

Human-in-the-Loop Tax Auditing: Optimal Triage Between AI and Human Inspectors

Kun Huang

huangkun123huang@163.com

Wuhan University

Research Article

Keywords: tax enforcement, tax auditing, artificial intelligence, human-in-the-loop, algorithmic triage, public economics, audit capacity, resource allocation

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11252754/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Human-in-the-Loop Tax Auditing: Optimal Triage Between AI and Human Inspectors

Kun Huang
Wuhan University
huangkun123huang@163.com

October 3, 2026

Abstract

Tax administrations must allocate scarce human inspection capacity after screening using artificial intelligence (AI). We develop a formal economic model and large-scale synthetic computational experiments comparing release, automated document verification, and human investigation. Human review creates value through two distinct mechanisms: additional evidence and a more effective recovery procedure. Optimal allocation ranks their combined incremental economic value relative to the best available nonhuman route. Equal recovery technologies can produce an intermediate-posterior review interval, whereas procedural advantages can generate disconnected review regions. Higher error dependence eliminates information value in an equal-accuracy clone benchmark, but need not eliminate human recovery value; dependence is not generally monotone in informativeness when marginal accuracies differ. A separate commitment model incorporates concealment costs, induced voluntary tax payments, and deterrence, and shows why frozen-behavior recovery rankings can be reversed. Across 50 independent replications of 100,000 synthetic cases, welfare-optimal triage exceeds risk-only and uncertainty-first allocation by 1.662 and 2.258 arbitrary value units per case, respectively. These differences evaluate specified model environments rather than estimate effects in a tax administration. Capacity has diminishing marginal value. In the baseline quality sweep, better AI tends to reduce positive review demand while increasing the value of retained reviews. The results identify mechanisms and counterexamples for studying human allocation, without administrative taxpayer data, empirical calibration, or claims of universal human–AI superiority.

Keywords: tax enforcement; tax auditing; artificial intelligence; human-in-the-loop; algorithmic triage; public economics; audit capacity; resource allocation.

1 Introduction

Tax administrations face a scarcity problem: the number of potentially auditable returns exceeds the capacity of human inspectors. Automated screening changes the information available for allocating that capacity, but leaves the allocation problem unresolved. A high-risk case may already be readily resolved through document matching. A less obvious case may require an inspector to obtain evidence unavailable to the machine. Another case may be straightforward to diagnose but difficult to recover from without an intensive investigation. Deciding which cases warrant human attention therefore requires an economic comparison of alternative procedures rather than a ranking of noncompliance probabilities alone.

The relevant institutional choice involves at least three routes. Some cases receive no additional action. Others enter a low-cost verification process, such as information matching or a request for

supporting documents. A subset receives an intensive human investigation. In the model, automated processing is a procedural route with a cost, a recovery technology, and a burden on compliant cases. It is not an automated finding of criminal liability. Human investigation supplies additional evidence and can make recovery more effective, but consumes an inspection slot and imposes its own resource and compliance costs. Treating these routes explicitly allows an authority to ask where human intervention changes the expected value of the case relative to its best feasible nonhuman alternative.

Tax enforcement provides a substantive setting for this question because the objective of an authority differs from classification accuracy. Recovered tax is a fiscal receipt and a transfer from a taxpayer; inspection time and document-production burdens consume real resources. An intervention that raises collections can consequently have a negative value under a specified welfare accounting. Moreover, announced enforcement procedures change incentives to conceal taxable activity. Evaluating an intervention only at the existing distribution of noncompliance omits that response. These distinctions are central to the economics of enforcement (Slemrod, 2019; Keen and Slemrod, 2017; Boning et al., 2025), and they remain relevant when the initial screening information is produced by an algorithm.

The tax-audit literature already studies information, commitment, and endogenous reporting. Allingham and Sandmo (1972) provide a foundational account of evasion under uncertain enforcement. Auditing with informative signals and prior commitment appears in Macho-Stadler and Pérez-Castrillo (2002); verification and reporting interact strategically in Reinganum and Wilde (1985). Empirical work documents that information reporting, imperfect recovery, and post-audit behavior matter for enforcement (kle, 2011; Pomeranz, 2015; Carrillo et al., 2017; Advani et al., 2023). Information and audit investment choices are jointly studied by Kuchumova (2017). These contributions preclude a claim that information-sensitive auditing or deterrence is itself new. They motivate specifying which information and recovery margins are altered when an already screened case is escalated to a human inspector.

Recent research also evaluates machine learning in tax-audit selection. Battaglini et al. (2025) study audit targeting and the problem of observing outcomes selectively. Bachas et al. (2025) compare algorithmic and discretionary selection and examine how implementation by officials affects the cases ultimately investigated. Barwahwala et al. (2024) investigate a field evaluation in which model accuracy alone does not characterize operational value. Work on audit models also incorporates economic objectives and distributional concerns (Black et al., 2022). Accordingly, this article does not claim that maximizing economic value instead of predictive accuracy is a new principle for tax administration. Its narrower question concerns the incremental contribution of an intensive human route after machine information is already available.

The machine-learning literature provides an equally important set of antecedents. Learning to defer explicitly models whether a prediction should be made by a machine or passed to an expert (Madras et al., 2018; Mozannar and Sontag, 2020). Human–AI complementarity depends on the differences between the available information and capabilities, rather than on separate accuracies alone (Wilder et al., 2020; Donahue et al., 2022). Most directly, Okati et al. (2021) derive a capacity-constrained threshold rule based on the difference between conditional human and model losses. Cost-sensitive deferral and benchmarking with expert capacities further narrow the possible novelty of an economic triage application (Alves et al., 2024, 2025). The basic sorting rule used here is an adaptation of this existing allocation logic, not a newly invented principle. Evidence from medicine, moderation, or credit provides methodological precedents; it is not evidence about tax inspectors.

Tax-specific precedents also constrain the contribution. Malashin et al. (2024) study a rejection interval for classification based on indirect-check features, routing uncertain legal-entity cases toward

direct tax inspection. [Das et al. \(2026\)](#) analyze commitment auditing of strategic agents with a limited budget. Earlier costly-verification research studies audit design and redistribution ([Border and Sobel, 1987](#); [Mookherjee and Png, 1989](#)). Thus neither a tax inspection gray zone, a capacity constraint, nor strategic commitment is presented as a new standalone idea. The distinction pursued here is the separation of marginal conditional evidence from procedural recovery, combined with transparent welfare accounting and explicit boundaries on their allocation implications.

We develop a model that separates the information supplied by an inspector from the inspector’s procedural recovery advantage. The separation matters because error redundancy eliminates the first margin without necessarily eliminating the second. It also distinguishes a case that is valuable because its outcome is uncertain from a case that is valuable because an intensive procedure can recover more of an already likely liability. We then compare fiscal and welfare objectives using explicit transfer accounting. Finally, a separate commitment extension permits concealment and total tax receipts to respond to the announced review allocation. It characterizes the marginal incentive-adjusted value of review without pretending to solve a complete strategic equilibrium of the binary evidence model.

The computational component evaluates these mechanisms in transparently specified synthetic environments. It uses 100,000 cases in the baseline, 50 independent replications, a two-dimensional AI-quality and capacity grid, fixed-marginal error-dependence experiments, recovery and burden sensitivity checks, and a Latin hypercube over legal parameter combinations. All cases, signals, and outcomes are generated by the supplied code. There are no administrative returns, participating firms, observed inspectors, or estimates of institutional treatment effects. The experiments compare economically optimized allocation with plausible risk, uncertainty, and recovery-based heuristics, and distinguish the value of a rule under known model primitives from the problem of estimating those primitives.

