Decision Justification Engineering (DJE)

What would justify allowing an AI system to act autonomously?

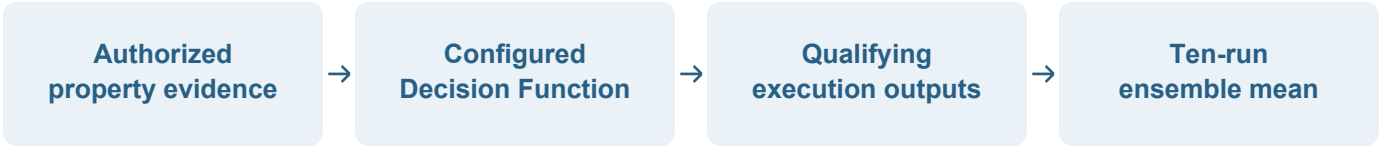
DJE explores how to establish the evidence and operating conditions that justify delegating a defined task to an automated system, and how to determine whether that justification continues to hold as conditions change.

A concrete task with observable outcomes

The laboratory uses ChatGPT to estimate and predict residential sale prices in one Florida community. Actual completed sales provide outcomes against which those results can be evaluated. The case makes choices about evidence, instructions, execution, and evaluation available for inspection.

Most individual-property predictions were examined retrospectively, with sale prices already known to the investigator but withheld from the AI. One prediction was made before its sale outcome was known. Separate held-out estimates characterize the eligible evidence population.

From evidence to an operative result



The configured Decision Function transforms authorized evidence into outputs. Each reported ensemble uses ten qualifying executions; their arithmetic mean is the primary operative result. Quality control determines which executions may enter that construction. Excluded records and their replacements remain part of the inspectable history.

What the work asks beyond prediction accuracy

When is the evidence adequate for the intended use? How do changes in its representation affect execution integrity? What changes require renewed evaluation? Who may authorize autonomous operation, and within what limits? What changes must automatically suspend that authority, even before prediction performance deteriorates?

Explore Decision Function engineering

Follow one case through its evidence and execution records

What the laboratory made observable

Selected results illustrate why the engineering case involves prediction performance, execution integrity, and the evidence state supporting the task.

Individual property predictions

9 retrospective predictions

1.22% mean APE

0.78% mean CV(F)

All nine were within **5%** of their sale prices.

1 prospective prediction

0.80% APE

0.41% CV(F)

$714,250 predicted; **$720,000** realized.

Reports + table results. Absolute percentage error (APE) compares each ensemble mean with its realized sale price; mean APE averages those errors across properties. CV(F), the coefficient of variation, measures individual-output dispersion, not variation of repeated ensemble means. The prospective case is excluded from retrospective aggregates.

Inspect prediction results and their workbook sources

Evidence representation and detected invalid execution

With the same written specification (S0), adding a structured quantitative table alongside the property reports was associated with a much lower detected exclusion rate in the chronological operational-evaluation records.



Each condition retained 90 qualifying executions. This is a descriptive historical comparison, not a controlled causal estimate. The rates concern detected validity failures, not prediction error.

Inspect exclusions, denominators, and QC limitations

A changed evidence population needs its own assessment

Leave-One-Out Cross-Validation (LOOCV) characterizes an evidence state using estimates compared with observed sale prices.

Governing standard: Mean APE ≤ 3.5% (illustrative)

Evidence state	Measured mean APE	Meets requirement?
State 0	3.23%	Yes
State 1	3.72%	**No — autonomous HOLD**

Under the proposed operating rule, failure of this governing requirement withdraws autonomous authority before the next prediction. Meeting it does not, by itself, establish full engineering qualification or human permission.

The errors shown were measured in the two states characterized in the laboratory; the 3.5% threshold is illustrative.

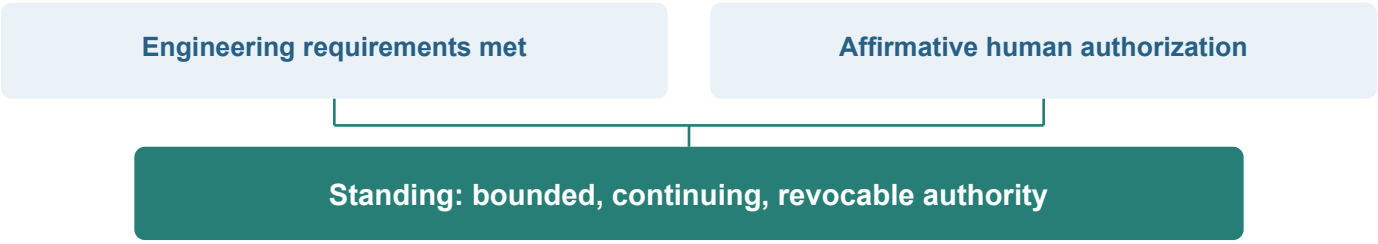
Explore state reassessment and the presentation's worked standards

From evidence to continuing authority

The proposed governance architecture connects the empirical engineering case with an independent human judgment about whether the contemplated autonomous use is justified.

Qualification and permission have distinct responsibilities

Engineering assessment asks whether the governing requirements are satisfied for the defined task, evidence, and configured method. The authorized human steward must also decide whether the evidence, limitations, and operating conditions justify permission. Human permission cannot waive a failed engineering requirement; favorable empirical performance does not compel permission.



Govern the state and the individual candidate

Continuing authority depends on the conditions supporting it. Changes in the eligible evidence population can require fresh characterization against governing standards. An individual candidate also needs adequate evidence, a place within the qualified operating domain, and a valid ensemble before release.

A candidate-specific HOLD need not withdraw authority for other candidates. A failed governing condition or withdrawal of human permission can stop autonomous operation more broadly. Deployment must specify those distinctions and the triggers, retry limits, and release checks that make them operational.

Explore Standing, candidate controls, and continuing engineering

What this release does and does not establish

The laboratory examines a limited property population, predominantly retrospective predictions, detected execution failures, and two evidence states. It does not establish broad independent validation or demonstrate an end-to-end autonomous governance system. Additional deployment controls and any application to another task require their own engineering and assessment.

The materials invite scrutiny of the evidence and reasoning, and discussion of what would be required to instantiate the approach for another consequential decision task.

Choose your next depth of review

Presentation
Website introduction
Supporting repository