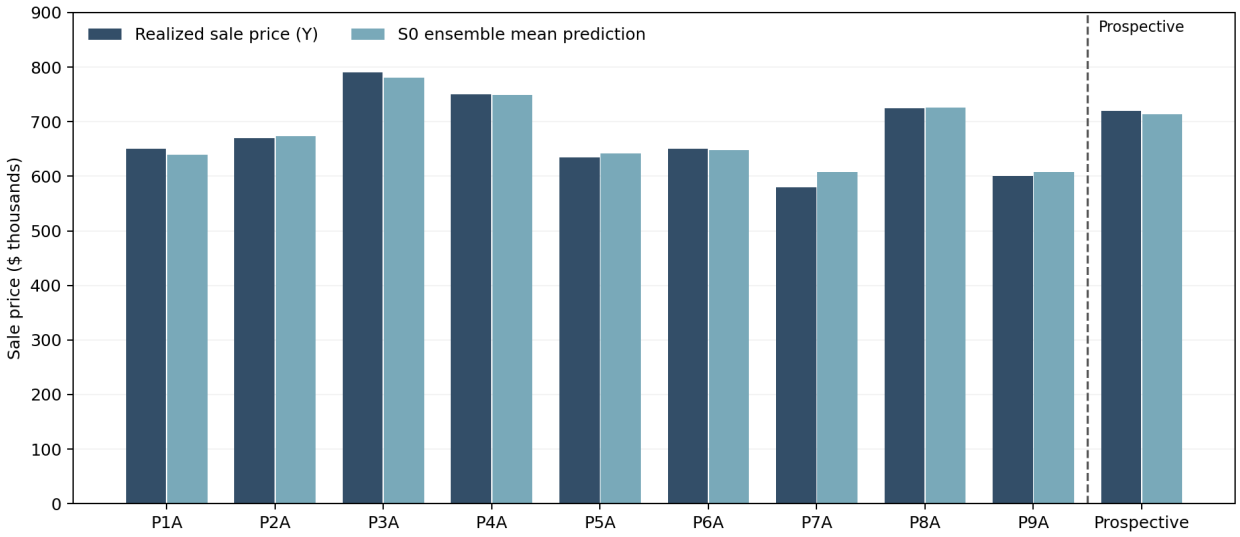
# The Decision Function Appears to Work

## Why Not Trust It?

### Nine historical operational valuations.
### One genuinely prospective valuation.




An autonomous AI Decision Function was tasked with predicting the eventual sale prices of listed homes in Seaside at Ponte Vedra Beach, Florida. It operated within defined experimental boundaries, rather than as a human appraiser or an already-authorized real-world valuation system. The chart shows how its predictions compared with the prices at which those homes ultimately sold. The first nine are retrospective predictions made with the sale outcomes already known to the investigator but withheld from the Decision Function; the tenth was made before its sale outcome was known.

The historical predictions track realized prices closely. All nine are within 5% of their respective outcomes, with mean APE(F) of 1.22% and mean CV(F) of 0.78%. The prospective prediction was $714,250; the property subsequently sold for $720,000—an error of approximately 0.80%.

The dashed line marks a meaningful evidentiary boundary. The prospective observation is not included in historical aggregate statistics. These are ten property outcomes, not a large independent validation population.

**Explore the supporting work:** Page-specific EXPLORE links open material related to that page. Evidence & Methods, at the bottom right of every page, opens the DJE overview, with access to companion documents, workbooks, and the worked case.

EXPLORE / Numerical Verification · Predictions

# Historical Performance Alone Does Not Warrant Deployment

### An encouraging result warrants further scrutiny.

The correspondence between predictions and realized sales is striking. It gives us a reason to investigate this Decision Function seriously. It does not, by itself, tell us how much of that performance is repeatable, which conditions made it possible, or whether those conditions will remain in place.

A reliable-looking result can arise from a sound decision process, a favorable evidence environment, chance, or some combination of these. Our task is not to assume the first explanation—or to discount it reflexively in favor of the others.

The more useful question is: what about the engineered system and its operating environment gives us reason to expect this performance to continue?

If conditions change, how would we recognize the change and determine whether the next prediction remains justified?

# There Are Engineering Reasons for Confidence

### The laboratory did more than ask an AI for a price.

- **A bounded model of the valuation setting.** Defined property information, comparable-sale eligibility, and evidence-window rules determined what the Decision Function received.
- **A well-defined Decision Function.** Clear instructions specified the task, restricted evidence use to the supplied inputs, and required an inspectable Experimental Record (ER) of outputs and reported reasoning.
- **Checks on evidence and execution.** Source comparisons and ER review addressed preparation errors, misreadings, and departures from the intended procedure.
- **Repeated executions and outcome comparisons.** Qualifying ensembles made variation observable; comparison with realized sale prices measured accuracy across the reported cases.
- **Outcome concealment and execution isolation.** Subject sale prices were withheld, and temporary chats separated experimental executions to reduce opportunities for leakage and carryover.

Their effects and limits will be examined in the engineering sections that follow.

# Protecting Experimental Isolation

## Outcome isolation required more than withholding a number.

- **Conceal the outcome.** The subject's realized sale price was redacted from the Decision Surface Snapshot (DSS).
- **Bound the evidence.** The DF was restricted to the supplied DSS and authorized table. The table restated selected DSS evidence; internet searches for additional information were prohibited.
- **Separate executions.** A subject's otherwise withheld sale price could appear in another experiment's evidence. Each experiment therefore used a separate Temporary Chat to reduce opportunities for cross-experiment carryover.
- **Check evidence before execution.** Source comparisons checked the prepared evidence.
- **Review execution afterward.** Each execution produced a frozen Experimental Record. QC assessed adherence to the authorized procedure before admitting its output to the qualifying population.

Realized sale prices were used afterward to evaluate qualifying results.

EXPLORE / Worked Case · P4A

# The Operating Environment Matters Too

*Reliability belongs partly to the conditions in which a decision is made.*

The laboratory was conducted within a relatively coherent residential community. Its properties offered useful comparative evidence, and the relevant market period was comparatively stable. Those circumstances likely made the valuation problem more tractable.

This is not a weakness to be explained away. Defining a bounded operating field is itself an important engineering choice. A system may be reliable within that field without being reliable everywhere.

The observed results do not establish that the same Decision Function would perform comparably for unusual properties, thinly supported markets, or rapidly changing conditions.

How much of the observed reliability belongs to the engineered Decision Function, and how much depends on the evidence environment in which it operates?

We need not fully separate those contributions before proceeding. But a system authorized to operate autonomously must take that dependency seriously.

# The Decision Justification Problem

**Uncertainty must be characterized. Sufficiency must be assessed. Authority must be governed.**

A decision can be accurate without being sufficiently justified. A sufficiently justified decision does not, by itself, authorize autonomous execution.

- **Uncertainty** concerns what remains unknown about the decision, its outcome, and the consequences of failure.
- **Sufficiency** concerns whether the available evidence and engineering controls adequately support the contemplated use.
- **Authority** concerns who may permit autonomous execution, under what conditions, and what happens when those conditions cease to hold.

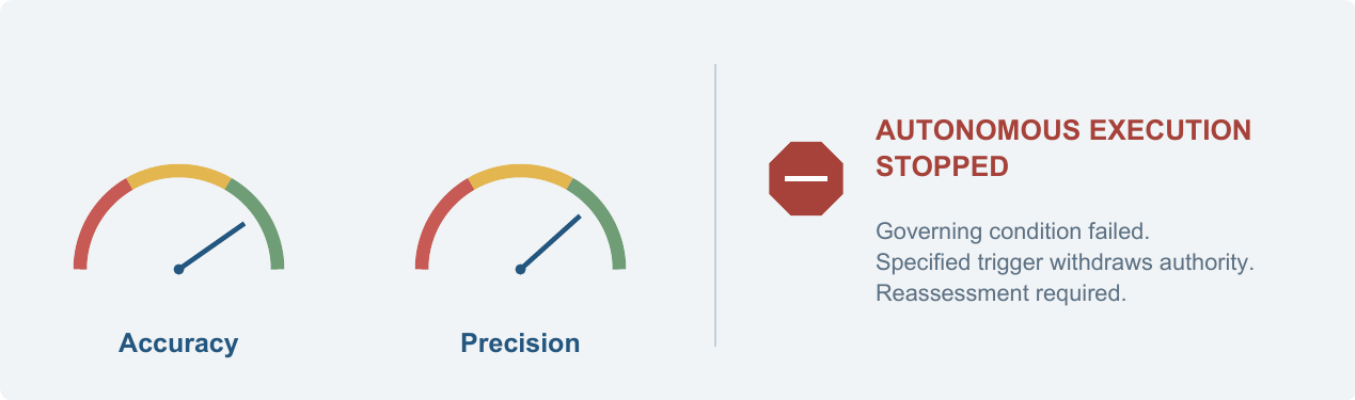
Decision Justification Engineering (DJE) treats these as related but distinct engineering responsibilities. Its concern is not merely whether a Decision Function can produce an answer, but whether autonomous execution can be justified, authorized, and kept conditional as circumstances change.

# Reliable Predictions Are Not Reliable Governance

## Two distinct kinds of robustness matter.

The first concerns **performance**: can the system produce estimates and predictions that are sufficiently reliable within a defined operating field?

The second concerns **governance**: once responsible decision-makers have authorized autonomous execution, can the surrounding system determine whether the conditions supporting that authority still hold—and automatically withdraw autonomous Execution Authority when they do not?




## Performance can remain reassuring while its justification changes.

A Decision Function may remain unchanged and continue to produce plausible, tightly grouped estimates and predictions while the evidence supporting them becomes less representative or less informative. Apparent confidence in the output cannot substitute for renewed justification.

