Measuring Judgment Latency and Net Time Recovery

A Staged Pilot Protocol for Role-Specific AI Decision Support

Roman Bodnarchuk | WisdomTwin, Inc. | ORCID 0009-0004-3113-2118

WT-300-001 | Version 1.0 | September 24, 2026

Status of record. Working-paper protocol; not peer reviewed and not preregistered. No participants have been recruited, no pilot conducted, and no effectiveness or safety results are reported. Host, approvals, sample size and acceptance margins must be specified before implementation. No DOI has been assigned.

Abstract

Claims that role-specific AI can save hours per day require a design that distinguishes elapsed waiting from human effort, counts review and maintenance burden, and treats decision quality and confidentiality as constraints. This paper proposes a staged evaluation of the AI Judgment Twin concept. Retrospective testing establishes bounded task performance, prospective shadow mode estimates feasibility without affecting decisions, and an authorized controlled pilot compares assisted work with the existing workflow. Decision readiness is defined independently of treatment assignment. The primary outcome is elapsed time to accountable human disposition, supplemented by the narrower judgment-latency interval and net human time per eligible episode. All assigned episodes, including declines and failures, remain in the analysis. The protocol specifies quality review, cluster-aware analysis, missing-data treatment, serious-incident stopping rules, and claim wording tied to observed evidence. It is a research design and does not establish that WisdomTwin reduces time, preserves quality, or is ready for regulated deployment.

Keywords: judgment latency; enterprise AI; pilot evaluation; net time recovery; controlled trials; human oversight; decision quality.

1. Objective and evidentiary position

WT-100-001 proposes that organizations may have adequate information yet wait for an authorized person's attention [1]. That mechanism is plausible but not sufficient to establish a product benefit. An assistant might accelerate the first response while increasing correction, reducing quality, or shifting work to another role. The present protocol is designed to make those competing outcomes visible.

The primary research question is: within predefined low-consequence, recurring decision classes, does access to bounded role-specific AI support reduce time to accountable disposition compared with the existing workflow, without an unacceptable decline in independently assessed decision quality? Secondary questions concern human effort, protected focus time, reliance, error detection, appropriate decline and confidentiality failures.

The intervention is decision support. An AI answer alone does not count as the authorized person's attention or final disposition. A delegation study would require a separately defined estimand and authority model; this protocol does not silently combine human-reviewed and autonomous decisions.

2. Design overview and stage gates

Stage	Exposure	Main purpose	Permitted inference
A: retrospective	Approved historical cases and controlled synthetic challenges	Test factual support, policy resolution and authorization	Performance on the sampled cases only
B: prospective shadow	Assistant outputs hidden from operational decision makers	Establish timing feasibility, coverage and likely review demands	Feasibility; no causal time-saving inference
C: controlled assistance	Eligible teams use assistance under human authority	Estimate effects against a concurrent comparator	Causal interpretation only if the assignment and analysis support it
D: follow-up	Continued observation after initial adoption	Assess drift, workload displacement and downstream rework	Persistence within the observed setting

Progression is conditional on a documented risk assessment and stage-specific evidence. Stage B must preserve normal decision making; hidden output is logged for evaluation, not delivered to influence the decision indirectly. Stage C begins only after the host approves its data, employee-participation, incident, security and study-governance arrangements. A study in one organization cannot establish suitability across government, banking, health care or other regulated sectors.

3. Setting, eligibility and intervention

The host, jurisdiction, participating roles, source systems and decision classes must be named in the preregistration. Initial eligibility should favor recurring decisions with explicit policy, accessible evidence, reversible dispositions and a named human authority. Exclude decisions involving clinical treatment, weapons employment, hiring or dismissal, credit eligibility, legal commitments or comparable high-consequence outcomes unless separately reviewed and authorized in a protocol designed for that domain.

Each class needs a readiness checklist, quality rubric, consequence tier, required evidence, responsible role and escalation route. Novel, disputed, insufficiently documented and permission-uncertain cases remain observable but receive a decline or escalation. Eligibility is assessed using rules fixed before outcomes are seen. The assistant cannot choose which completed cases become the denominator.