The article contributes a bounded analysis of how scarce inspection capacity interacts with evidence acquisition, recovery procedures, and enforcement incentives. Its analytical results identify both useful restrictions and failures of familiar intuitions. The numerical comparisons quantify the consequences of those restrictions within the modeled environments and provide a reproducible test bed for future empirical work. This scope is deliberately narrower than a recommendation for a particular tax authority, staffing level, or commercial AI system. Applying the framework would require estimates of conditional inspector performance, route-specific recovery, taxpayer burden, and behavioral responses that the synthetic experiment does not supply.

The remainder of the article states the economic environment and computational design, presents analytical characterizations and numerical results, and discusses identification, institutional interpretation, and limitations. Mathematical proofs and an implementation appendix accompany the article, while the computational supplement contains the generator, optimization routines, tests, and machine-readable results.

2 Methods

2.1 Economic environment and institutional routes

There are n cases indexed by i , or an atomless unit mass for the continuous characterization. The state $\theta_i \in \{0, 1\}$ indicates whether a case contains a recoverable material discrepancy. The authority initially observes X_i , which contains prior risk covariates, the machine signal, a potential recoverable amount $R_i > 0$, and an observable complexity index C_i . The benchmark posterior is $p_i = \Pr(\theta_i = 1 \mid X_i)$. It is calibrated by construction, conditional on all information used to assign a

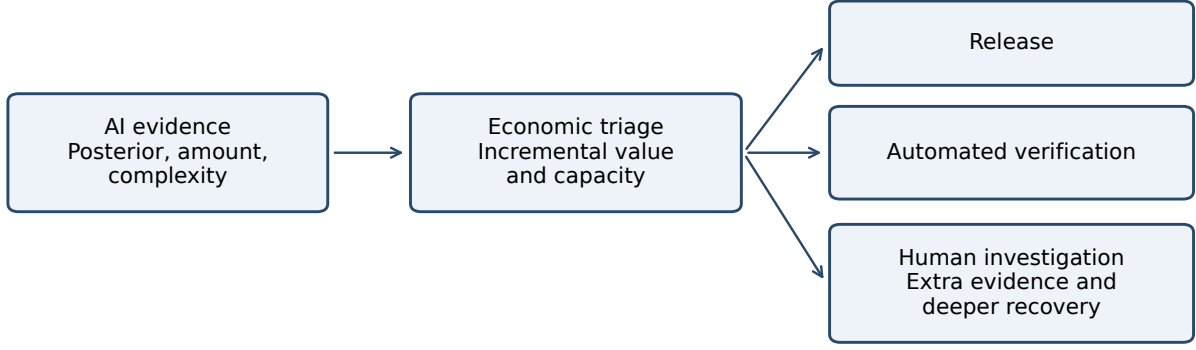


Figure 1: Model schematic. Observed machine evidence and case characteristics enter an economic routing rule. Scarce human investigation provides additional evidence and a deeper recovery procedure before a terminal processing decision. The schematic represents the formal model, rather than observations from a tax administration.

route. A raw uncalibrated score is not interchangeable with this posterior; a robustness experiment explicitly distinguishes the authority’s score-based belief from the true conditional probability.

The route N closes the case without further intervention and has normalized incremental payoff zero. Route A conducts low-cost verification at administrative cost c_A . If the discrepancy exists, it recovers $r_{Ai}R_i$, where $r_{Ai} \in [0, 1]$. If no discrepancy exists, it recovers no tax but may impose a document-production burden. This is a stylized reconciliation mechanism. The model does not authorize automatic penalties, assume that a score establishes liability, or represent an appeal process. These omissions delimit the institutional interpretation of route A .

Route H first conducts human investigation at additional administrative cost c_H . Investigation supplies a binary signal Z_i , after which the authority chooses whether to close or process the case. When it processes the case, it incurs c_A and can recover $r_{Hi}R_i$. The benchmark assumes $r_{Hi} \geq r_{Ai}$, allowing deeper investigation to improve the recoverable fraction. Human signal quality can nevertheless fall with complexity and need not exceed machine quality. The capacity constraint counts cases entering H , including reviewed cases subsequently closed. Counting only cases ultimately processed would incorrectly treat an unproductive investigation as free capacity.

A processing burden b_1 falls on a discrepant case, and b_0 falls on a compliant case. Every investigated case additionally bears b_H , whether or not processing follows. These are real private resource costs, not fines transferred to the government. The baseline uses identical terminal burdens in the two processing technologies to isolate recovery and evidence margins; heterogeneous burdens can be accommodated by changing the route payoffs. One slot per investigation is the principal restriction behind the exact sorting solution. Variable investigation hours or inspector specialization would require a different allocation problem.

2.2 Fiscal value, welfare value, and transfers

The fiscal authority values recovered tax net of administrative spending. For a processing technology $a \in \{A, H\}$, let $g_a^F = r_a R$ and $f^F = 0$. Processing has conditional fiscal value $pg_a^F - c_a$. Private burdens are recorded as outcomes but receive zero weight in this fiscal benchmark. This exclusion defines the objective; it does not assert that the burdens are socially irrelevant.

For welfare, the tax receipt and the taxpayer’s payment first cancel as a transfer. A net premium $\mu \geq 0$ can then attach to fiscal resources, representing the excess social value of public funds relative

to the private resources transferred. Thus a receipt T contributes μT after cancellation, rather than both T as a government gain and another T as a taxpayer loss. The static processing primitives are

$$g_a^W = \mu r_a R - b_1, \quad f^W = b_0. \quad (1)$$

Real administrative inputs and their financing enter as $(1 + \mu)c_a$: the unit resource cost plus the net fiscal premium of the expenditure. Compliance burdens enter once as private real costs. Thus static welfare is μ times recovery minus $(1 + \mu)$ times administrative expenditure minus private burden. The welfare calculation is a partial accounting with a constant net premium; it is not an estimate of the marginal value of public funds in any jurisdiction. We use arbitrary value units and report sensitivity to μ . No redistribution weights across taxpayers are inferred from the synthetic cases.

In the static model there is no deterrence benefit and no other real benefit from a tax transfer. Consequently, when $\mu = 0$ and every cost and burden is nonnegative, no costly intervention can improve welfare. This limiting case is economically informative: a positive static enforcement objective requires an explicit fiscal premium or another stated real mechanism. We do not insert an unexplained “compliance benefit” to evade that implication. The commitment extension below separately introduces real concealment costs and induced tax payments.

For either objective, write the corresponding primitives as g_A, g_H, f , and let $\bar{c}_A = c_A$ under fiscal value or $(1 + \mu)c_A$ under welfare. Define the machine-only processing value and its best alternative by

$$V_A(p) = pg_A - (1 - p)f - \bar{c}_A, \quad B_A(p) = \max\{0, V_A(p)\}. \quad (2)$$

The review cost entering the objective is $k_H = c_H$ for fiscal value and $k_H = (1 + \mu)c_H + b_H$ for welfare. The separate outcome calculation continues to report c_H as administrative expenditure and b_H as private burden. This convention avoids charging the burden both inside k_H and a second time in the payoff.