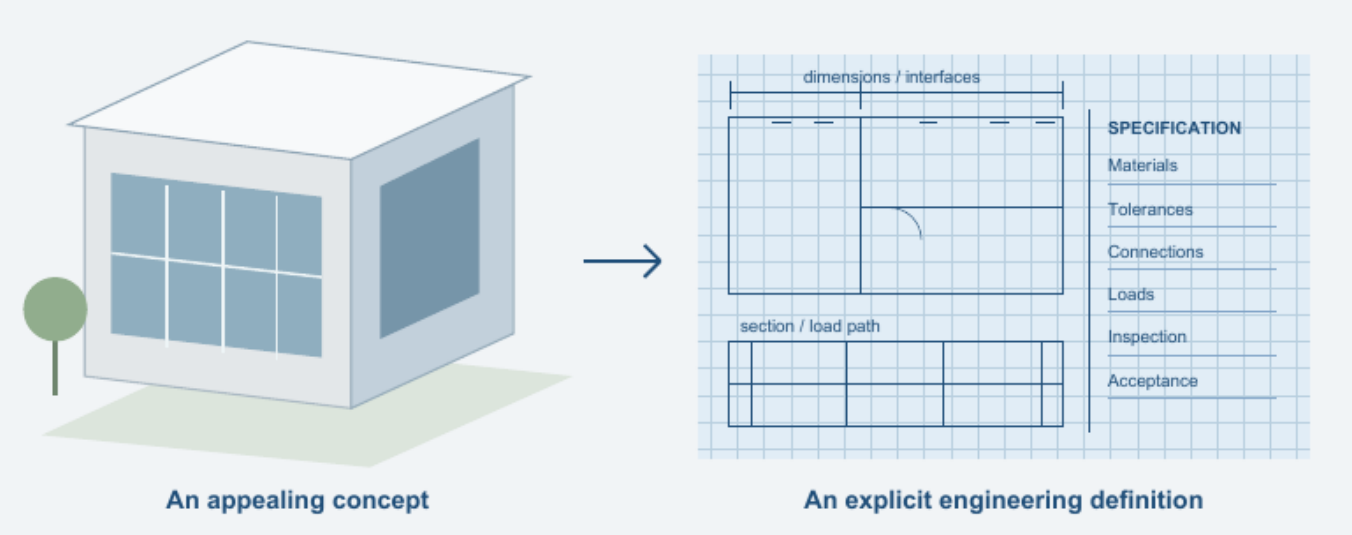
The governance objective requires specified triggers that automatically disengage autonomous execution when governing conditions fail. A dashboard warning alone, relying solely on the steward to act, is insufficient for that objective. These deployment controls remain engineering proposals rather than demonstrated operational mechanisms.

This is the transition from predictive reliability to governance reliability.

# From Performance to Justified Authority

**Governance objective: Establish when autonomous execution is justified, determine whether that justification continues to hold as conditions change, and withdraw authority when it no longer does.**

**The governing idea is straightforward.**
**Implementation is demanding.**
**Engineering is the discipline that connects the two.**




The objective becomes operational only when the conditions supporting authority, the means of checking them, and the required responses to change are specified.

The engineering task is to connect those requirements to evidence, execution controls, empirical assessment, and continuing permission. The next questions make that task concrete.

# Making Governance Operational

## From abstract principles to engineering questions.

The purpose of this instantiation is not simply to illustrate individual governance principles. It is to make those principles operationally concrete within a single, coherent decision process.

Questions about evidence, changing conditions, qualification, authority, and revocation cannot be resolved independently when they must coexist in an operating system. Their relationships become visible, their assumptions become contestable, and their consequences become observable.

**Consider a seemingly straightforward principle: material changes in operating conditions should trigger reassessment.**

**What does that require in practice?**

- What constitutes a material change in the evidence state?
- Does a change require reassessment of the candidate, the evidence state, the configured DF, or some combination?
- Which changes can be handled within approved procedures, and which require further engineering or renewed authorization?
- When does a finding call for a local correction, and when does it require broader review of prior results, downstream decisions, or continuing authority?

The instantiation does not establish that every proposed mechanism is sufficient for deployment. It does, however, move the governance discussion from abstract principles toward engineering propositions that can be specified, tested, challenged, and refined.

# From Governing Principle to Engineering Discipline

## The laboratory makes these responsibilities inspectable.

Making the governing objective operational requires evidence boundaries, execution controls, empirical characterization, sufficiency standards, and continuing authorization.

The Residential Property Valuation Decision Node (RPVDN) laboratory makes these responsibilities inspectable in one concrete application: predicting residential sale prices. The following pages introduce the properties, evidence, configured Decision Function, and experimental procedures, then examine what the results support—and what they do not.

# The Laboratory: Seaside at Ponte Vedra Beach

## Before the framework can be understood, the laboratory has to be visible.

The empirical work was built around one deliberately bounded **Residential Property Valuation Decision Node (RPVDN)** in one residential community: Seaside at Ponte Vedra Beach.

A Decision Node is a bounded decision task with a defined output, an identifiable evidence basis, and a specified decision procedure. Here, the task is estimating the eventual sale price of an actively listed residential property.

Nine properties sold in Seaside during 2025. All nine completed sales were used to construct the original comparable-sale evidence population. Together, these nine completed sales formed the initial empirical evidence state, State 0.

The laboratory then followed the community into 2026. Each retrospective operational prediction was produced by the Decision Function with the subject's actual sale price withheld. The resulting predictions were compared with the realized sale prices.

# The Laboratory: How the Engineering Case Unfolds

## Core components of the Decision Node—and a guide to the investigation ahead.

**THE MODEL**

**General:** a bounded depiction of Reality, including the rules governing evidence states and their evolution.

**RPVDN:** listed-property evidence, MLS records, and a rolling completed-sale population.

**THE DECISION FUNCTION (DF)**

**General:** a configured mechanism that transforms authorized evidence into outputs; AI is not required.

**RPVDN:** a configured ChatGPT model operating under the selected specification.

**THE UNIT OF MEASUREMENT**

**General:** the defined operative result derived from DF outputs.

**RPVDN:** $\overline{F}$, the arithmetic mean of ten qualifying outputs, used for LOOCV estimates and subject-property predictions.

**THE INVESTIGATION**

- Represent and empirically characterize the initial evidence state using outcome-redacted Leave-One-Out Cross-Validation (LOOCV).
- Verify execution validity through experimental controls and quality control before admitting results into performance evaluation.
- Evaluate empirical sufficiency against selected standards; distinguish evidence-state qualification from subject-property qualification as the laboratory proceeds.
- Follow operational predictions and the changing completed-sale population; reassess qualification and revoke automated authority when governing requirements fail.

# The Model: A Bounded Depiction of Reality

**A Model selects the aspects of Reality relevant to a defined decision.
Its adequacy is judged against that decision's purpose.**

Here, "Model" describes the decision's defined scope and evidence structure, distinct from the AI model or other computational mechanism used to implement the Decision Function. It specifies what kinds of evidence and circumstances are relevant, how they are represented, and the temporal rules that determine what is eligible at each decision point.

In RPVDN, that depiction includes listed residential properties, their MLS characteristics and photographs, completed-sale evidence, structured quantitative representations, and the rolling 12-month eligibility rule.

The Model is dynamic. Each decision point has an eligible evidence state; newly completed sales enter, and older sales leave. A state is not simply a fixed list of comparable properties.

The DF operates on represented evidence within this Model. Consistent DF behavior does not by itself establish that the Model is complete or adequate for a particular decision. That is a distinct question for later sufficiency assessment.

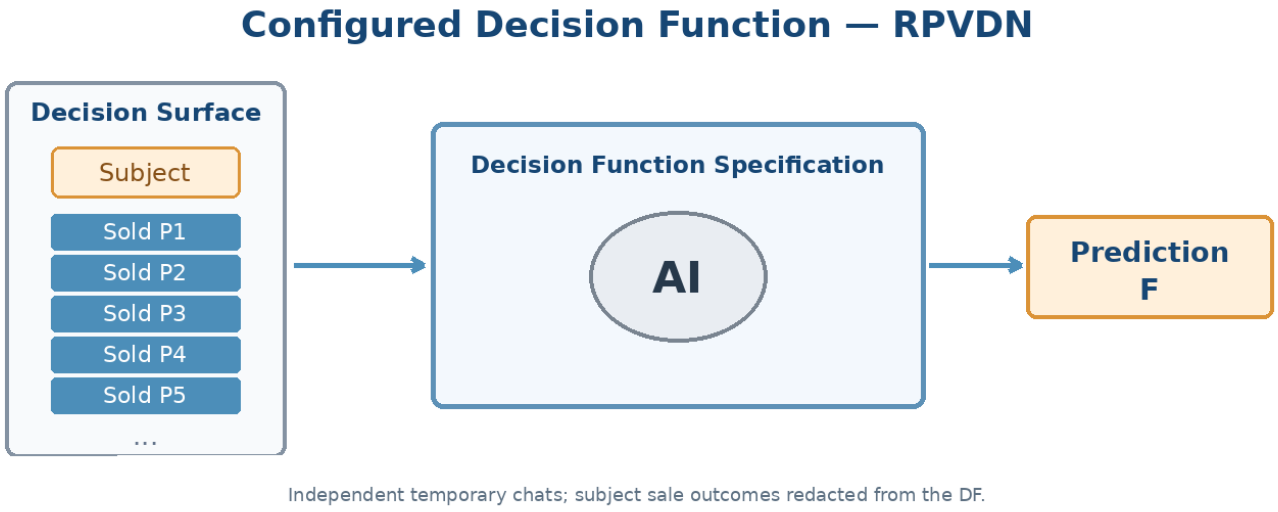
# Defining the Decision Function

## A specified decision procedure is part of the tested implementation.

The laboratory used a selected specification to direct a configured ChatGPT model in estimating or predicting listed-property sale prices. The specification constrained the task, evidence use, reasoning procedure, and required frozen output.

The specification is not interchangeable with the evidence or its representation. Changing the specification changes the configured Decision Function in use, even when the evidence state remains the same.

Changing the AI model or model version also changes the configured Decision Function. Results obtained under one configuration do not automatically establish the behavior of another.




Independent temporary chats; subject sale outcomes redacted from the DF.

The released S0 experiments ran from August 22 through October 4, 2026, in separate ChatGPT Temporary Chats, using the default model with Instant selected throughout. Experimental Records did not capture model/version identifiers.