The companion Trust Layer specification defines the proposed control boundary [4]. The assisted workflow provides an evidence packet, source links, discrepancies, a proposed disposition if permitted, and a route to the human decision maker. The human can accept, edit, reject or escalate it. Meetings and messages can be reduced only when the host's process permits this; the intervention must not suppress mandatory reviews to improve timing metrics. The comparator is the organization's actual current workflow, documented with the same timestamps and quality evaluation.

An intervention manifest freezes model and prompt versions, corpus snapshot rules, connectors, retrieval settings, permission policy, user interface and permitted tools. Material changes are dated and analyzed as protocol amendments rather than silently pooled as a single intervention.

4. Time origin and outcome definitions

An independent workflow rule declares readiness when the class-specific evidence checklist is complete. Record who or what applied the rule and whether later review found it incorrect. Readiness must not be generated differently by the treatment and comparator, because that would move the time origin in a way that can manufacture an apparent benefit.

Timestamp or measure	Operational definition
t_ready	First time the predeclared readiness criteria are met
t_attention	First substantive review by the authorized human, not receipt of a notification or AI response
t_disposition	Recorded human approval, rejection, justified deferment or escalation under the role mandate
t_execution	Actual implementation when applicable; distinct from disposition
Judgment latency	t_attention minus t_ready
Disposition latency	t_disposition minus t_ready
Implementation latency	t_execution minus t_ready, for applicable cases

The primary time outcome is disposition latency. Judgment latency is a prespecified secondary outcome, preserving the construct's narrower meaning [1]. Faster unsupported rejection is not a successful outcome: quality review and downstream rework are evaluated alongside timing. Reopened episodes retain links to the original disposition.

Define calendar-hour and business-hour clocks separately and designate one as primary before analysis. Report time zone, business calendar, weekends and out-of-hours handling. Cases still open at the analysis cutoff are right-censored, not dropped or assigned zero time. Some cases may never reach execution; implementation analysis must use its own risk set and competing-disposition description.

5. Human effort and net time recovery

Elapsed latency is not labor time. Ten people waiting overnight do not necessarily incur ten person-nights of work. Capture human effort across requester, decision maker, reviewer, administrator and support roles using a common task taxonomy. Each person-minute belongs to at most one bucket within an episode:

1. Evidence preparation and search. 2. Communication and coordination, including meetings. 3. Substantive review, judgment and confirmation. 4. Correction, escalation, incident handling and downstream rework. 5. Allocated system administration, governance and maintenance.

For episode e, total effort H(e) is the sum of these mutually exclusive buckets. The primary effort contrast is the mean total effort under comparator assignment minus the mean total effort under assisted assignment. Keep one-time implementation cost separate from recurring effort and report the amortization assumptions in any economic scenario. Do not add a separate “interruption recovery” estimate to already measured work time.

Synthetic arithmetic example, not a forecast: if comparable episodes require 18 person-minutes under the existing workflow and 10 person-minutes including review and allocated operations under assistance, the difference is 8 minutes per episode. At six eligible episodes per day, that corresponds to 48 minutes of capacity, not several hours. The conversion assumes stable volume and does not establish that the freed capacity becomes focused work or cash savings.

Daily time-use sampling should ask whether recovered capacity became concentrated work, additional communication, new tasks, rest or unmeasured activity. Prespecify the duration used to define a focus block and report sensitivity to that choice. Worker diaries and telemetry both have measurement limitations; disagreement is informative and should not be resolved automatically in favor of the instrument that shows larger gains.

6. Quality, safety and human factors

Independent reviewers score de-identified evidence packets and dispositions against a decision-class rubric. Where feasible, hide assignment and remove branding, while documenting any content cues that could unblind review. Two reviewers assess an overlap sample and adjudicate material disagreement. Report agreement and uncertainty; do not imply that reviewer consensus is objective ground truth.

The quality rubric should cover factual support, policy consistency, completeness, defensibility, recognition of uncertainty and suitability of escalation. Severity is assigned separately. A minor wording defect and unauthorized disclosure cannot be collapsed into the same average score. Human acceptance of an output is not a substitute for independent quality assessment.