2.3 Human information and procedural comparative advantage

Conditional on the existing information, let $\alpha = \Pr(Z = 1 \mid \theta = 1, X)$ and $\beta = \Pr(Z = 1 \mid \theta = 0, X)$. These probabilities describe additional evidence, not the reviewer’s unconditional classification accuracy. Expected human-route value is obtained before the extra signal is observed. Let

$$\begin{aligned} J_1 &= p\alpha(g_H - \bar{c}_A) - (1 - p)\beta(f + \bar{c}_A), \\ J_0 &= p(1 - \alpha)(g_H - \bar{c}_A) - (1 - p)(1 - \beta)(f + \bar{c}_A), \\ V_H(p) &= [J_1]_+ + [J_0]_+ - k_H. \end{aligned} \quad (3)$$

The positive-part operator incorporates the option to close after either signal. It also places processing cost inside the state-weighted signal probabilities. Adding c_A unconditionally to all reviews would overcharge cases closed after investigation; omitting it would overstate review value.

Define the value of human review as $v = \text{VOHR} = V_H - B_A$. This is a net economic value relative to the best nonhuman route. It is an application of expected information and intervention value, rather than a claim to introduce a new general concept. To distinguish mechanisms, let $B_H(p) = \max\{0, pg_H - (1 - p)f - \bar{c}_A\}$ denote a counterfactual intensive recovery procedure without the extra evidence. Then

$$v = \underbrace{\mathbb{E}[B_H(p_Z) \mid X] - B_H(p)}_{I: \text{information value}} + \underbrace{B_H(p) - B_A(p)}_{P: \text{procedural value}} - k_H. \quad (4)$$

This counterfactual is a decomposition, not an additional route available for free. Information value is nonnegative because the reviewer can ignore the signal. Procedural value is nonnegative in the

maintained recovery-ordering benchmark. The total can be negative after review cost, and neither component is identified by unconditional accuracy alone.

2.4 Capacity-constrained optimization and policy comparators

The authority maximizes the sum of expected route values subject to $\sum_i 1\{a_i = H\} \leq H$. Without an inspection slot, the case receives A if $V_A > 0$ and N otherwise. Relative to these defaults, allocation chooses at most H increments v_i . The implemented optimum sorts the positive increments in decreasing order, assigns the first H , and leaves the rest on their default routes. Sorting costs $O(n \log n)$; no neural network or mixed-integer solver is needed when economic values and equal slot requirements are known. Deterministic ties use original case order. The continuous result permits randomized allocation on a cutoff tie.

We compare eleven rules. No action is a normalized zero-value benchmark. AI only uses the best nonhuman route. Random triage selects a random set of inspection slots. Risk-only triage gives slots to the largest posterior probabilities. Uncertainty-first triage ranks $4p(1-p)$. Human only allocates investigation to the highest positive human-route values and disallows machine-only processing elsewhere. Welfare-optimal and fiscal-optimal triage use their respective increments. A recovery-complexity hybrid ranks $pR(0.1 + C)$ among cases with positive estimated welfare increments. Two diagnostic rules first exclude nonpositive welfare increments and then rank by risk or uncertainty. These value-screened rules isolate the ranking error from the cost of filling slots with harmful reviews. The hybrid uses an economically motivated ranking without the full value formula and provides a stronger comparison than random assignment.

Risk-only, uncertainty-first, and random triage fill the stated capacity and may review negative-increment cases. They represent simple capacity-filling rules, not optimized members of every possible risk-threshold family. For all reviewed cases their terminal processing decisions still use the specified welfare payoff. Using the same terminal decision isolates the assignment margin. Fiscal-optimal triage instead uses fiscal decisions throughout, so its welfare comparison also includes the objective difference. These distinctions are stated because an apparently large performance difference can otherwise reflect a change in downstream behavior rather than assignment alone.

2.5 Known values, estimation, and allocation sensitivity

The sorting method is exact only for the values supplied to it. In the benchmark those values are computed from known primitives of the data-generating process (DGP). This should be distinguished from learning a policy with observed expert decisions, where estimation, selective labels, and limited training capacity can dominate the assignment problem (Mozannar and Sontag, 2020; Narasimhan et al., 2022; Mozannar et al., 2023). The algorithm therefore provides an oracle allocation benchmark and a transparent way to diagnose which estimated primitive is causing a routing error. It does not introduce a learned rejector, evaluate a neural architecture, or report a trained model’s generalization performance.

A simple finite-sample bound clarifies this distinction. Suppose terminal rules and nonhuman defaults are fixed and known, but each review increment is estimated with $|\hat{v}_i - v_i| \leq \varepsilon$. Let S^* optimize true increments and $\hat{S}$ optimize estimated increments under the same capacity. Then the true allocation loss is at most $2H\varepsilon$. Indeed, add and subtract the estimated sums on both sets: the estimated optimizer makes the middle comparison nonpositive, while each outer error sum has at most H terms. This bound concerns allocation with fixed terminal rules. If an estimated channel also changes the action after a human signal, the rule’s true expected value must first be evaluated, and the same bound does not automatically apply to the oracle route value.

The bound is conservative and gives no estimate of an institution’s error level. It shows why a known-primitives comparison and an implementable routing system are different scientific objects. Accurate probabilities are insufficient when amounts, recovery fractions, or conditional human evidence are poorly measured. The exploratory misspecification experiment directly evaluates terminal-rule changes under a fixed true DGP instead of presuming that an estimated value remains the true value of the implemented route.

2.6 Synthetic data-generating process

The baseline generator draws prior probabilities $\pi_i \sim \text{Beta}(2, 5)$, complexities $C_i \sim \text{Uniform}(0, 1)$, and recoverable amounts $R_i = 100 \exp(0.8\varepsilon_i)$ with $\varepsilon_i \sim N(0, 1)$. These variables are independent in the benchmark. It draws $\theta_i \sim \text{Bernoulli}(\pi_i)$. The AI supplies a binary signal B_i with state-symmetric correctness $q = 0.8$. The posterior conditional on π_i and B_i follows Bayes’ rule. A continuous posterior therefore results even though the machine signal is binary. This is a deliberately tractable information experiment, not a fitted tax classifier or an empirically calibrated risk model.

Human correctness is $h_i = 0.94 - 0.22C_i$. Machine and human errors are jointly drawn conditional on the state and observables. Set $e_A = 1 - q$, $e_H = 1 - h_i$, and

$$c = \Pr(E_A = 1, E_H = 1 \mid \theta, X_{\text{prior}}) = e_A e_H + \rho \{\min(e_A, e_H) - e_A e_H\}. \quad (5)$$

The same error table is imposed in both states. Its four cells are $c, e_A - c, e_H - c, 1 - e_A - e_H + c$ and are nonnegative for $\rho \in [0, 1]$. The baseline uses $\rho = 0.25$. Marginal accuracies remain q and h_i as dependence changes. Here ρ is normalized positive dependence, not Pearson correlation. Pearson error correlation equals ρ times the feasible maximum, which is generally below one when accuracies differ. The code reports both parameters to avoid implying that correlation one is feasible for arbitrary marginals.

Conditioning this joint table on B_i determines α_i and β_i . If $B = 1$, they are $(1 - e_A - e_H + c)/q$ and c/e_A . If $B = 0$, they are $(e_A - c)/e_A$ and $(e_H - c)/q$. Impossible state–signal combinations at a perfect machine boundary have zero probability and receive harmless numerical conventions. Routing uses these conditional channels; it never treats the human signal as independent simply because a marginal accuracy has been specified.

Recovery fractions are $r_A = 0.85 - 0.55C$ and $r_H = 0.95 - 0.15C$. Administrative costs are $c_A = 3$ and $c_H = 7$. Processing burdens are $b_0 = 35$, $b_1 = 2$, with review burden $b_H = 1$, and net fiscal premium $\mu = 0.6$. Available capacity is $H/n = 0.1$. Parameter choices create both economically valuable automation and heterogeneous human advantages. They are illustrative rather than estimates, and their sensitivity is part of the experiment. The full parameter configuration and package versions are saved in a run manifest.