EXPLORE / Conceptual Companion · Decision Function Engineering

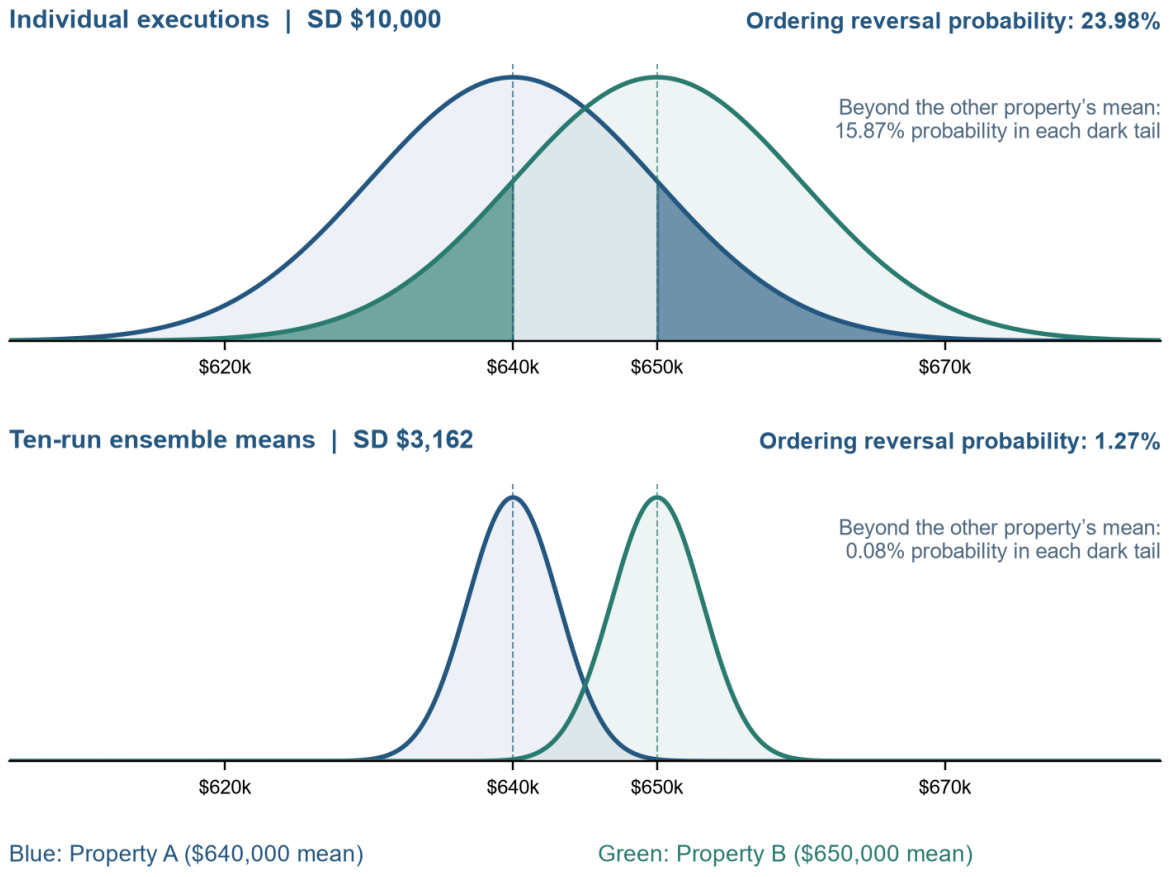
# Valid Executions Are Stochastic

## The configured RPVDN Decision Function can produce different outputs from the same evidence.

Variation can occur among valid executions even when the specification and authorized evidence remain unchanged. The laboratory averages ten qualifying outputs to produce $\overline{F}$, its operative result.

## Why averaging matters when relative valuations are close

A conceptual example: expected outputs of $640,000 and $650,000, with individual-run standard deviation (SD) of $10,000 for each property.



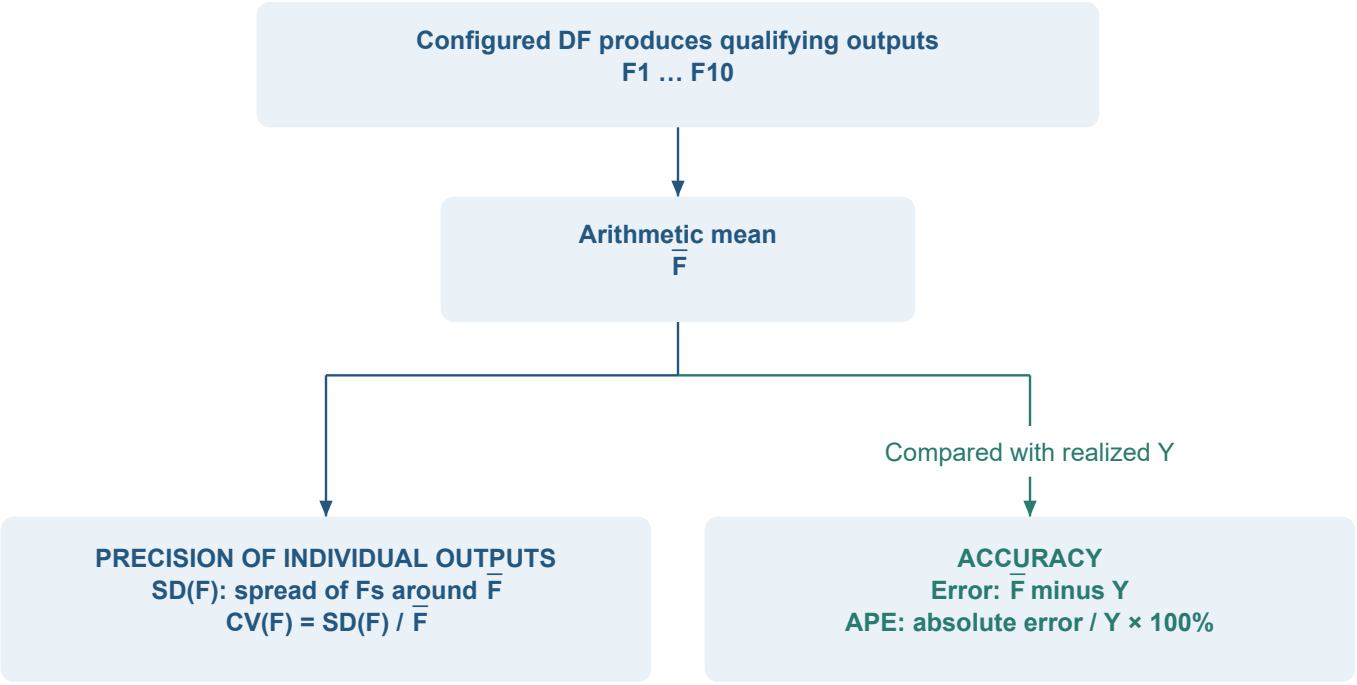

Averaging narrows variation around unchanged expected values. It reduces the risk of reversing their ordering, but does not by itself remove systematic error or bias.

Illustrative normal distributions. Runs are independent and identically distributed within each property, and the two properties' outputs are independent. Dark tails cross the other property's mean; probabilities apply separately to each tail. Ordering reversal compares both draws. Panels use different vertical scales; each distribution has total probability 100%.

EXPLORE / Conceptual Companion · Decision Function Engineering

# How Is the Operative Result Produced?

In RPVDN, ten qualifying outputs from the configured Decision Function form an ensemble. Their arithmetic mean, $\bar{F}$, is the primary operative result. The ensemble supports two distinct summaries: precision of the individual outputs and accuracy of the mean against the realized outcome.




Property-level ensemble statistics

Y enters the evaluation after the result is formed. It is not an input to the DF or the arithmetic mean. CV(F) describes dispersion among individual outputs; it is not the relative SD of repeated ensemble means.

The same measurement procedure supports held-out estimates for state characterization and subject-property predictions. These uses answer different questions.

The engineering case concerns this particular implementation: its authorized evidence, configured DF, and output transformation. Changing any of these may change what is being evaluated. Evidence qualifying one implementation does not automatically qualify another.

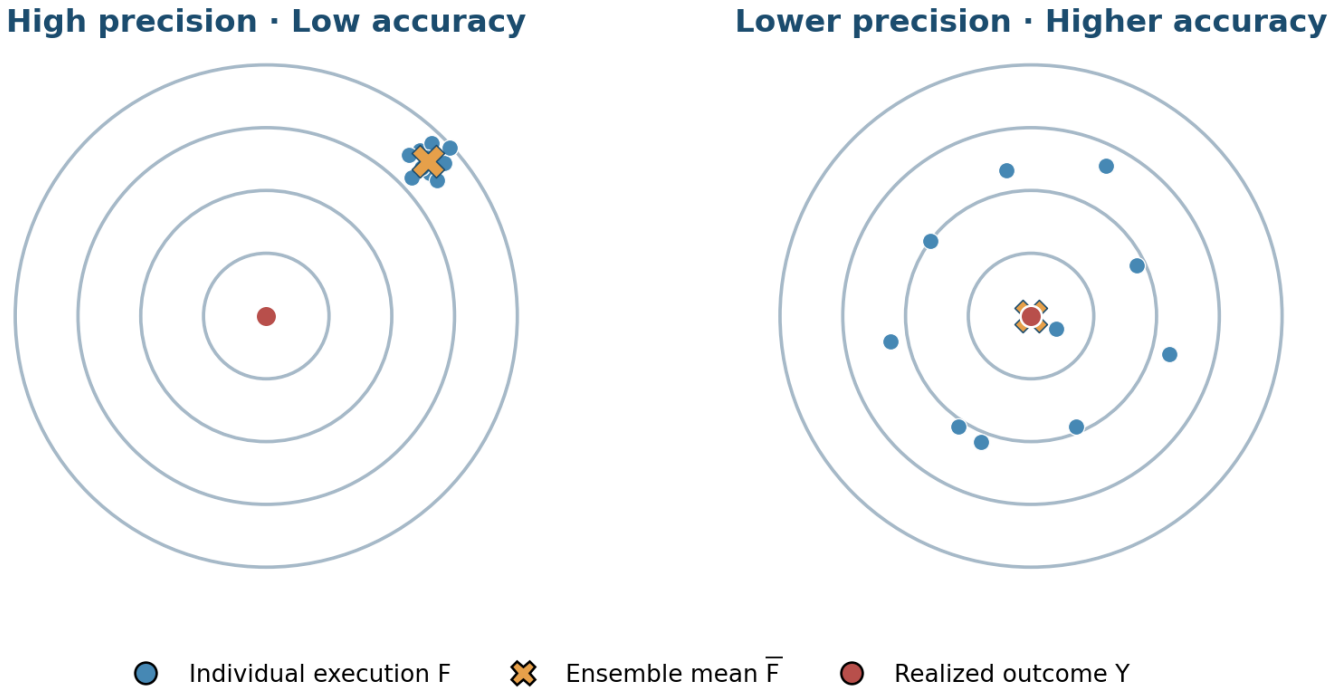
# Precision Is Not Accuracy

### A highly stable decision quantity can still be wrong when reality answers.

A tightly clustered set of estimates can be consistently wrong. Precision describes how consistently the Decision Function reconstructs its central estimate; accuracy describes correspondence with the realized sale price.