Safety outcomes include unauthorized disclosure, prohibited action, fabricated evidence, missed mandatory escalation and failures to honor revocation. The denominator must reflect genuine opportunities for the failure: access-control challenges for leakage testing, action attempts for action-gate testing, and eligible operational episodes for operational incidents. Report both challenge-set and operational rates without combining them.

Zero observed events does not establish zero risk. For illustration, with 300 independent and identically distributed trials and no events, the exact one-sided 95% upper binomial bound is 1 minus 0.05 raised to the power 1/300, approximately 0.994%. This arithmetic illustration is not a sample-size recommendation. Correlated tests, incomplete attack coverage or heterogeneous conditions weaken that interpretation.

Human-factors measures include review effort, perceived workload, trust calibration, ability to detect errors, and willingness to challenge a recommendation. Include deliberately flawed outputs only in an approved non-operational evaluation, with appropriate study safeguards. NIST's risk-management resources provide governance context [2,3]; the specific measurements and stop rules here are author-proposed.

7. Assignment and contamination

Randomize at the smallest unit that limits contamination. If coworkers share decisions or recommendations, team or workflow-cluster assignment may be preferable to individual assignment. Document the assignment generator, blocking factors and allocation process before exposure. Balance important pre-treatment characteristics such as decision class, baseline latency, workload and role seniority where feasible.

A parallel cluster design is the default proposal. A phased or stepped rollout may be appropriate if operational constraints require it, but needs explicit control for calendar trends and sufficient clusters. A simple before-and-after comparison is vulnerable to seasonality, staffing changes, learning, backlog clearance and concurrent process reform. If randomization is impossible, label the study observational and report a

credible comparison strategy without claiming randomization-level causal certainty.

Record spillovers, off-protocol assistant use and differential adoption. Analyze all eligible episodes according to assigned condition for the primary estimate. Per-protocol or usage-based analyses are secondary and vulnerable to selection bias. Declines, tool outages and abandoned drafts remain in the assigned denominator rather than being treated as nonexistent interactions.

8. Sample size and statistical analysis

This is not an executable preregistration until sample size is justified. A preliminary baseline period or shadow study must estimate outcome distribution, cluster sizes, intracluster correlation, expected eligible volume, censoring and loss to follow-up. The host must select the smallest meaningful time benefit and an acceptable non-inferiority margin for decision quality using substantive judgment. Those choices cannot be inferred from a marketing target.

Calculate power or precision using the planned assignment unit and analysis method. A large episode count cannot compensate for too few independent clusters. If the feasible design is underpowered, designate it a feasibility study and avoid efficacy conclusions. Report all assumptions and code used for the design calculation before assignment.

For a controlled trial, estimate the intention-to-treat effect with cluster-appropriate uncertainty. Time-to-event methods should account for right-censoring; a prespecified restricted mean time to disposition over a common horizon can provide an interpretable contrast without assuming proportional hazards. Also report medians and tail behavior when estimable. The exact model, covariates, horizon, interval construction and handling of repeated episodes must be fixed in the analysis plan.

For human effort, estimate the mean difference including operations and rework, with clustering and skew handled explicitly. Do not claim a quality-preserving benefit merely because a quality difference is not statistically significant. The confidence interval must be assessed against the prespecified quality margin, with its direction and scale clearly stated. Time and quality success criteria should be declared jointly before the data are examined.

Treat subgroup analyses as exploratory unless adequately powered and prespecified. Report multiplicity handling for confirmatory endpoints. Publish model diagnostics, protocol deviations and sensitivity analyses needed to address identified threats, rather than selecting the analysis that yields the most favorable result.

9. Missingness and measurement integrity

Log missing timestamps, missing effort reports and unavailable quality reviews by assignment and decision class. Identify whether missingness arises from integration failures, participant burden, case complexity or withdrawal. Do not replace missing effort with zero. Set a prespecified strategy for imputation or bounds and disclose the assumptions under which it is defensible.