2.7 Computational design, uncertainty, and robustness

The production experiment first generates 100,000 baseline synthetic cases with seed 20261003. Fifty independent replications use seeds 20262003 through 20262052. Every policy in a replication is evaluated on the same cases, giving paired numerical differences. The main outcomes integrate over the conditional true state and signal probabilities. We also draw realized states and signals and report realized outcomes as an independent implementation diagnostic. Conditional expectation reduces simulation variance without converting the exercise into an empirical estimation design.

For a scalar outcome Y , the Monte Carlo standard error of its replication mean is

$$\text{MCSE}(\bar{Y}) = \frac{\text{sd}(Y_1, \dots, Y_{50})}{\sqrt{50}}.$$

For a policy difference, we calculate the paired difference within each seed before computing that standard error. These errors describe numerical integration variability across synthetic samples. They exclude parameter uncertainty, institutional uncertainty, and sampling error in administrative estimates because none of those quantities is estimated here. An optimized rule’s advantage under known primitives is partly a property of the optimization problem; repeated simulations quantify its magnitude, not the existence of a general theorem about actual authorities.

Capacity varies from zero to full availability. AI quality uses 26 values from 0.5 to 0.99, crossed with 16 capacities between zero and 0.3, giving 416 grid cells. The same seed is used across grid cells as a common-random-number device. Changing AI quality also changes the coupled conditional human channel, so this grid is not a clean Blackwell experiment holding every other signal law fixed. Error dependence varies over 21 points for three recovery and accuracy settings. Human cost uses 31 values from zero to 60. Twenty-four named environments vary prior risk, recovery distribution and scale, calibration, AI and human quality, dependence, costs, burdens, and fiscal valuation. A 200-draw Latin hypercube jointly varies ten parameters on 10,000 cases per draw.

Calibration sensitivity transforms the true posterior using $\tilde{p} = \text{logit}^{-1}(\gamma \text{logit}(p))$, with $\gamma = 1.8$ or 0.6. Policies choose using $\tilde{p}$, while outcome evaluation integrates under true p . This is a specified belief misspecification, not a realistic model of every form of miscalibration. A random scan of 100,000 legal scalar cases searches for failures of monotone risk intuitions. Analytical counterexamples independently address disconnected regions, error dependence, and AI information effects. Numerical checks corroborate calculations but do not substitute for proofs.

2.8 A separate deterrence commitment extension

The static case distribution is fixed. To examine incentives without disguising them as an extra term in static prediction, we introduce a separate population of observable types j with masses m_j . A type chooses concealment intensity $e_j \in [0, 1]$ to maximize $(T_j - F_j D_j)e_j - \kappa_j e_j^2/2$. Here T_j is the tax saving per unit concealed, F_j is the total private sanction conditional on detection, and κ_j governs real concealment cost. The total private sanction includes tax recovery and any penalty, so tax saving and recovery are not counted independently in the private response. The government commits to a review intensity $t_j \in [0, 1]$, giving $D_j = d_{Aj} + \delta_j t_j$ and consuming $\sum_j m_j t_j \leq K$ capacity. We restrict parameters so the response $e_j = (T_j - F_j D_j)/\kappa_j$ remains interior for every feasible t_j .

Let $e_{0j} = T_j/\kappa_j$ denote no-enforcement concealment. The specified incremental welfare is

$$W_j(t_j) = \frac{\kappa_j}{2}(e_{0j}^2 - e_j^2) + \mu T_j\{(e_{0j} - e_j) + e_j D_j\} - c_{Aj} - \Delta c_j t_j. \quad (6)$$

The first term is saved real concealment cost. The fiscal term includes both induced voluntary compliance and recovered evaded tax; omitting voluntary payments would undercount the revenue response. Baseline voluntary tax paid absent enforcement cancels in the incremental comparison. Penalties enter private incentives through F_j but receive no separate public-funds premium in this particular welfare specification. This treats their receipts as unweighted transfers; an alternative fiscal treatment would change the derivative and must be stated explicitly.

In this extension, c_{Aj} and Δc_j denote welfare-valued administrative costs inclusive of financing costs. Detection satisfies $0 \leq d_{Aj} \leq d_{Aj} + \delta_j \leq 1$, with $F_j \geq T_j > 0$, $\kappa_j > 0$, $\delta_j \geq 0$ and $\mu \geq 0$. The optimizer solves the concave commitment problem by a scalar dual search. One thousand equally weighted synthetic types are generated with a separate seed. We compare the continuous optimum to a binary capacity-filling rule ranked by the frozen-response marginal recovery $\mu T_j e_j(0) \delta_j - \Delta c_j$. This comparison combines ranking and intensive-margin flexibility, so the numerical gap is not

attributed exclusively to deterrence ranking. A two-type analytical example isolates a ranking reversal. The extension does not endogenize the static posterior distribution or prove uniqueness of a unified machine-screening equilibrium.

An exploratory input-sensitivity analysis was added after internal model review. It keeps the true baseline population and calibrated posterior fixed, while the decision maker assumes a different human accuracy, dependence parameter, or recovery fraction. Choices and terminal decisions use the assumed model; outcome evaluation uses the unchanged true DGP. It also compares a positive-value-screened expected recovery-increment ranking. These are single-seed diagnostics of model-input risk, distinct from the independently replicated baseline experiment and not a preregistered confirmatory analysis.

2.9 AI assistance and reproducibility

OpenAI Codex assisted primary-source literature discovery and verification, model formulation, analytical derivations and proof checking, code development, synthetic simulation implementation, counterexample and robustness analysis, manuscript organization and drafting, and English editing. Multiple AI reviewer perspectives were used to critique the work; they are internal diagnostics rather than external peer review. AI is not an author. The human author has confirmed the scientific content and retains responsibility for its accuracy and integrity. The disclosed role exceeds grammar correction and is recorded in the computational supplement.

The accompanying code regenerates the synthetic cases, route assignments, expected and realized metrics, parameter sweeps, tables, and figures with one reproduction command. Tests include exhaustive small-instance comparisons, accounting identities, signal-law boundaries, zero and full capacities, and perfect and uninformative information. Primary-source documents retained for private research are excluded from the distributed code supplement. No confidential or identifiable taxpayer data are used at any stage.

3 Results

3.1 Analytical characterization of scarce review

The first result states the allocation benchmark. It applies to both government objectives once their expected values have been specified. Its economic content is the counterfactual comparison: a case allocated to a human forgoes its best feasible nonhuman route, not necessarily zero payoff. Its mathematical content is standard unit-slot selection, closely related to capacity-constrained deferral (Okati et al., 2021; Alves et al., 2024).

Theorem 1 (Optimal unit-slot allocation). *With separable ex ante case payoffs, fixed nonhuman defaults, and one slot per human review, an optimum selects up to H largest strictly positive values $v_i = V_{Hi} - B_{Ai}$. In an atomless population with integrable gains, it selects $v > \lambda^*$, excludes $v < \lambda^*$, and may randomize at equality. The multiplier is nonnegative and satisfies complementary slackness. Finite deterministic allocations permit a cutoff interval rather than a unique multiplier.*

The statement does not require a calibrated score to be a sufficient statistic by itself. The gain can depend on amount, complexity, route-specific recoverability, burden, and the human signal law conditional on all retained machine information. A case with higher noncompliance probability can have a lower gain because the machine-only alternative is already effective or because recovery is small. Likewise, a high-uncertainty case need not deserve review if neither available procedure

can deliver enough value to cover its cost. These comparisons explain why using a probability or entropy ranking alone discards economically relevant variation.

