, requirements, and evidence.

A new evaluation setting may have no comparable baseline. That absence is neither evidence of regression nor evidence that everything is satisfactory. It is a limit on what comparison is available. New validation can establish a baseline, but the absence of earlier comparable results should remain visible in the claim.

All rights reserved · Research exposition

Reproduction and authority also have different roles. Reproducing an old evaluation can help confirm what it established. It does not, by itself, renew permission to use the artifact. Permission can be withdrawn for reasons independent of the artifact's past behavior. Conversely, a newly granted permission does not repair missing evidence. Continued use needs both questions answered under the applicable conditions.

This connects the paper to the collection's architecture and evidence framework. Records support accountability when their subject, scope, and current relevance are clear. The existence of a record alone does not establish that its conclusion governs the next action.

6. Worked conceptual cases

Four hypothetical cases illustrate the conditions under which an evaluation could support continued use. Their outcomes are not reported measurements.

6.1 A supported transformation

A records team needs a transformation that preserves customer identifiers, represents missing values consistently, and rejects malformed dates. Several generated artifacts differ in organization and naming. The evaluation is explicitly limited to a declared input language and the states relevant to this transformation.

The team finds supported behavioral agreement and separate evidence that the selected behavior satisfies every declared requirement. The artifact delivered for use remains tied to that evaluated behavior and to the conditions of the decision. The result is useful: the team has a reasoned basis for this transformation within the stated boundary.

The conclusion does not expand to every possible date representation or every future data source. A later source with a different date convention requires attention to the changed requirements. The successful case demonstrates scoped usefulness, rather than a universal claim about correctness.

6.2 Agreement that must not authorize use

A generated reporting routine is intended to summarize authorized records. Several artifacts produce the same totals, but one obtains missing information by reading another organization's records. Its final output looks plausible and may match the others.

The unauthorized read defeats a requirement that output agreement does not address. Repetition of the same result cannot compensate for crossing the information boundary. Use of the artifact is denied on that basis. An acceptable continuation would require an artifact whose behavior satisfies both the reporting requirement and the applicable information restrictions.

The case illustrates why denied use is an evaluation result rather than the completion of the underlying business objective. The report may still be achievable through a permitted artifact. The evidence should support that continuation without reclassifying the prohibited behavior as acceptable.

All rights reserved · Research exposition

6.3 Recovery after a changed evaluator

A previously evaluated parser is considered for reuse after the evaluation has been strengthened. The new conditions expose an ambiguity that the earlier comparison did not resolve. The old decision is no longer sufficient for the new claim.

The objective remains to obtain a parser suitable for the task. Re-evaluation can separate the newly distinguishable behaviors and establish which one meets the requirements. If that evaluation succeeds, a new scoped decision can support continued work. If evidence remains insufficient, the unresolved obligation stays visible.

The recovery is successful when the task regains an appropriate basis for use. Merely retaining the earlier certificate or reporting that it became unsuitable would not establish the required result. This example does not assume that every environment or historical dependency will always be available for reproduction.

6.4 A log that changes significance

Two artifacts produce the same requested report but write different diagnostic messages. The original evaluation excludes specified operational log differences from its behavioral equality claim. That restriction is made clear to the consumer.

A later workflow begins reading the messages to decide whether the report should be distributed. The log now affects a consequential outcome. Its earlier exclusion cannot justify treating the expanded workflows as equivalent. The requirements and evidence must address the new consumer.

The case shows why a narrow exception must remain attached to the claim. A distinction can be irrelevant for one purpose and decisive for another. The evaluation does not give the difference a permanent status independent of use.

7. Threat model and trust assumptions

The source’s threat model covers misleading agreement, hidden behavioral differences, contaminated evaluation, inappropriate reuse of evidence, and overstatement of guarantees. Generators can produce unsuitable artifacts. Evaluation inputs can be chosen to flatter a result. An instrument can be weakened. A consumer can receive something different from the evaluated subject. An apparently independent group can share the same failure. Source: threat and conformance catalogs.

The proposed response relies on trusted evaluators, soundly governed evaluation instruments, controlled execution, appropriate isolation, and retained evidence. These dependencies are substantial. A corrupted evaluation authority or an unsound checker can undermine the argument even if the surrounding records are intact. A record of a wrong conclusion remains a wrong conclusion.

The source also treats the language of the claim as an assurance boundary. Calling bounded unsuccessful falsification a proof, implying independence without measurement, or claiming complete equality while excluding operational differences can cause downstream misuse without changing the underlying artifact at all. Evaluation and publication therefore share a responsibility to state what was actually established.

All rights reserved · Research exposition

The conformance program defines obligations for an implementation. Demonstrating conformance requires executed evaluations and their results. Historical comparisons with other governed systems provide context but cannot supply that evidence for SEPION.

8. Evidence boundary and research questions

The package specifies a proposed protocol and records corrections to several evidence definitions. Operational conformance, interoperability, and practical performance remain to be demonstrated in an identified implementation. The conformance and negative-test catalogues define what such an evaluation would need to cover; no completed protocol evaluation or deployment qualification is established here. Source: edition and evidence scope.

Several questions determine the practical value of the approach. How much of an ordinary task can be specified precisely enough for useful comparison? How can the adequacy of those requirements be evaluated without merely reproducing the same omissions? What costs arise when distinguishing difficult artifacts requires extensive search or proof? How should a consumer reason about evidence that is useful but remains short of proof?

Other questions concern change. When do stronger instruments permit comparison with old results? How much evidence is needed to characterize correlation among generators? How can an evaluation maintain useful coverage as the space of inputs, states, and relevant consumers grows? How should historical reproducibility be handled when a required environment is no longer available?

The practical hypothesis is that these distinctions improve decisions about generated artifacts: a smaller, carefully scoped claim can be more useful than an expansive but ambiguous assurance. Establishing that value requires implemented evaluation, representative workloads, measured costs, and evidence that the resulting decisions improve outcomes. The source provides a research program for those questions; it does not establish their empirical answers.

Source guide

The following references identify sections of *SEPION ACP/1.1 Canonical Package* underlying the conceptual argument. Proprietary construction and selection methods are excluded. Terms used across the collection are defined in the glossary.

Scope and foundations. Artifact 1, “Normative Protocol Specification,” §§1–4, “Scope,” “Normative Language,” “Guarantee Canon,” and “Design Laws.”

Behavioral comparison. Artifact 1, §§5.1–5.7 and §6, “Formal Model” and “Primitive Definitions”; §9, “Protocol Non-Guarantees.”

Requirements and measurement. Artifact 1, §§5.11–5.13, “Compliance,” “Convergence,” and “Anisotropy”; §12, “Known Limits.”

Effects and scope. Artifact 1, §§5.8–5.10.1, “Observation Partitioning,” “Operational Effect Quarantine,” “Effect Origin Attribution,” and its verdict table; §§9 and 12.

All rights reserved · Research exposition

Change and continued use. Artifact 1, design laws 9–12 and 15, §5.14, “Drift”; Artifact 5, §§5 and 8, “Deterministic Replay Tests” and “WORM and Replay Integrity Validation.”

Assurance obligations. Artifact 5, “Conformance and Validation Suite,” §§1–11; Artifact 6, “Threat Catalog,” “Threat Table” and “Residual-Risk Register.”

Edition and evidence scope. Package opening evidence boundary, “Package Manifest,” and Artifact 1, §§8–12. The eight-artifact package includes implementation and runtime contracts.

All rights reserved · Research exposition