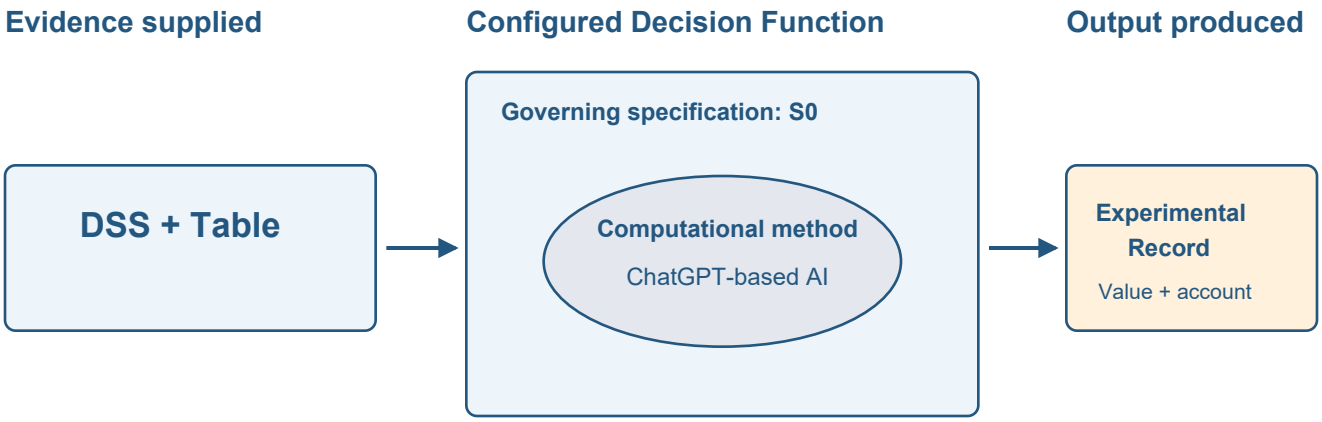
Decision Function Engineering

DJE | Working reader edition | October 28, 2026

A Decision Function (DF) transforms authorized evidence into specified outputs. Engineering it requires identifying both the configured procedure and the evidence presented to it, investigating their effects, and defining how execution outputs become the quantity used by the Decision Node (DN).

The Real Property Valuation Decision Node (RPVDN) provides a concrete illustration of these engineering choices.

Evidence Structure and DF Configuration



NT/T varies the input; S0 and the ChatGPT-based AI are retained in the configuration.

The diagram separates two engineering choices: **how authorized evidence is structured and presented**, and **how the DF is configured to act on it**. The reported experiments vary the first through NT and T conditions. The specification and computational method define the configured DF; neither should be confused with the evidence supplied to an execution.

Evidence Structure: DSS Alone versus DSS + Table

The released experiments examine a consequential intervention: supplying a structured quantitative table alongside the Decision Surface Snapshot (DSS), compared with supplying the DSS alone. The governing specification is S0 in both conditions. Other specification and computational-method choices remain part of the engineering design space, but they are not the comparison reported here.

The Decision Surface Snapshot (DSS) supplies MLS property detail reports containing property fields, remarks, photographs, and private comments. It presents the subject and eligible completed-sale evidence with the subject outcome withheld as required by the experimental task. The structured table additionally organizes selected quantitative information from the DSS. In the diagram, DSS + Table represents the T condition; NT supplies the DSS alone.

The released comparison asks what changes when the same kind of authorized property evidence is made more accessible through a structured table. **The intervention is the additional representation of evidence presented to the DF.**

Both conditions use S0 and the selected ChatGPT-based computational method—the configured DF portrayed in the diagram. The structured table is intended to restate selected quantitative information already present in the DSS, rather than introduce additional market observations. Its preparation and consistency with the source evidence are therefore front-end QC concerns.

The most relevant direct comparison in this release is OPVAL, the retrospective prediction experiments. Each subject is valued using eligible evidence through its closing date, with its realized sale price concealed from the DF. Both NT and T experiments were performed, and a comparable QC process was established and applied across them.

Condition	Evidence supplied	Qualifying executions	Excluded attempts	Total attempts	Exclusion rate
NT	DSS	90	32	122	26.2%
T	DSS + Table	90	2	92	2.2%

This is a consequential engineering result: the Table condition was associated with substantially fewer detected execution failures and fewer replacement attempts needed to assemble the qualifying population. The comparison concerns execution validity as assessed by the applied QC process; it does not by itself establish that every qualifying execution was free of error, that the table guarantees validity, or that lower exclusion rates necessarily imply better price accuracy.

The QC companion develops the detected failure mechanisms, the review process, and its qualifications. The P4A worked example makes one DSS + Table evidence package, its execution records, and the retained exclusion and replacement directly inspectable.