Proposition 2 (Capacity and diminishing marginal value). *Let $d_1 \geq \dots \geq d_n \geq 0$ be the sorted positive parts of the fixed case gains, including zeros. Optimal value at integer capacity H is $J(H) = \sum_i B_{Ai} + \sum_{k=1}^H d_k$. Hence $J(H+1) - J(H) = d_{H+1}$ is nonnegative and nonincreasing. With a fixed tie order, selected sets are nested as capacity expands. The continuous fractional relaxation is concave; additional capacity has zero value once all positive-gain cases are reviewed.*

This result is about the optimized objective. Welfare-optimal allocation need not make every fiscal or classification outcome monotone. Nor does the result imply that a finite staffing expansion has a large effect: gains may already be near zero, and indivisible case slots create discrete changes. In the atomless version the shadow value falls as the review threshold moves down the gain distribution. In the finite computations, the reported shadow value is the exact benefit of the next slot. Calling this the unique dual price would overstate what the discrete optimization identifies.

Proposition 3 (Information, procedure, and review geometry). *Equation (4) exactly separates information value, procedural value, and review cost. If processing technologies are identical and homogeneous, let $G = g - \bar{c}_A > 0$, $L = f + \bar{c}_A > 0$, and $\alpha > \beta$. Gross information value is triangular in p , maximized at the economic processing threshold $t = L/(G + L)$, with peak $(\alpha - \beta)GL/(G + L)$. A nonempty human region at shadow price λ exists if and only if this peak exceeds $k_H + \lambda$. Unequal recovery technologies need not produce a connected review interval.*

The relevant ambiguity concerns an action threshold determined by amounts and burdens. It need not be greatest at posterior one-half, where entropy is largest. If G substantially exceeds L , the action threshold lies below one-half; a marginal signal can then matter most for apparently low-risk cases. Conversely, a large compliant-case burden raises the threshold. Figure 2 illustrates the equal-recovery benchmark over posterior and conditional human quality. The vertical axis changes the human evidence channel, holding economic primitives fixed. It is not a separately estimated uncertainty measure or an assertion that inspection quality can be observed costlessly.

In the general model, the procedural component can sustain review when additional information is unhelpful. An analytical example has $G_A = 1$, $G_H = 2$, $L = 1$, $\alpha = 0.9$, $\beta = 0.1$, and $k_H + \lambda = 0.83$. The strict review region is $(0.48947, 0.7) \cup (0.83, 1]$. As risk increases, the sequence can be release, human, automated, and human again. This example refutes a universal two-threshold ordering in either direction. It does not say that every tax environment has disconnected regions. Its role is to prevent the benchmark interval from being mistaken for a general restriction on the full allocation problem.

Proposition 4 (Human quality and review cost). *A Blackwell refinement of the human signal conditional on existing information weakly raises human value when processing options and costs are unchanged. Heterogeneous improvements can change the selected cases without increasing a binding review count. A common increase in review cost lowers every gain by the same amount, leaves rankings unchanged, and causes nested exits once positive-gain demand falls below capacity.*

More accurate human evidence does not generally imply a larger observed review region under fixed capacity. The number of reviews can already equal H , and an improved subset may displace previously selected cases. In contrast, a uniform cost change affects the ranking only through changes in positive eligibility, because all increments move together. Costs that vary with complexity have no such ranking invariance. The main simulation keeps the review cost common so that cost and case-selection changes can be interpreted separately.

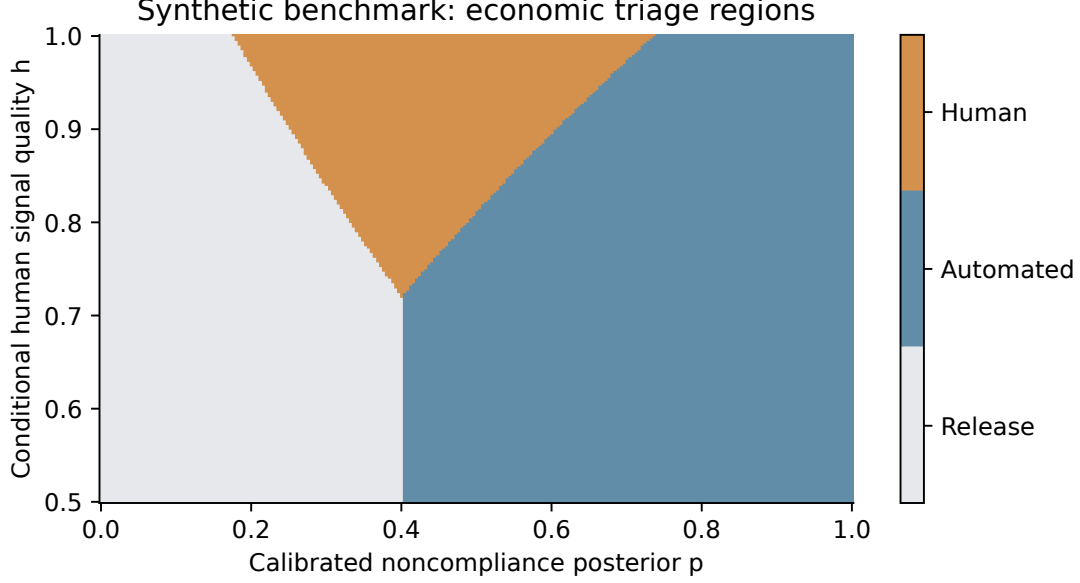


Figure 2: Economic triage regions in a synthetic computational experiment. The homogeneous equal-recovery benchmark uses $g = 60$, $f = 35$, objective-valued processing cost 3, review cost 7, and shadow price 3. Human evidence has conditional probabilities $\alpha = h$, $\beta = 1 - h$. The horizontal axis is calibrated posterior; the vertical axis is additional signal quality, not posterior uncertainty.

Proposition 5 (Dependence and information redundancy). *In the state-symmetric coupling of equation (5), when $q = h$, increasing normalized dependence is a conditional garbling toward a copy of the observed AI signal. Information value weakly declines and is zero at $\rho = 1$. At that endpoint $v = P - k_H$; review vanishes only when procedural advantage is insufficient to cover cost. When marginal accuracies differ, information value need not decline monotonically with dependence.*

The qualification is substantive. For $q = 0.9$, $h = 0.7$, observed $B = 1$, prior $\pi = 0.1$, and equal technologies with $G = L = 1$, the posterior is one-half. Gross information value is $\frac{1}{2}|\frac{2}{5} - \frac{28}{45}\rho|$. It decreases to zero at $\rho = 9/14$ and then increases to $1/9$. At high dependence the extra signal is informative in the opposite direction conditional on the already observed machine output. An authority able to interpret that conditional signal optimally can use it. A rule that always acts after the human’s positive classification would miss this possibility.

Proposition 6 (AI improvement and human demand). *If richer AI information literally retains the former AI information as an observed component, while the joint law of that former information, the state, and physical human signal, payoffs, and capacity is held fixed, optimized ex ante total value weakly increases. Neither human review uplift, positive review demand, nor the capacity shadow value has a universal comparative-static sign.*

The appendix provides constructive substitution and complementarity examples. Perfect AI evidence can eliminate an information-only human advantage. Alternatively, a more informative machine can isolate a subset on which human evidence becomes decision-relevant, creating positive review opportunities that did not exist with coarse machine information. These examples concern information systems with retained alternatives. They do not prove monotonicity for every scalar “accuracy” change, especially when a changed machine also changes the conditional distribution of inspector evidence.