Audit a sample of timestamp events against source workflow records. Record clock drift, duplicate events, backfilled timestamps and incorrectly declared readiness. An assistant-produced log is not independent evidence that its own latency or quality claim is correct. The evaluation team needs access to a separate verification path, subject to privacy and retention controls.

Freeze an analysis dataset with a documented cutoff and version history. Retain a change log for data corrections. Where real enterprise records cannot be shared, publish the schema, protocol, analytic code, aggregate results and a synthetic demonstration dataset clearly separated from the real analysis. No synthetic outcome may be substituted for an unavailable empirical result.

10. Stopping, incident handling and governance

A confirmed unauthorized disclosure, prohibited consequential action or fabricated evidence that materially affects an operational decision triggers immediate pause of the affected workflow, containment and review by the accountable risk owner. The scope of pause may need to expand if a shared component is implicated. Restart requires documented remediation and relevant retesting. A productivity target cannot override the stop rule.

Prespecify further quantitative stop boundaries for quality and operational burden using baseline evidence and the chosen monitoring design. Repeated unplanned significance testing can distort inference; monitoring frequency and decision authority need documentation. Safety investigation can require an urgent stop regardless of the statistical plan.

Before participant exposure, determine the applicable ethics review, worker notice or consent, data-protection assessment and organizational authorizations. The protocol does not assert that formal ethics review is universally mandatory or that organizational approval is equivalent to independent ethics review. Participation data should not be reused for undisclosed performance management. Withdrawal and record-retention rules must be explained in advance.

11. Reporting and economic interpretation

A final report should disclose recruitment and exclusions, assignment, baseline balance, system versions, protocol deviations, all prespecified endpoints, uncertainty, adverse events, missingness and the flow of cases through decline and review. Publish negative and null results as well as favorable findings. Author and vendor interests remain visible.

An appropriate eventual claim might take this form: “In [setting], over [period], among [eligible classes and participants], assigned access to [version] changed [outcome] by [estimate and interval], under [review conditions], with [quality and incident results].” Bracketed fields are a reporting template, not placeholder findings for public promotion.

Recovered person-hours are capacity. Wage-equivalent value requires a stated costing model; cash savings require evidence of an actual avoided expenditure. Revenue, patient outcomes, legal quality or other downstream value require separate measurement. Neither the USD 250 million consulting estimate discussed in WT-100-001 nor a “one day per week” arithmetic restatement establishes WisdomTwin's savings.

12. Preregistration completion checklist

Before registration and enrollment, supply the host and accountable investigators; decision classes; recruitment and approvals; frozen readiness rules; allocation and analysis units; baseline data; sample-size calculation; primary time horizon; quality margin; intervention versions; comparator definition; safety taxonomy; stopping and restart authority; missing-data plan; retention and disclosure policy; and a dated analysis script. Record the registration URL only after an actual registry record exists.

The protocol is deliberately incomplete on these host-dependent fields. Supplying arbitrary values would create an apparently finished design without an evidentiary basis. Its present contribution is a reusable framework for making those choices explicitly and testing the original thesis honestly.

Declarations

Interest and authorship: Prepared for Roman Bodnarchuk, WisdomTwin co-founder and CEO, with a financial interest in the proposed intervention. Generative AI assisted in drafting and production. Independent methods and domain review have not been completed.

Data and ethics: No empirical data, participants or live enterprise interventions. The numerical examples are synthetic arithmetic. No ethics approval, participant consent, preregistration or trial registration is claimed.

Funding: No additional research-funding information was supplied for this companion draft.

Publication: Working paper, not peer reviewed; DOI unassigned.

References

1. Bodnarchuk, R. (2026). Judgment latency in regulated enterprises: A conceptual framework and research agenda for role-specific AI decision support. WT-100-001, v2.0. https://wisdomtwin.ai/research/wt-100-001

2. Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

3. National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

4. Bodnarchuk, R. (2026). The WisdomTwin Trust Layer: A testable control specification for role-specific AI decision support. WT-200-001, v1.0. Companion working paper; no DOI. The intervention's control model should be read alongside this protocol.