DF Configuration: Specification and Computational Method

The DF configuration includes **the governing specification and the computational method**. The specification defines the task, evidence-use rules, required outputs, and execution constraints. The computational method performs the task; for an AI-based method, the identified model and version and relevant execution settings belong to the configuration.

In this release, S0 governs the selected ChatGPT-based method. The NT/T intervention changes the supplied evidence representation rather than the specification or the selected computational method. Historical model identification and execution conditions must nevertheless be stated only as specifically as the retained records support. The released S0 executions were conducted August 22–October 4, 2026 in standard ChatGPT Temporary Chats, using the default model with Instant selected throughout. Experimental Records did not capture model/version identifiers, so this account does not assign a model revision to an individual execution. OpenAI's release history supplies availability context, not proof of the model serving a particular ER.

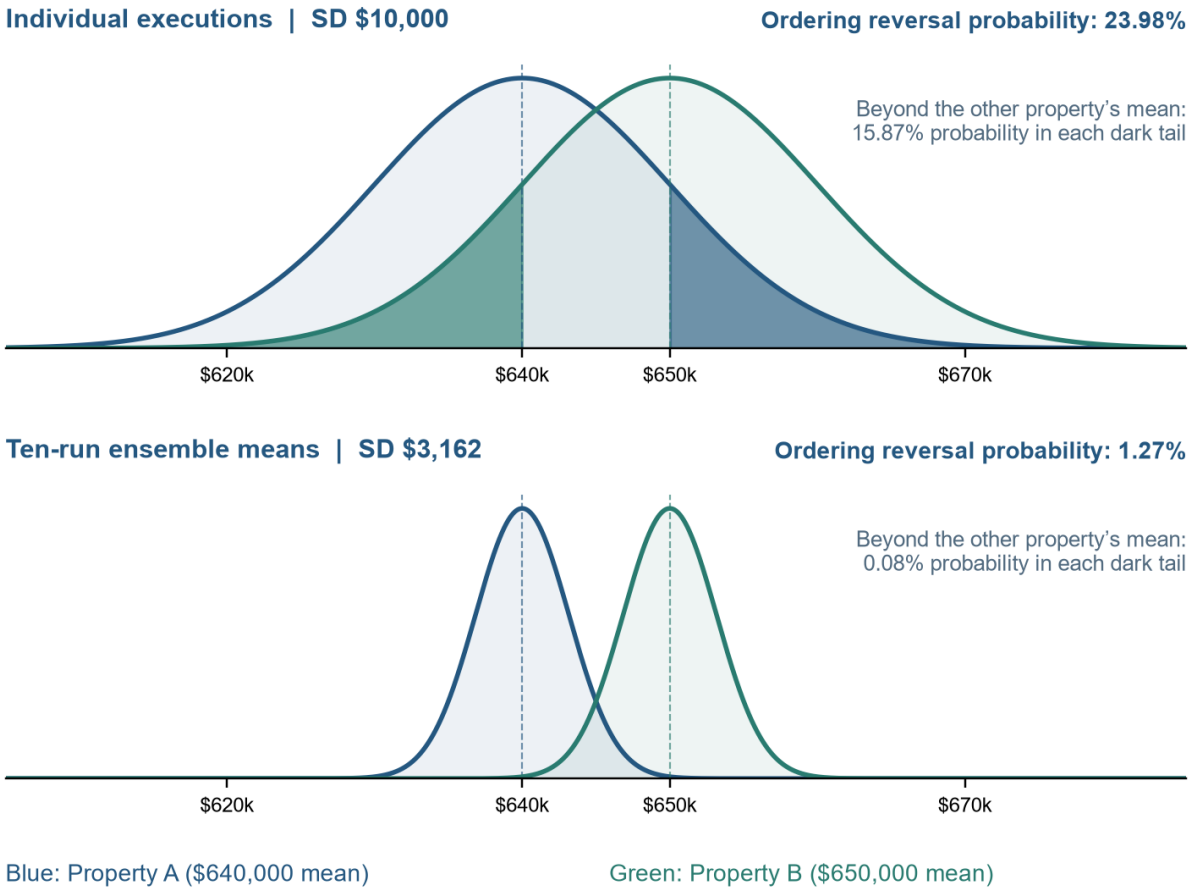
An investigator can revise the specification, change the AI model or version, or select a deterministic method suited to the task. Each change creates a configuration whose behavior needs examination. Other specifications were investigated in the broader laboratory; model substitution and deterministic alternatives are engineering possibilities, not experiments reported here. Evidence supporting one configuration does not automatically qualify another.

Stochastic Variation and the Ensemble Response

The selected AI-based method can produce different values across executions with the same authorized inputs. That stochastic variation creates an additional engineering task: define the operative result and establish how its variability will be measured and managed.