Proposition 7 (Fiscal and welfare disagreement). *Fiscal-optimal and welfare-optimal allocations can rank the same cases differently. In the static transfer-neutral benchmark $\mu = 0$, with nonnegative burdens and positive administrative costs, release is optimal for all cases. A positive welfare rationale then requires a stated real mechanism, a fiscal premium, or an incentive response outside that static benchmark.*

Proposition 8 (Incentive-adjusted review allocation). *In the interior commitment model with $\mu \geq 0$, the type-level marginal welfare value is*

$$M_j(t) = \delta_j \left[F_j e_j + \mu T_j \left(e_j + \frac{F_j(1 - D_j)}{\kappa_j} \right) \right] - \Delta c_j, \quad (7)$$

and $W_j'' = -\delta_j^2(F_j^2 + 2\mu T_j F_j)/\kappa_j \leq 0$. *An optimum equalizes $M_j(t_j)$ to the capacity multiplier on interior types, with the usual boundary inequalities. Ranking only frozen-behavior expected recovery can reverse the allocation order.*

The response term incorporates saved concealment resources and additional voluntary payments as well as recovery. A two-type counterexample holds $\delta = 0.05$, $\Delta c = 0.01$, and $\mu = 1$ fixed. Type A has $(T, F, \kappa, d_A) = (2, 2, 4, 0.1)$ and type B has $(6, 12, 12, 0.4)$. Their frozen recovery gains are 0.035 and 0.020, but their marginal welfare gains at zero review are 0.125 and 0.260. With equal masses and capacity 0.25, the optimum sets $t_A = 0, t_B = 0.5$, reversing the frozen ranking. The extension uses commitment and known responses; it does not establish that an authority without commitment can implement the same outcome.

3.2 Baseline synthetic policy performance

Figure 3 shows the full model’s comparative-advantage map over calibrated risk and observed complexity. Posterior uncertainty is already determined by p , so it is not treated as an independently varying dimension. Complexity changes both the conditional evidence channel and the recovery advantage. The plotted shadow value comes from the baseline population, while amount and the observed machine signal are held fixed in this cross-section. Thus its regions explain routing mechanisms without claiming to display the entire heterogeneous population.

Table 1 reports the baseline environment. Each outcome is an expected amount per synthetic case except human share. Welfare-optimal triage uses all 10,000 available slots and produces welfare 8.379, recovered tax 20.984, and net fiscal value 19.712. These numbers are consequences of the declared primitives and remain in arbitrary units. The fiscal optimizer collects more tax but incurs much larger compliant-case processing burdens; under the stated welfare weights its welfare is negative. The difference reflects objectives and terminal actions, not poorer optimization of the fiscal objective.

Across the 50 replications, welfare-optimal triage exceeds risk-only triage by 1.662 units per case, with paired Monte Carlo standard error 0.00272. It exceeds uncertainty-first triage by 2.258, with standard error 0.00379. With a positive-value screen applied first, the remaining gaps are 0.957 for risk ranking and 1.297 for uncertainty ranking. Thus the benchmark differences include both reviewing cases with negative increments and misranking cases with positive increments. The stronger recovery–complexity hybrid falls short by 0.139, with standard error 0.00051. Its small gap relative to the simple rules is useful evidence about which modeled primitives account for allocation value.

The expected and realized calculations use different integration methods but similar replication means. This agreement checks the signal table, routing, and accounting implementation. It does

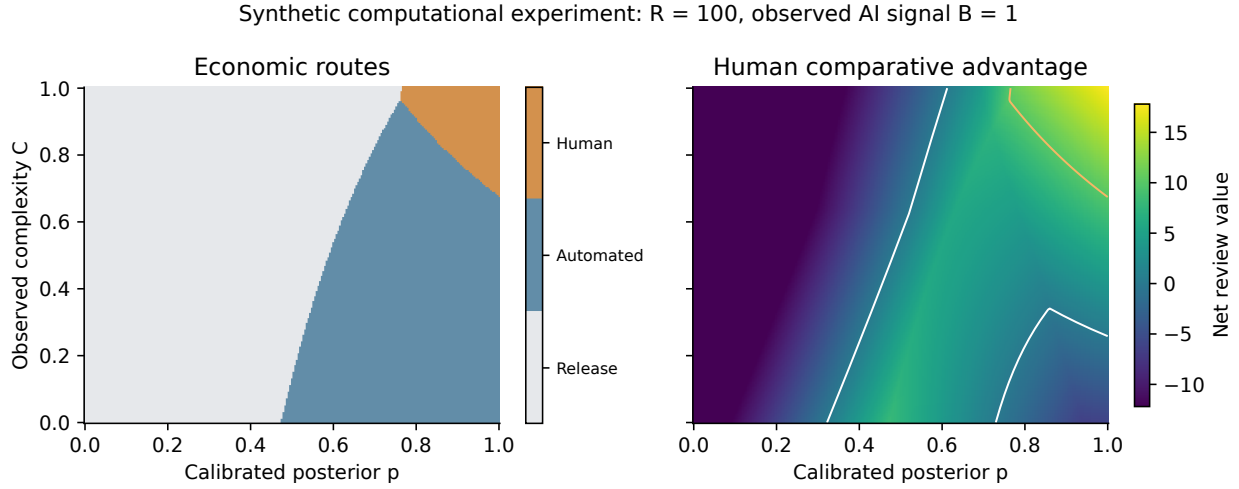


Figure 3: Human comparative-advantage map in a synthetic computational experiment. Both panels fix $R = 100$, observed machine signal $B = 1$, and baseline primitives. Left: optimal routes under the baseline next-slot capacity value. Right: net review value relative to the best nonhuman route; white and orange contours mark zero and the baseline cutoff. Complexity changes recovery fractions and conditional human evidence.

Table 1: Baseline synthetic computational experiment, 100,000 cases. Recovery, net fiscal value, and welfare are arbitrary units per case; human share is a fraction. Welfare includes financed administrative costs and private burdens.

Policy	Recovery	Net fiscal	Welfare	Human share
No action	0.000	0.000	0.000	0.000
AI only	15.430	14.827	5.835	0.000
Random triage	16.737	15.412	5.340	0.100
Risk only	18.792	17.472	6.711	0.100
Uncertainty first	18.019	16.664	6.138	0.100
Human only	16.097	15.140	7.232	0.100
Risk + value screen	19.730	18.442	7.426	0.100
Uncertainty + value screen	18.732	17.443	7.080	0.100
Optimal welfare	20.984	19.712	8.379	0.100
Optimal fiscal	28.231	25.327	-4.581	0.100
Recovery-complexity hybrid	21.364	20.056	8.239	0.100