For a frozen ensemble estimate $\bar{F}$ and realized sale price Y, Absolute Percentage Error (APE) measures $|\bar{F} - Y| / Y \times 100\%$. Low dispersion among runs does not guarantee low APE when the realized sale price becomes known.

This is why stochastic precision is useful but insufficient evidence for autonomous authority.



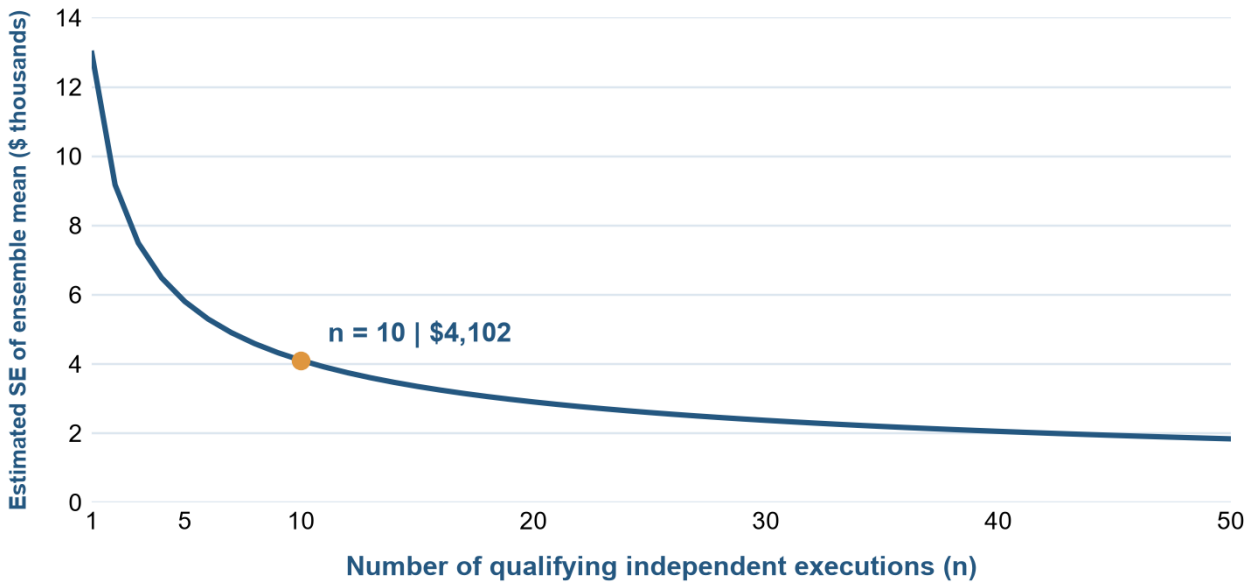

Conceptual illustration · Ten qualifying executions per target · Not RPVDN observations

# Ensemble Size Affects the Precision of the Mean

Repeated runs reveal how consistently the Decision Function behaves. Ten qualifying runs form an ensemble; their arithmetic mean, $\bar{F}$, is the ensemble estimate. The coefficient of variation, CV(F), is the standard deviation of the run predictions divided by their mean, expressed as a percentage. Lower CV indicates more consistent predictions, not necessarily greater accuracy.

The chart illustrates the theoretical reduction in variation of the ensemble mean as qualifying runs are added, with diminishing gains from each additional run.

Only qualifying runs enter the ensemble. Averaging does not repair invalid execution, inadequate evidence, or a defective specification. Ten qualifying runs were a practical laboratory convention, not a validated minimum or optimal ensemble size.




Theoretical 1/√n scaling; observed P3 SD = $12,971 (50 qualifying S0 + Table runs).

EXPLORE / Conceptual Companion · Decision Function Engineering

# Representing the Initial State

### State 0 is a defined eligible completed-sale evidence population, represented under the Model's rules.

In this RPVDN instantiation, State 0 consisted of nine eligible Seaside properties sold during 2025. Each property contributed authorized MLS characteristics, photographs, listing and transaction evidence, represented in a Decision Surface Snapshot (DSS) and a structured quantitative evidence table.

The eligible sales population is the evidence state. Its accuracy and precision characteristics are not intrinsic properties of those sales: they emerge when a particular configured Decision Function and measurement procedure are used to examine the state.

The laboratory therefore used the same DF in two roles: first as an observer generating LOOCV estimates to characterize the state, and separately to produce predictions for individual listed subject properties. The next page introduces the first role.

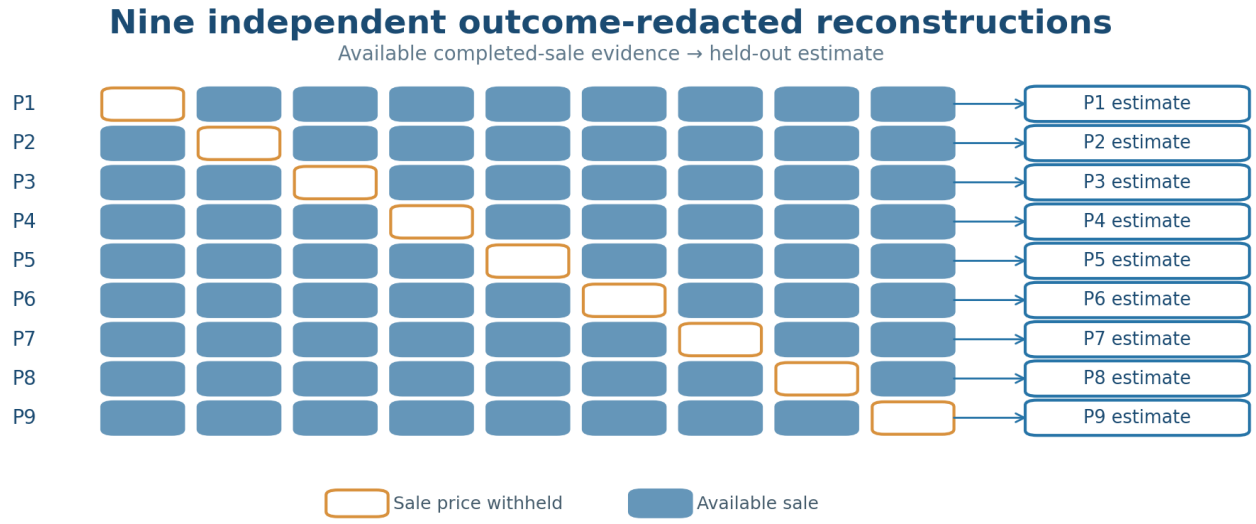
# Hide One Outcome. Estimate Its Price. Reveal Reality.

## Outcome-redacted estimates characterize the evidence state under the selected DF.

Leave-One-Out Cross-Validation (LOOCV) rotates each completed sale into the held-out position. Its realized sale price Y is withheld while the configured DF uses the remaining authorized evidence to generate qualifying independent outputs and an $\overline{F}$ estimate.

After the ensemble estimate $\overline{F}$ is frozen, Y is revealed and compared with it to calculate error. Within-ensemble CV(F) measures dispersion among the qualifying outputs.

Across the nine held-out properties, accuracy, bias, and execution-precision statistics characterize State 0 for this DF, evidence representation, and measurement procedure. These measured characteristics can be assessed against defined sufficiency standards.




EXPLORE / Conceptual Companion · LOOCV and Evidence State Characterization

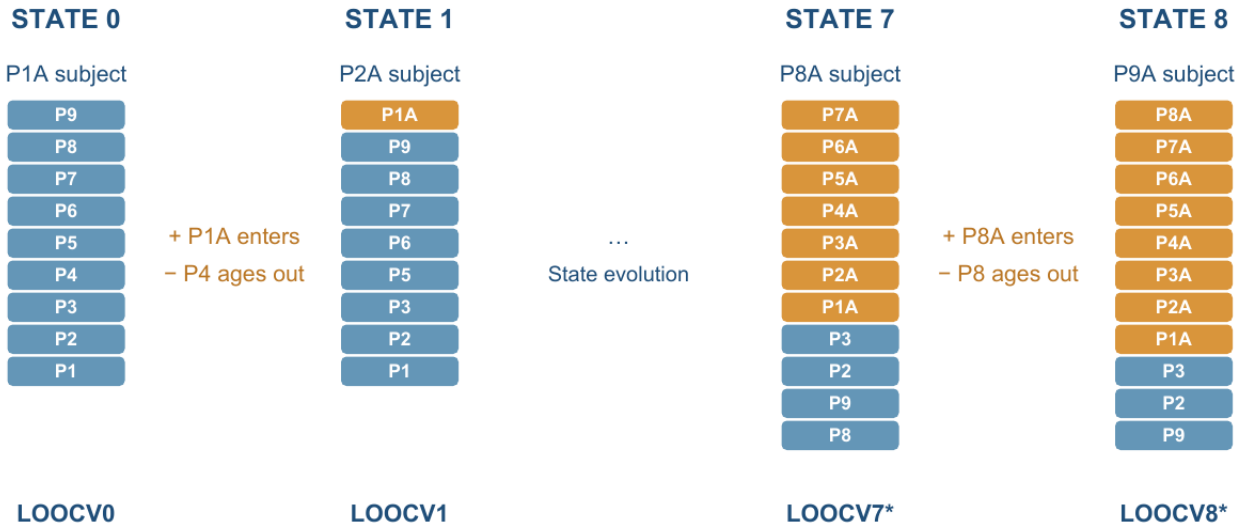
# The Laboratory Moves Forward in Time

## The Model includes the rules by which eligible evidence states evolve.

For each retrospective sale-price prediction, the laboratory assembled the completed-sale population eligible under the rolling 12-month rule.