RPVDN chose an ensemble approach: the **arithmetic mean of ten qualifying executions** is its operative result. This choice defines how stochastic execution outputs are combined; post-production QC determines which records qualify for that construction.

The benefit depends on execution dispersion relative to the difference that matters for the decision. When two properties have nearby expected valuations, individual outputs can reverse that ordering. Averaging qualifying outputs reduces this sensitivity under the stated assumptions.



Conceptual illustration, not fitted laboratory results. The displayed probabilities assume normal outputs, independent and identically distributed runs within each property, and independence between properties. Averaging leaves the expected centers unchanged. Reversal concerns expected output ordering, not whether that ordering agrees with realized sale prices.

Each dark tail gives the probability of crossing the other property's mean, rather than the total pairwise reversal probability. The two tail probabilities apply separately, not as a combined event. Panels use different vertical scales; each distribution has total probability 100%. Averaging reduces execution variability; it does not by itself remove systematic error or bias.

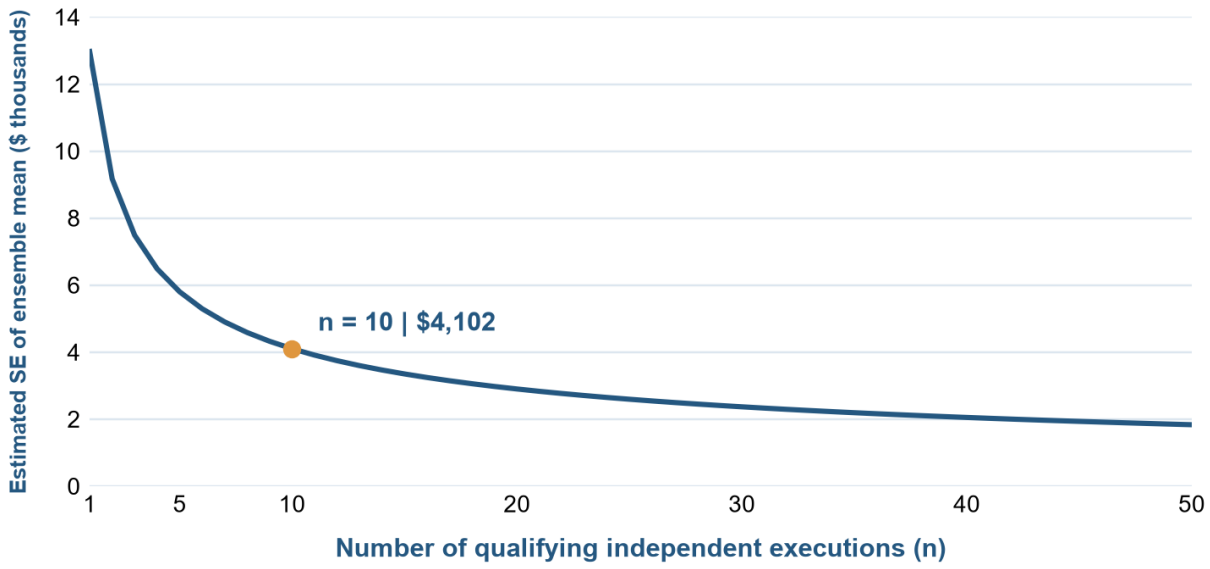
This ensemble approach responds to the selected Decision Function configuration and its stochastic outputs. A deterministic method with identical inputs and execution conditions returns the same output; repetition would not supply the same means of managing stochastic variation. Such a DN might use a single execution as its operative result. Neither an ensemble nor a single output is a universal requirement for every DF.

Precision and accuracy remain separate. Dispersion among qualifying outputs describes execution variation. Comparison of the operative result with realized sale price Y describes its error. A tightly clustered set of outputs can still produce an inaccurate mean, and increasing the ensemble size does not remove shared evidence defects or systematic interpretation errors.

Ensemble Size Is an Engineering Variable

The number of qualifying executions, n, affects the stability of the mean and the resources needed to produce it. Larger ensembles require additional execution time and QC work, including any excluded attempts and replacements. Ten qualifying executions is the adopted laboratory construction, not an established optimum for deployment.

Under an idealized model of independent outputs with stable finite variance, the standard error of the mean decreases in proportion to $1/\sqrt{n}$. This supplies a reference for investigating diminishing returns. It does not establish that every set of repeated AI executions meets those assumptions, or that additional runs necessarily improve accuracy.



The chart uses the observed P3 LOOCV0 SD of $12,971 from 50 qualifying S0 + Table executions to illustrate theoretical standard-error reduction. The curve is not a set of separately measured ensemble errors at each n, and does not establish an optimal ensemble size or improved accuracy.

The selected ensemble construction is therefore part of the measurement procedure. Its stability, cost, and limitations need to be assessed against the intended use. Changing n changes the construction to which the resulting performance evidence and standards apply.