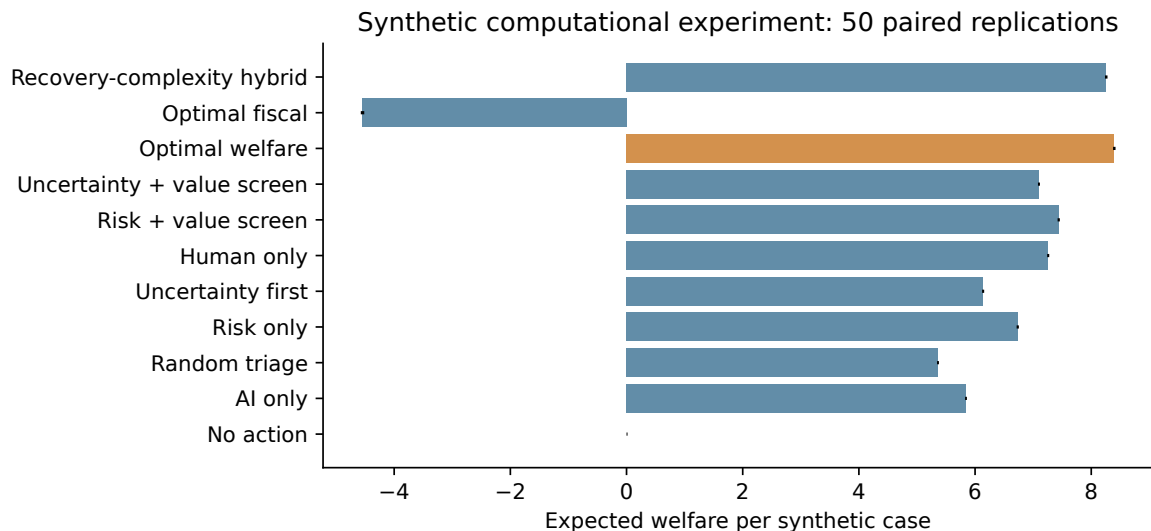


Figure 4: Expected welfare in a synthetic computational experiment with 50 independent replications of 100,000 cases each. Error bars are 1.96 times the Monte Carlo standard error of the replication mean and describe numerical variability, not empirical uncertainty about a tax authority.

not validate the DGP against observed returns. The optimizer’s weak superiority under correct primitives follows from feasibility; the experiment measures how far the particular benchmarks lie from the optimum. The computation should not be read as an out-of-sample claim about an algorithm trained with scarce inspection labels.

3.3 Capacity, AI quality, and error dependence

Figure 5 shows the capacity curve and next-slot value. Welfare rises as positive increments are admitted in descending order and eventually saturates. The curve’s diminishing increments match Proposition 2. Net fiscal value is shown as a secondary outcome under welfare-optimal assignment, rather than as the maximized objective. A large nominal pool of available inspectors can remain partly unused if no additional investigation covers its economic cost. This behavior reflects the upper-bound capacity constraint; it would not follow from a requirement to conduct a fixed number of audits.

The AI-quality experiment shows generally declining positive review demand alongside an increasing value of the retained review portfolio over much of the quality range. The small demand rebound at the final quality point precludes a claim of strict monotonicity. The additional-capacity value is nonmonotone: it first declines and then rises at high machine quality in the baseline parameterization. The recovery gap is larger for complex cases by construction, and better machine signals help identify discrepant cases on which that gap pays. This is a modeled selection mechanism, not a measured productivity response of real inspectors. Figure 6 presents review uplift, unconstrained positive demand, and the next-slot value separately, because a fixed 10% capacity can conceal changes in desired review demand.

Figure 7 distinguishes information and procedural margins under fixed marginal accuracies. In the equal-accuracy, equal-recovery setting, review uplift approaches zero as the reviewer becomes a clone. Under equal accuracy with procedural advantage, review remains valuable at the clone endpoint despite zero incremental information. Under heterogeneous unequal accuracy, the total effect of dependence need not be monotone, consistent with the analytical counterexample. Comparing these

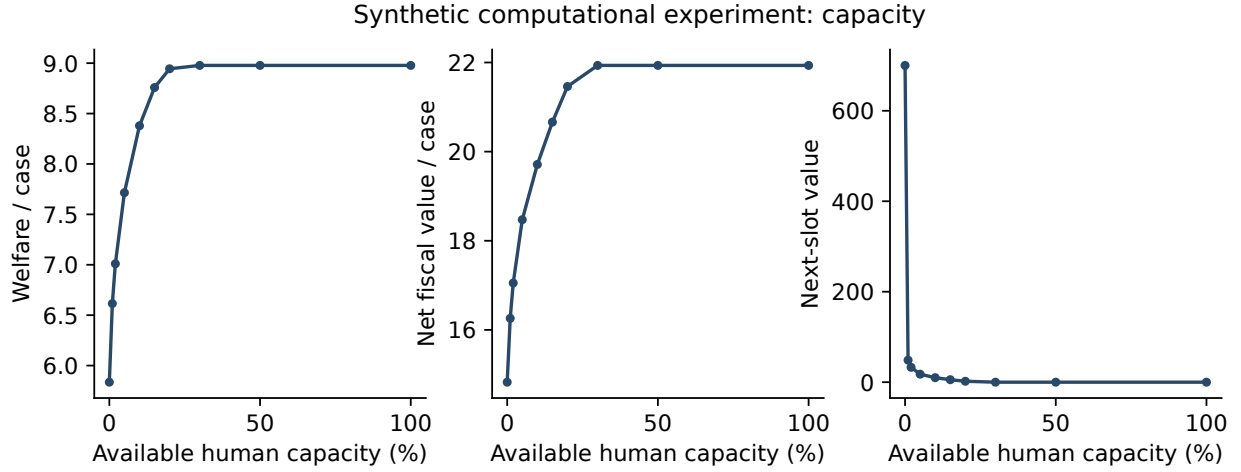


Figure 5: Human capacity and economic outcomes in a synthetic computational experiment. Capacity is the available fraction of 100,000 cases. The final panel reports the true-DGP value of the next unit review slot and is zero after positive gains are exhausted.

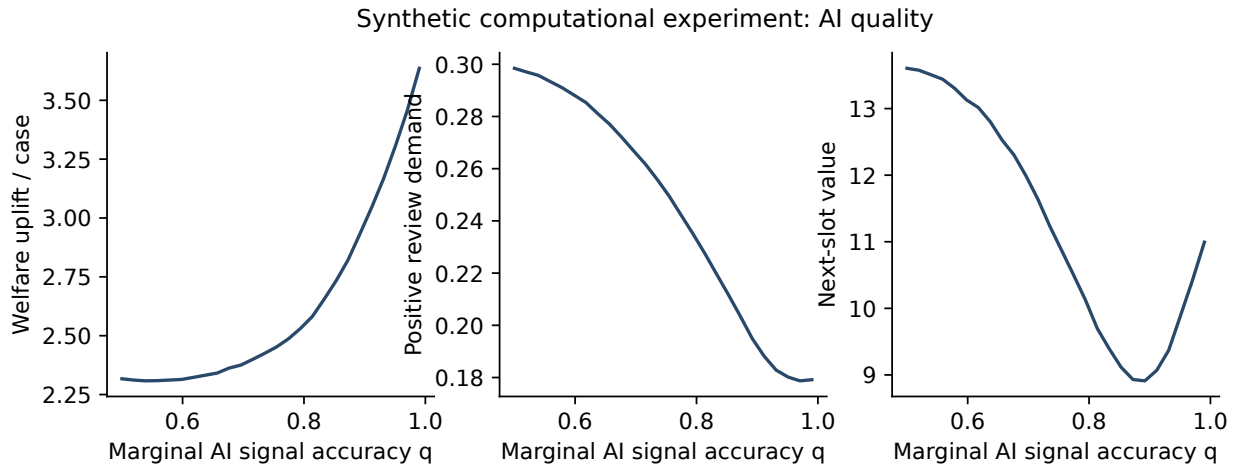


Figure 6: AI-quality sensitivity in a synthetic computational experiment, 26 machine accuracy values and fixed 10% available human capacity. The joint evidence channel is regenerated at each value, so the sweep is not a retained-information Blackwell experiment.

settings requires care: the equal-recovery ablation changes the no-human recovery technology and therefore the baseline welfare level. The plotted review uplift is relative to each setting’s own best nonhuman value.