Under the pre-established evidence rule, P1A entered after its sale became eligible, while P4 aged out.

### Evidence-State Transitions at Prediction Checkpoints




Blue: original completed sales  Orange: subsequent completed sales

* LOOCV7 and LOOCV8 were not performed; they illustrate proposed RPVDN state characterization.

# One Decision Function: Estimates and Predictions

**SAME CONFIGURED DF · SPECIFICATION S0**

Each reported result is the mean of ten qualifying executions.

| EVIDENCE AND TASK | OUTPUT | USE OF THE RESULT |
|---|---|---|
| **LOOCV — Estimates & State Characterization**<br>Each completed sale becomes a held-out subject; its outcome is withheld. The remaining eligible state supplies comparative evidence. | **Fresh held-out sale-price estimates** | Compare with Y to produce state metrics; assess those metrics against prospectively selected ES standards. |
| **OPVAL — Retrospective Predictions**<br>P1A–P9A: use MLS evidence available through the subject's closing date, with its actual sale price withheld. | **Frozen retrospective sale-price predictions** | Y was already known to the investigator. Compare after freezing the prediction to evaluate operational performance. |
| **PROSP — Prospective Predictions**<br>P1P: use the candidate evidence and eligible completed-sale state available before the subject's sale outcome is known. | **Frozen prospective sale-price prediction** | Y arrives later. P1P's prediction was $714,250; the realized sale was $720,000 (0.80% absolute error). |

**A candidate's earlier prediction does not enter a later state's LOOCV metrics. Fresh held-out executions produce the estimate, using that state's authorized evidence.**

Y is the realized sale price in all three settings. The purpose of the output and the timing of outcome availability differ.

# Valid Execution Before Performance Measurement

## Evidence and Representation Integrity → Valid Execution → Performance Evaluation

A fluent Decision Function can produce a coherent and numerically plausible result while failing to perform the decision it was authorized to make. It may use evidence that should have been excluded, omit evidence that should have been included, misread material evidence, violate outcome redaction, or depart materially from the governing specification.

That creates an ordering requirement. Performance cannot be meaningfully evaluated until the execution itself qualifies as an observation of the intended system.

Quality Control (QC) therefore asks a deliberately narrower question than “Was the estimate good?”

It asks:

“Did this execution faithfully perform the decision we intended to evaluate?”

# A Correct Answer Can Come From an Invalid Execution

### Outcome correspondence does not establish procedural validity.

Accuracy cannot retroactively validate an execution that did not perform the authorized decision. Nor should an invalid execution enter the population used to characterize the behavior of the properly engineered Decision Node merely because its output later looks favorable.

QC is therefore outcome-independent. A valid but inaccurate execution remains evidence. An invalid but accurate-looking execution does not become valid.

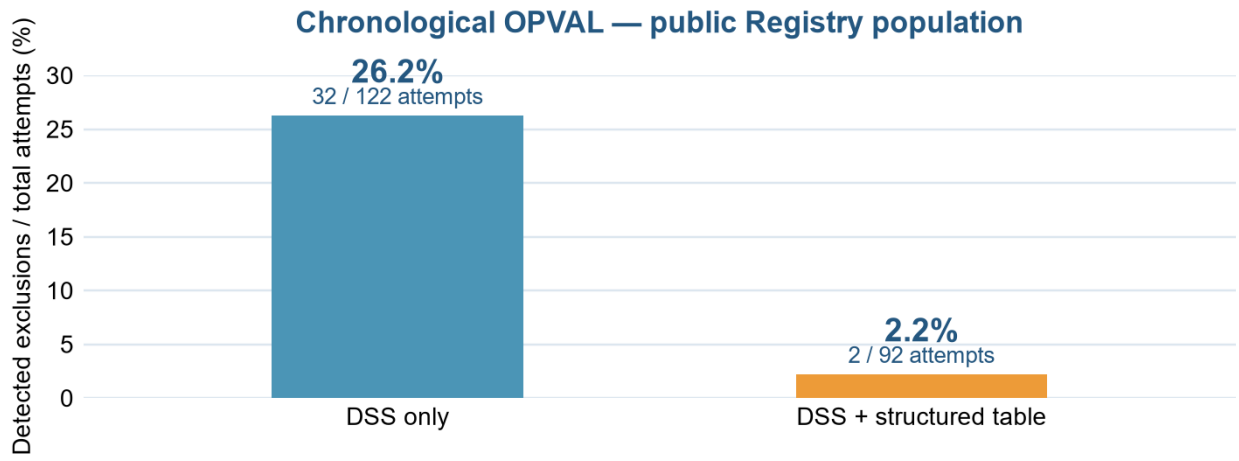
This separation is foundational: otherwise, knowledge of outcomes can influence which executions are retained for performance evaluation.

# Representation Became an Engineering Variable

The early Decision Surface Snapshot (DSS)-only experiments exposed a substantial burden of detected invalid execution. A Structured Quantitative Evidence Table was introduced to test whether a more explicit representation could reduce that burden.

The intervention did not add new market evidence. It changed how selected authorized evidence was represented to the Decision Function.

In the chronological OPVAL population, detected exclusions were 32 of 122 attempts (26.2%) with DSS-only evidence and 2 of 92 attempts (2.2%) with the structured table. These populations were reviewed under the most comparable QC process.




OPVAL only. Descriptive comparison, not a controlled causal estimate. Rates reflect detected exclusions.

Finding: structured representation was associated with substantially lower detected invalid execution in chronological OPVAL.

EXPLORE / Numerical Verification · QC Exclusions

# Representation Helped. It Did Not Solve Execution Risk.

## Better representation reduced detected failure.
## Valid-execution controls remained necessary.

The Table result should not be read as “representation problem solved.” Subsequent experiments continued to expose reconstruction and transformation errors even when structured evidence was available.

QC can detect only the failures its procedures are capable of identifying. Admission means that no identified disqualifying error was found under the applicable process; it does not prove that an execution is error-free.

The narrower empirical conclusion is stronger precisely because it is bounded: structured representation materially reduced detected invalid execution in this implementation, while residual execution risk remained.

EXPLORE / Conceptual Companion · Quality Control and Execution Integrity

# Retain the Full Experimental Record

### Exclusion from an ensemble does not remove a record's evidentiary value.

Only qualifying executions enter ensemble calculations and the performance measures derived from them. The repository retains excluded executions alongside their QC dispositions and replacement histories. This makes the construction of the qualifying population open to inspection.

The preserved record connects an ensemble result to its governing specification, supplied evidence, individual executions, QC decisions, and any replacements. A reviewer can recalculate the qualifying ensemble from its constituent outputs and examine why other executions were excluded.

Retaining excluded records permits investigation of:

- **Failure mechanisms.** Examine weaknesses in the specification, evidence representation, interpretation, or output handling that qualifying results alone may conceal.
- **Engineering responses.** Develop corrections, plan further experiments, and identify any required requalification.
- **QC judgments.** Challenge an exclusion and inspect its replacement, rather than treating a recorded disposition as proof that the QC judgment was correct.

The P4A case makes that scrutiny concrete by retaining the excluded execution alongside its replacement and the recorded exclusion rationale.

Traceability supports audit, independent inspection, and the continuing engineering cycle. It does not by itself establish accuracy, adequate controls, or sufficiency for autonomous execution.

EXPLORE / Worked Case · P4A

# LOOCV: Empirically Characterizing an Evidence State

## Held-out estimates provide evidence for assessing the state against defined sufficiency requirements.

An evidence state comprises the eligible completed-sale population. Under the LOOCV procedure, the configured DF produces held-out sale-price estimates from which accuracy, bias, and execution-precision statistics are calculated to characterize that state.

For each eligible completed sale in the state, withhold its realized sale price and use the DF to estimate it from the remaining authorized evidence. Freeze the qualifying ensemble estimate, then compare it with the realized outcome. Repeat across the population.

The resulting measurements support assessment of the state against defined sufficiency requirements. They do not, by themselves, establish the entire engineering case.

# Nine Outcomes, Not Ninety Predictions

For each held-out property, ten qualifying executions produce sale-price estimates whose variability is measured. Their arithmetic mean forms one frozen ensemble estimate.

The nine held-out properties provide nine realized sale outcomes. The constituent executions do not create ninety independent observations of estimation accuracy.

**Before Y:** the individual estimates and CV measure variation among the DF's estimates for that property.

**After Y:** the frozen ensemble estimate is compared with Y to calculate Absolute Percentage Error (APE).

Across the nine held-out properties, property-level results are summarized to characterize the state's accuracy, bias, and execution precision.

EXPLORE / Conceptual Companion · Decision Function Engineering

# State 0: Initial Empirical Characterization

The original evidence State 0 contained nine completed sales: P1 through P9. The initial cross-validation exercise on this state is designated LOOCV0; the corresponding exercise on the later State 1 is LOOCV1. These labels name the tests, not the evidence states themselves.

The initial evidence state was empirically characterized using the configured DF under the specified LOOCV procedure. These measurements describe the state for this implementation, not a universal performance guarantee.

| Metric | State 0 (LOOCV0) |
|---|---|
| Mean APE(F) | 3.23% |
| Median APE(F) | 2.70% |
| Worst APE(F) | 6.94% |
| Share with APE(F) ≤ 5% | 6 / 9 |
| Share with APE(F) < 10% | 9 / 9 |
| Mean CV(F) | 0.93% |
| Mean signed % error | +2.24% |

These metrics illustrate one approach to state characterization. Other metrics may be selected to suit the decision and governing requirements; alternative characterization methods may call for different measures.