Experimental Records, QC, and Measurement

Each DF execution produces an Experimental Record (ER), containing specified values together with reported reasoning, confidence and evidence-sufficiency assessments, and qualifications. The ER provides observable claims for inspection: did the execution identify the correct subject, use the supplied evidence consistently, respect the specification, and return the required outputs?

The explanation is a reported account, not an independently verified trace of the model's internal computation. A persuasive explanation can contain a consequential misreading. Conversely, an

inaccurate price estimate can come from an execution that meets the validity requirements. Execution validity and price accuracy answer different questions.

Post-production QC precedes admission to the qualifying ensemble. Review can admit the ER or exclude it under the applied criteria, with replacement executions used to complete the ten-member population. Defined extraction and aggregation procedures then produce the operative mean. Dispersion among qualifying outputs measures execution variation; comparison with realized Y permits error measurement. The NT/T exclusion-rate comparison examines this admission process, rather than treating excluded attempts as part of the qualifying price ensemble.

This control point complements front-end QC of the supplied evidence and table. A representation that improves accessibility is useful only if it accurately preserves the intended source information. A later review of an ER can reveal an evidence defect, but does not substitute for establishing the quality of what entered the DF.

Two Roles for the Measured Outputs

The configured DF produces held-out **sale-price estimates for state characterization** and **sale-price predictions for individual subjects**. LOOCV uses the identified state's eligible completed-sale population to construct held-out estimates. QC, calculation, and aggregation turn those outputs into accuracy, bias, and precision metrics for the state.

Predictions instead use a subject-specific twelve-month evidence window. OPVAL evaluates predictions retrospectively with realized Y concealed; PROSP produces and freezes a prediction before Y is known. The LOOCV companion explains the evidence-window distinctions and the historical cutoff used for P1P.

When a previously predicted property enters a later characterization population, the adopted LOOCV procedure generates a fresh qualifying ensemble for its held-out estimate. It does not carry forward the prior prediction. This consistently constructs the characterization population, even when the authorized evidence supplied to the Decision Function is identical. Stochasticity explains possible differences between fresh outputs; it does not itself require rerunning. The LOOCV companion explains this protocol choice and its distinction from other engineered reuse procedures.

Retain the Record and Feed Findings Back into Engineering

Retained records connect observed behavior to engineering choices. Qualifying ERs show how the operative quantity was constructed; excluded ERs and replacements preserve detected defects, judgments, and additional execution burden. Together with the supplied evidence and configuration, these records can motivate changes to evidence presentation, instructions, extraction, or QC. The table intervention illustrates this process: change the evidence interface and investigate the consequences.

LOOCV-derived state characterization contributes evidence to Engineering Sufficiency (ES) assessment for the configured DF and measurement procedure. More broadly, findings from held-out estimates, subject predictions, exclusions, and QC can inform continuing engineering across the DN. These findings help identify what to revise and what to investigate; they do not automatically qualify a revised configuration. An engineering change requires assessment of which earlier evidence remains applicable. The Governance companion develops the associated authorization responsibilities.

Scope and Limitations

The released evidence documents S0 and the selected ChatGPT-based method. Historical model and version identification must reflect what the records establish; a common configuration label is not a substitute for that identification. Other specifications and ensemble sizes up to n = 50 were investigated in the broader laboratory but are outside this released experimental record.

The OPVAL NT/T comparison is emphasized because both conditions were run with a comparable QC process. LOOCV1 and PROSP did not include NT runs. LOOCV0 included NT and T work, but its QC process was less consistent as a basis for the same comparison. These differences limit comparisons across experimental populations.

Observed exclusions are detected and recorded QC dispositions, not a complete census of all possible errors. QC itself evolved, and retained judgments can be challenged. The P4A excluded execution is an example: retaining its record does not establish that the exclusion judgment was correct. The QC companion and case guide explain the disagreement.

Ten qualifying executions define the adopted operative quantity, not an optimal deployment design. Repetition does not supply additional independent property outcomes. Shared evidence defects, systematic interpretation errors, and dependence among outputs constrain what ensemble precision establishes. Reported confidence is not an independently calibrated probability.

P1P used a purchase-contract-date cutoff rather than a closing-date cutoff. Its ERs are dated September 11, 2026. The historical prediction is retained unchanged; refining the prospective eligibility rule is a subsequent engineering task. The LOOCV companion discusses this distinction.

The demonstrated state characterization covers State 0 and one transition to State 1. It does not characterize every state encountered in prediction experiments. LOOCV's suitability is an engineering choice for this RPVDN, not a universal procedure for other DNs. The LOOCV companion addresses that choice and the interpretation of overlapping evidence populations.

The materials form a bounded, inspectable laboratory illustration, not a validated automated deployment or proof of generality. Operating controls, standards, and the sufficiency of evidence after a consequential change require assessment for the intended use.