# Measurement Does Not Choose the Sufficiency Standard

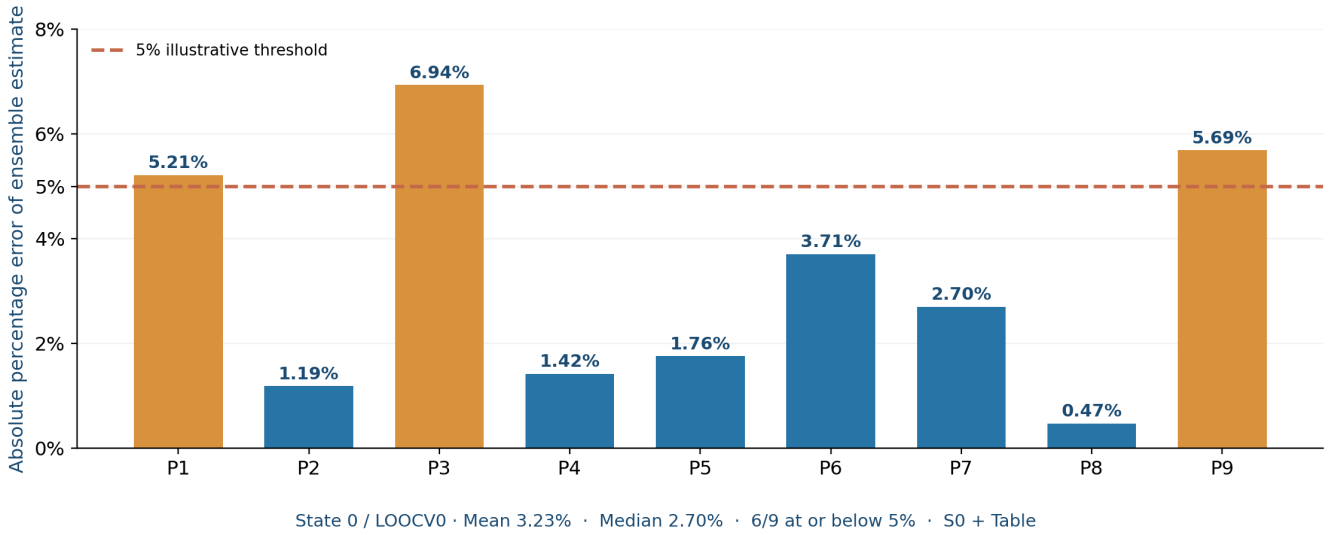
### Measurements characterize the evidence state for the configured Decision Function. They do not, by themselves, establish sufficiency for the intended use.

Estimates produced by the configured DF under the LOOCV procedure are used to calculate accuracy, bias, and precision statistics for the evidence state. Those measurements do not, by themselves, establish whether autonomous execution should be permitted.

That requires standards established in advance: what level of error is acceptable, how much variability can be tolerated, and what empirical findings would withhold or withdraw qualification.

The thresholds that follow are illustrative governance choices, not universally validated limits. They show how the same empirical measurements can support different prospective decisions.

Authority to establish or prospectively revise standards is distinct from Interpretive Sufficiency (IS), which judges whether reliance is justified under the governing standards. IS does not itself confer permission to rewrite the standards.




State 0 / LOOCV0 · Mean 3.23% · Median 2.70% · 6/9 at or below 5% · S0 + Table

# Two Illustrative Standards at State 0

Two illustrative prospective standards differ only in the required share of held-out outcomes within 5% of realized sale price.

Standard A requires at least 50% within 5%. Standard B requires at least 66.7% within 5%. Both also require mean APE ≤5%, every held-out APE <10%, mean CV ≤1%, and absolute mean signed percentage error ≤3%.

The State 0 results satisfy both standards. This establishes illustrative empirical qualification under either rule, without asserting that the rules themselves have been validated for real deployment.

| Requirement | Standard A | Standard B | Observed State 0 |
|---|---|---|---|
| Mean APE(F) | ≤ 5% | ≤ 5% | 3.23% |
| Mean CV(F) | ≤ 1% | ≤ 1% | 0.93% |
| \|Mean signed % error\| | ≤ 3% | ≤ 3% | 2.24% |
| Share with APE(F) ≤ 5% | ≥ 1/2 | ≥ 2/3 | 6 / 9 |
| Share with APE(F) < 10% | 100% | 100% | 9 / 9 |
| Illustrative result | PASS | PASS | PASS A / B |

EXPLORE / Numerical Verification · State Characterization

# Qualification Is Not Permission to Automate

### Meeting an empirical standard is necessary.
### It is not sufficient to authorize delegation.

Here, state qualification means that the evidence state, characterized using the configured DF and measurement procedure, meets the governing empirical sufficiency standards. It does not by itself establish the full engineering case, grant permission to automate, or admit an individual prediction candidate.

Empirical qualification must be considered in the context of the intended use, the consequences of failure, and the uncertainty that remains.

The authorized decision-maker or institution retains the authority to withhold autonomous execution, require collaborative decisions, or limit the scope within which autonomous execution is permitted.

When the engineering case is sufficient and the authorized decision-maker’s independent permission judgment is affirmative, the DN achieves Standing. Standing carries bounded Execution Authority for automated operation.

That permission is conditional and revocable. The next evidence state must not inherit qualification merely because the previous state satisfied the standard.

# State Characterization and Operational Prediction Are Distinct

### LOOCV: evidence-state characterization.
### OPVAL: candidate-prediction performance.

In the Operational Valuation (OPVAL) experiments, the configured DF produces sale-price predictions for individual candidates using each property's eligible evidence.

The retrospective predictions follow the 2026 sales sequence. Each uses MLS evidence available through the subject's closing date, with its actual sale price withheld. The prediction is frozen before comparison with Y.

State characterization and candidate-prediction performance provide distinct empirical evidence. The first supports assessment of the state against sufficiency requirements; the second shows how the DF's predictions compare with realized sale outcomes.

# Chronological Operational Evidence

Nine chronological 2026 properties, P1A through P9A, were evaluated under outcome redaction and later scored against their realized sale prices.

The configured Decision Function produced all nine retrospective predictions using the eligible evidence state for each subject property. All nine predictions fell within 5% APE in the current aggregate record.

This operational evidence is distinct from, and cannot replace, cross-validation of the current evidence state.

These summary metrics describe the same nine retrospective predictions shown on page 1; the prospective P1P result remains separate.

| Metric | OPVAL predictions |
|---|---|
| Mean APE(F) | 1.22% |
| Median APE(F) | 1.06% |
| Worst APE(F) | 4.81% |
| Share with APE(F) ≤ 5% | 9 / 9 |
| Mean CV(F) | 0.78% |

EXPLORE / Numerical Verification · Predictions

# Two Ways Operational Reality Returns

## A realized outcome provides evidence of past performance and may inform future qualification and authorization decisions.

**PERFORMANCE FEEDBACK** — A frozen prediction is compared with realized Y. In operational deployment, specific prospective operational-error revocation standards can be established. A threshold breach triggers the prescribed response under the governing ES requirements. Operational performance also informs the steward's continuing IS judgment, whether or not a specific revocation threshold has been established.

**EVIDENCE-STATE FEEDBACK** — An eligible completed sale can enter the rolling evidence window while an older sale expires; the changed state is then subject to fresh empirical characterization before further autonomous operation.

# Operational Flow Changes the Evidence State

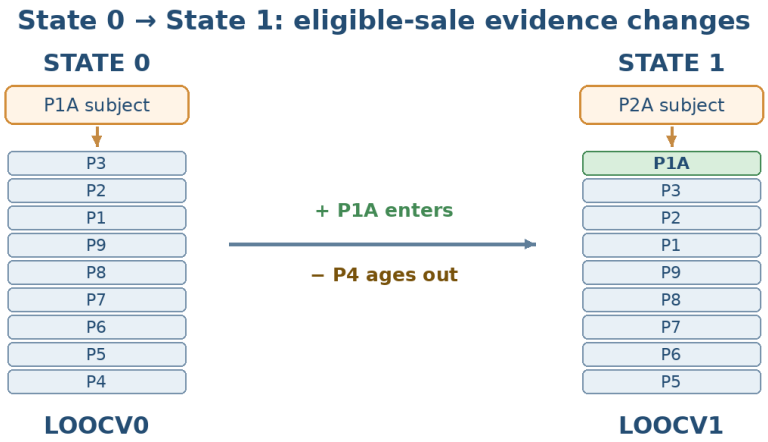
## State 0 + P1A − P4 = State 1

The original State 0 was {P1, P2, P3, P4, P5, P6, P7, P8, P9}.

Before P2A was to be evaluated, P1A had closed and became eligible completed-sale evidence. P4 had aged beyond the prospectively selected rolling 12-month completed-sale window.

The resulting State 1 is {P1A, P1, P2, P3, P5, P6, P7, P8, P9}

The transition was not selected because of its performance consequences. It followed the evidence-eligibility rule.



# Qualification Does Not Automatically Carry Forward

### Current authority requires current Standing on the evidence state governing the next automated decision.

The evidence available to the Decision Function has changed. The question before P2A is therefore not whether the prior State 0 qualification remains valid. The question is whether State 1, empirically characterized using the configured DF, satisfies the requirements for continued automated execution.

For this RPVDN implementation, the same outcome-redacted LOOCV procedure was rerun on State 1. Each eligible completed sale was treated in turn as the held-out subject, with its sale price withheld. The configured DF produced estimates using the remaining authorized evidence; each frozen ensemble estimate was then compared with the realized sale price.

# One Decision Function, Two Evidence States

The configured DF was used in the same LOOCV procedure to produce the sale-price estimates from which State 0's and State 1's empirical characterizations were calculated. The state changed; the configured Decision Function and the prospective standards did not.

The results of LOOCV0 and LOOCV1 therefore provide a matched comparison of empirical performance across two successive evidence states.

Mean APE rose from 3.23% to 3.72% while mean CV fell from 0.93% to 0.69%; held-outs within 5% fell from 6/9 to 5/9. Improved ensemble precision did not imply improved accuracy.

| Metric | State 0 (LOOCV0) | State 1 (LOOCV1) |
|---|---|---|
| Mean APE(F) | 3.23% | 3.72% |
| Median APE(F) | 2.70% | 2.74% |
| Worst APE(F) | 6.94% | 7.05% |
| Mean CV(F) | 0.93% | 0.69% |
| Mean signed % error | +2.24% | +2.81% |
| RMSE | $23,929 | $26,105 |
| Share with APE(F) ≤ 5% | 6 / 9 | 5 / 9 |
| Share with APE(F) < 10% | 9 / 9 | 9 / 9 |

# Apply the Standards Without Changing Them

### The evidence changed.
### The governing rules did not.

State 1’s measured characteristics satisfy Standard A but fail Standard B. Results must be assessed against the standards established in advance for that assessment. Standards may be revised for future assessments by the responsible authority; those revisions do not change whether earlier results met the standards then in force.

Standard A requires at least five of nine held-out properties within 5% of Y. Standard B requires at least six of nine. The remaining illustrative requirements are the same.

State 1 produces five of nine within 5%. The engineering consequence depends on which standard was adopted prospectively.

| Requirement | Standard A | Standard B | Observed LOOCV1 |
|---|---|---|---|
| Mean APE(F) | ≤ 5% | ≤ 5% | 3.72% |
| Mean CV(F) | ≤ 1% | ≤ 1% | 0.69% |
| \|Mean signed % error\| | ≤ 3% | ≤ 3% | 2.81% |
| **Share with APE(F) ≤ 5%** | ≥ 1 / 2 | **≥ 2 / 3** | **5 / 9** |
| Share with APE(F) < 10% | 100% | 100% | 9 / 9 |
| Illustrative state qualification | PASS | **FAIL** | Passes A; **Fails B** |

EXPLORE / Numerical Verification · State Characterization

# Five Qualify. Six Required. HOLD.

## A prospective standard can change the governance consequence without changing the measurements.

Under illustrative Standard A, State 1 meets the five-of-nine concentration requirement, subject to the other criteria and permission. Under illustrative Standard B, the six-of-nine requirement is not met.

Under illustrative Standard B, State 1 therefore fails empirical sufficiency, and automated execution would be placed on HOLD before P2A. This is a conditional governance illustration, not a claim that B was historically selected.

HOLD does not invalidate prior decisions or prove the Decision Function defective. It means the prospective empirical condition for continuing autonomous authority is no longer satisfied.

# From the Laboratory to the Architecture

## The RPVDN implementation is an illustration of DJE, not the definition of DJE.

The reader has seen evidence represented, execution paths rejected, uncertainty characterized, sufficiency standards applied, and autonomous standing withheld when a governing requirement is not satisfied.

The broader engineering questions are now visible:

- Is the selected model of Reality adequate for the intended decision?
- Is that Reality faithfully represented to the Decision Function?
- Does the engineering evidence justify automated execution?
- Does legitimate human judgment permit its consequences?

The principle is simple. Implementation is messy. Engineering is the discipline that connects the two.

# General Sufficiency Architecture Emerges

**Model Sufficiency (MS)** asks whether the bounded depiction of Reality selected for the application, including its evidence states and their evolution, is sufficient for its intended purpose.

**Representation Sufficiency (RS)** asks whether the information presented to the Decision Function faithfully preserves and conveys the intended evidence under the Model, without consequential errors, omissions, or misrepresentations.

**Engineering Sufficiency (ES)** asks whether the affirmative engineering case is sufficient for the defined Decision Function, operating within its qualified domain. In RPVDN, outcome-redacted LOOCV supplies empirical evidence for assessing evidence-state sufficiency within the broader engineering case; it is not a universal requirement for other Decision Nodes.

**Interpretive Sufficiency (IS)**: the authorized human steward's independent judgment that the evidence, limitations, and operating conditions justify granting or continuing bounded automated execution.

# Interpretive Sufficiency and Standing

**ES qualifies the engineering case.**
**IS permits the consequence.**
**Standing carries authority.**

Even when ES is satisfied, autonomous execution does not follow automatically.

**Interpretive Sufficiency (IS)** is the authorized decision-maker's independent judgment that the remaining uncertainty is acceptable for the contemplated consequence.

IS may say "not yet" before permission is granted—or "not anymore" when new evidence warrants withdrawal. Passing ES does not compel IS approval; IS cannot waive failure of a governing ES requirement.

Affirmative ES together with affirmative IS permits Standing: bounded, continuing, revocable authorization for autonomous execution. Standing carries bounded Execution Authority (EA).




Both ES and independent IS are necessary; neither alone grants Standing.

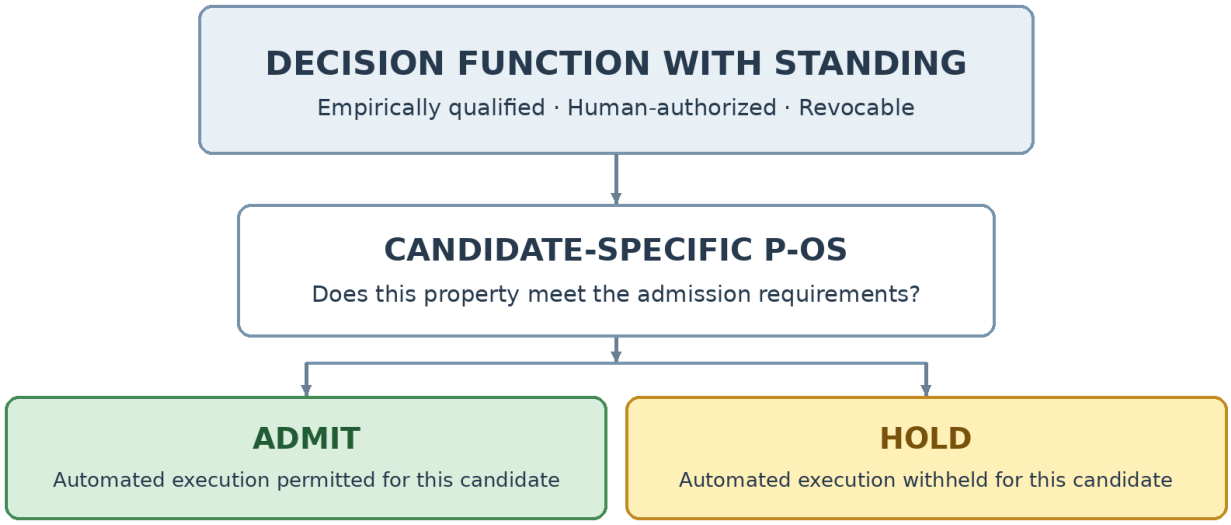
# Two Phases: Establish Standing, Then Govern Its Use

## A linear qualification sequence becomes a governed operating cycle.

The first phase establishes Standing through a clear progression: MS → RS → ES → IS → Standing. MS and RS address the adequacy of the selected Reality and its representation; ES establishes the engineering case; IS supplies independent permission in light of consequence. In this illustrative deployment, affirmative ES and IS establish bounded Standing, which carries Execution Authority (EA).

The second phase governs individual candidates through prospective Operating Sufficiency (P-OS), execution controls, and subsequent feedback. These considerations remain relevant even when automated Execution Authority is not engaged and a decision must proceed collaboratively. Standing determines whether the authorized automated path is available; it does not eliminate candidate-level obligations.

The levels must remain distinct: an individual candidate HOLD need not affect system Standing, while specified operating triggers or consequential observations may prompt system-level suspension or reconsideration.



# Prospective OS: Is There Enough Information?

## Missing essential evidence is different from being outside the autonomous operating envelope.

Standing establishes an authorized automated path. It does not establish that every candidate is ready or suitable to enter it.

The first Prospective Operating Sufficiency (P-OS) question is evidentiary: are the minimum facts required for a defensible valuation present and faithfully represented in the Decision Surface Snapshot (DSS) and structured evidence table?

If evidence required by the prospectively established P-OS standard is missing, automated execution for that candidate is placed on HOLD. The missing evidence may then be obtained, added to the evidence package, and reassessed. Automated execution for that candidate proceeds only after the P-OS requirements are satisfied. A candidate-specific HOLD does not itself revoke system-level Standing. A Decision Function might produce an estimate despite missing information; that does not establish permission for automated execution.

The governing P-OS standard specifies which evidence is required and what alternatives, if any, are acceptable. These checklist-style requirements can be prospectively specified and evaluated deterministically.

| ILLUSTRATIVE CANDIDATE EVIDENCE CHECK | Status |
|---|---|
| Gross Living Area (GLA) | **MISSING — REQUIRED** |
| Bedrooms and bathrooms | Available |
| Lot characteristics | Available |
| Condition evidence | Available |
| CANDIDATE DISPOSITION | **HOLD — obtain GLA and reassess** |

# Prospective OS: Is the Candidate Within the Envelope?

## The candidate may be sufficiently described but insufficiently qualified for automation.

An adequately described candidate may nevertheless fall outside the domain for which the Decision Function is qualified. This is a different P-OS question from information completeness.

Illustrative admission rules could compare Gross Living Area (GLA), listing price, configuration, or other material attributes with the qualified comparable-sale population. The permitted operating envelope and its measurement rules must be specified prospectively, before the candidate is evaluated.

In the hypothetical example below, the prospectively established GLA ceiling is 120% of the largest qualifying comparable sale. This 20% allowance is illustrative, not empirically calibrated. A candidate beyond the boundary is placed on HOLD for automated execution; this does not itself revoke system-level Standing. Boundaries may be revised only through an appropriately qualified and authorized engineering process, not relaxed ad hoc to admit a candidate.

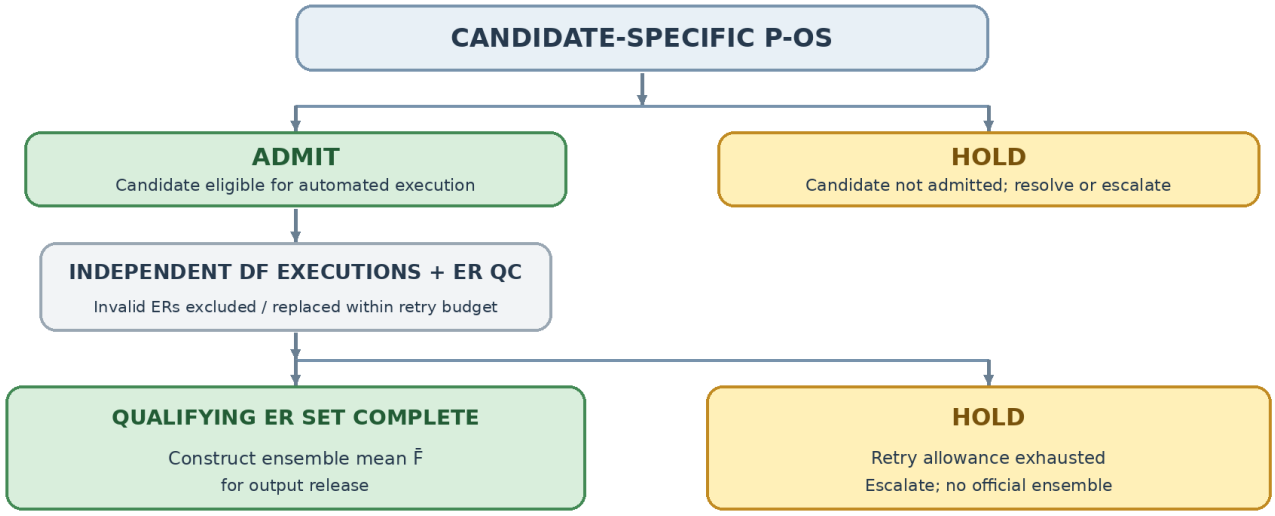
| ILLUSTRATIVE OPERATING-DOMAIN CHECK | Measurement |
|---|---|
| Candidate Gross Living Area (GLA) | **5,500 sq ft — OUTSIDE LIMIT** |
| Largest qualifying comparable | 4,400 sq ft |
| Prospectively specified ceiling | 120% of largest comparable |
| Maximum permitted GLA | 5,280 sq ft |
| CANDIDATE DISPOSITION | **HOLD — outside qualified envelope** |

# From Candidate Admission to Output Release

Passing candidate P-OS admits the property to the automated execution path; it does not guarantee a releasable ensemble. In the laboratory, only qualifying independent Experimental Records (ERs) entered the ten-run ensemble. Material QC failures required exclusion and replacement.

An illustrative prospective operational control could permit replacement executions within a defined retry budget. If the required qualifying ERs cannot be assembled within that budget, the candidate execution goes to HOLD or escalation. This does not itself revoke system-level Standing.

Repeated execution cannot guarantee detection or correction of persistent errors. The laboratory demonstrates disciplined experimental QC, not a validated deployment retry limit or comprehensive operational release protocol.



# Realized Performance: A Different Operating Feedback

### Candidate admission governs a decision.
### Realized performance can govern continuing authority.

When a frozen prediction is eventually compared with realized sale price Y, the result becomes evidence about the continuing performance of the Decision Function—not about whether the already-valued candidate should have been admitted.

An illustrative prospective rule might automatically revoke Standing and withdraw Execution Authority after two consecutive realized operational valuations each exceed 5% absolute percentage error (APE), requiring review and renewed authorization before automated execution resumes.

The laboratory does not validate “two” or “5%” as optimal thresholds. It does show how a clear rule could operate deterministically. A single exceptional failure or a troubling pattern may also warrant interpretive reconsideration before any rule is triggered.

Actual outcomes can thus affect ES through specified monitoring rules and IS through independent assessment of residual risk.

State requalification and candidate P-OS checks provide distinct prospective intervention opportunities. Realized outcomes provide an additional, later feedback path that can trigger renewed engineering or interpretive review. The laboratory has not demonstrated that its prospective controls will anticipate every deterioration in accuracy.

# Operating Evidence Returns to ES and IS

## Engineered triggers where supportable.
## Interpretive veto where consequential uncertainty remains.

Candidate admission patterns, execution observations, and realized outcomes can inform whether standing remains justified. A changed evidence state separately requires requalification under the governing state-evaluation procedure.

Where prospective engineering triggers exist, operating observations can invoke specified controls or ES reassessment. Where consequential observations lack an engineered automatic trigger, they may inform independent IS judgment.

These are complementary, not exclusive: the same observation may satisfy an engineered rule and independently prompt interpretive reconsideration. No dashboard design is needed to establish this architectural point.

IS may withhold or withdraw permission despite affirmative ES, but cannot override failure of a governing ES requirement.

# The Completed Governance Cycle

**Establish justification.**
**Authorize automation.**
**Govern its use.**
**Let reality answer.**
**Reassess.**

A defined Decision Function earns Standing through an affirmative engineering case and independent permission judgment. Standing carries bounded Execution Authority.

Standing authorizes a bounded class of automated decisions, not unconditional admission of every candidate. Prospective OS governs candidate eligibility; specified execution controls govern resulting records.

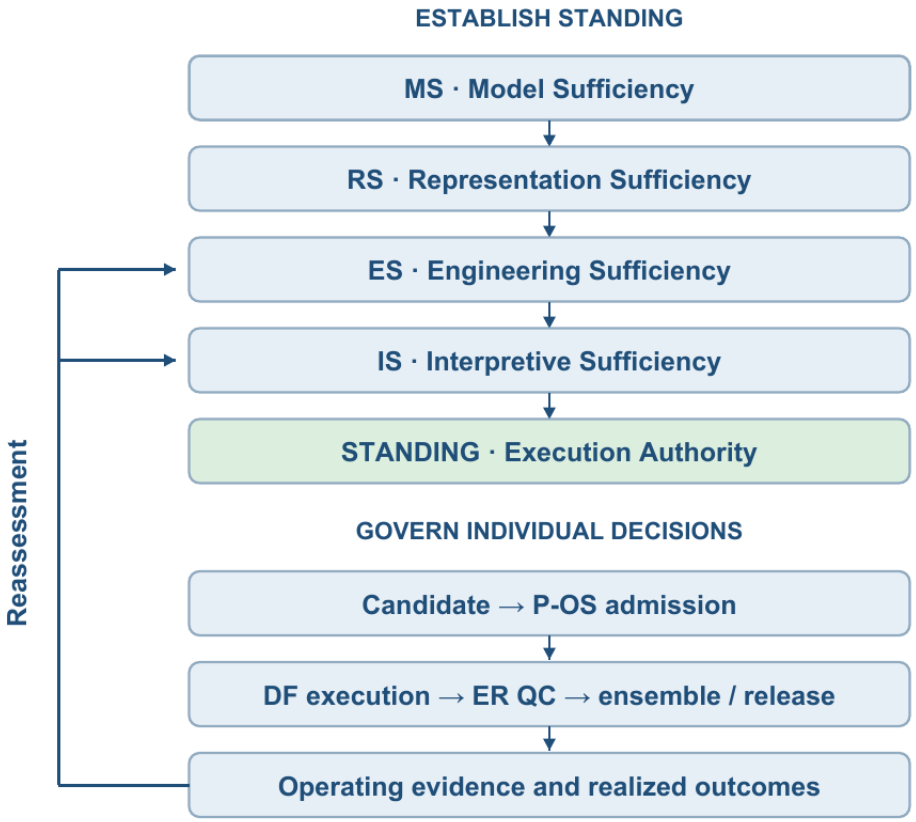
Operating observations may trigger engineered controls, inform IS, or both. Realized outcomes may change the eligible evidence state and require renewed empirical characterization. Standing continues while its governing requirements remain satisfied and affirmative IS remains in force; otherwise, Standing is revoked and automated Execution Authority is withdrawn.

The governing cycle includes opportunities to intervene before an individual prediction and to reconsider Standing as evidence evolves; later realized outcomes provide a distinct feedback path. These controls do not guarantee that every adverse outcome will be prevented.

# Governance Architecture — Overview

## Establish Standing. Govern individual decisions. Distinguish candidate disposition, system authority, and continuing engineering.

The architecture distinguishes qualification and authorization from the operational treatment of individual candidates. Standing permits bounded automated Execution Authority; candidate-specific sufficiency and execution controls remain necessary whether execution is automated or involves human–AI collaboration.




Operating evidence and realized outcomes have consequences at different levels.

**Candidate disposition.** P-OS or ER QC failures can place an individual candidate on HOLD without themselves revoking Standing.

**System authority.** Failure of governing ES requirements or withdrawal of affirmative IS revokes Standing and withdraws automated Execution Authority.

**Continuous DN engineering.** Experience informs changes to the Model, evidence representation, DF, QC procedures, and standards; material changes require renewed qualification and authorization.

EXPLORE / Conceptual Companion · Governance of Autonomous Execution

# Authority Must Remain Justified

## A system’s ability to produce a decision does not establish its authority to make one.

DJE treats autonomous authority as an engineered, bounded, and revocable permission—supported by evidence, conditioned on sufficiency, and subject to reconsideration as reality changes.

The question is not simply whether the system can decide. It is whether the justification for allowing it to decide continues to hold.

EXPLORE / Conceptual Companion · Governance of Autonomous Execution

# Scope and Limitations

*A reading note following the narrative conclusion*

## Purpose of the illustration

The RPVDN project makes governance and engineering responsibilities concrete in an inspectable Decision Node. It offers a bounded empirical case and a proposed operating architecture, rather than a complete production qualification or proof of generality.

## Coverage of the evidence

The completed-sale outcome populations contain nine properties. State characterization directly examines State 0 and one transition to State 1, not every state encountered during operational evaluation. Historical outcome-redacted predictions and the one genuinely prospective prediction provide different kinds of evidence.

## Status of the proposed controls

Standards A and B, the illustrative P-OS envelope, retry provisions, and the two-consecutive-errors operating trigger are not empirically optimized or deployment-validated. The state transition illustrates an opportunity to withhold authority before another prediction; it does not establish that prospective controls anticipate every deterioration.

## Interpretation and further engineering

QC detection is imperfect and the review process evolved during the laboratory. Execution validity, empirical performance, and permission for autonomous execution remain separate judgments. Identifying a failure, illustrating a control, and validating that control are distinct evidentiary achievements. The companion pieces and preserved records support further examination of these boundaries.