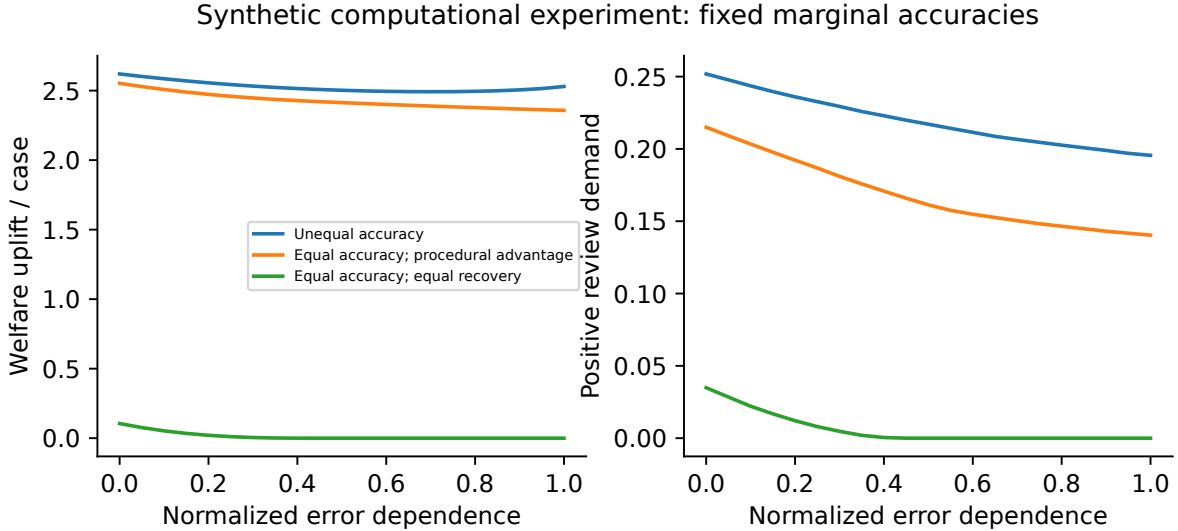


Figure 7: Fixed-marginal dependence sensitivity in a synthetic computational experiment. The horizontal parameter moves joint errors from independence to maximal feasible positive dependence. It is not Pearson correlation when marginal error rates differ. Each uplift uses that scenario’s own AI-only baseline.

The two-dimensional quality–capacity surface in Figure 8 reports shadow values rather than imposing named institutional regimes. It identifies parameter cells where human capacity is economically binding and cells where its marginal value is small. A numerical surface alone does not establish a discontinuous regime boundary or a universal threshold between AI substitution and complementarity. Its function is to display how the case-gain distribution shifts jointly with information and available review resources.

3.4 Robustness and strategic numerical extension

The named robustness environments alter risk distributions, recovery tails, cost levels, false-action burdens, and belief calibration. Table 2 reports absolute welfare for the main ranking comparisons. Correctly specified optimization weakly dominates feasible comparison policies by construction, including under the Latin hypercube’s 200 lawful parameter draws. That dominance is a verification property rather than a robustness discovery. The meaningful variation is in the gap size, the bindingness of capacity, and the decomposition of human gains. In the miscalibration environments the optimizer instead uses misspecified beliefs; true-DGP evaluation records the resulting losses and does not enforce dominance.

The exploratory input-sensitivity analysis makes the known-primitives limitation concrete. On the unchanged baseline seed, an expected recovery-increment ranking with a common oracle positive-value screen achieves welfare 8.263, close to the true optimum 8.379. An optimized rule that incorrectly assumes maximal dependence achieves 8.060, while assuming constant human quality 0.99 gives 7.769. Assuming constant recovery fractions 0.85 and 0.95 yields 7.296. These misspecified rules can lose to the simpler hybrid despite having correctly calibrated state probabilities. They are single-seed sensitivity constructions, not replication means or empirical estimates. Their purpose is

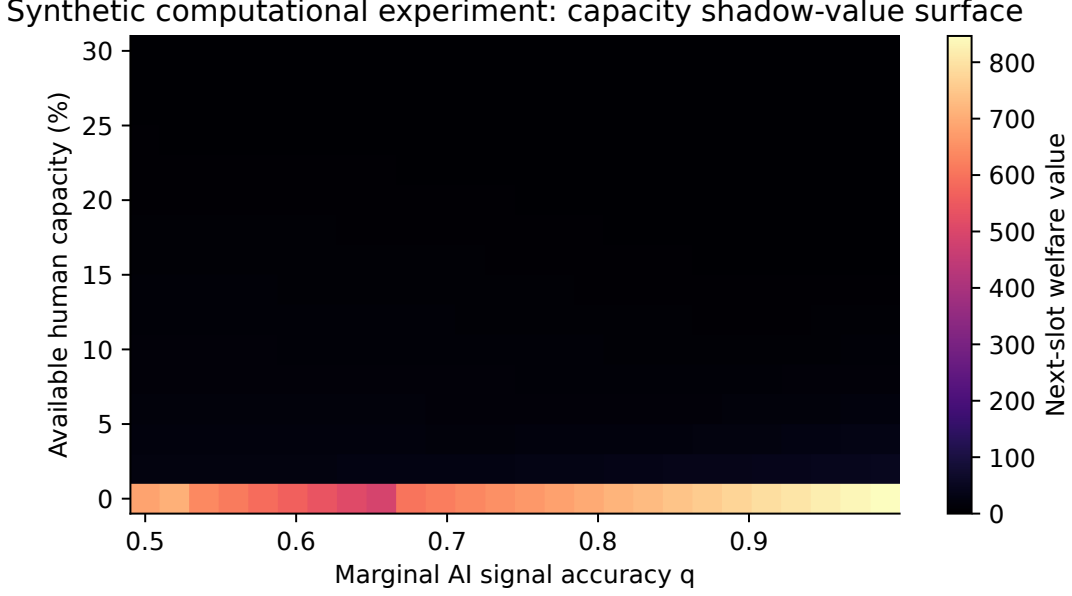


Figure 8: Capacity shadow-value surface in a synthetic computational experiment: 26 AI-quality values crossed with 16 capacity values, totaling 416 cells. Brighter cells have larger welfare value for the next review slot.

to identify conditional channel and recovery estimation as substantive requirements of value-based assignment.

Equal recovery removes the procedural advantage; zero net fiscal premium removes the static welfare rationale for costly processing. These ablations establish which mechanisms drive the baseline results. Changing recovery tails alters the concentration of large gains and hence the value of the earliest slots. The model cannot infer which tail or premium best describes an actual tax population. The sensitivity checks deliberately preserve that uncertainty rather than selecting the configuration with the largest reported policy gap.

Figure 9 reports the separate 1,000-type deterrence experiment. The continuous optimum improves the specified welfare relative to the frozen-response ranking and changes concealment intensity. The gap combines endogenous-response ranking and the ability to spread review intensity across types. Proposition 8’s two-type example provides the cleaner evidence that ranking alone can reverse. The computation is a tractable commitment benchmark, not an estimated deterrence multiplier or a simulation of interacting real firms.

4 Discussion

4.1 Economic interpretation and the scope of the contribution

The model’s central comparison is between human investigation and the best alternative that would actually be used for the case. A raw risk score omits that alternative. When inexpensive reconciliation already resolves a likely discrepancy, high risk need not imply high incremental information value. When recovery requires investigation even after diagnosis is straightforward, the procedural component can instead favor high-risk cases. The analysis therefore supports neither universal high-risk escalation nor universal uncertainty-first escalation. It specifies the margins

Table 2: Robustness in synthetic computational experiments, 100,000 cases per named environment. Values are expected welfare per case. Belief-calibration variants choose with a transformed posterior and are evaluated under the true DGP.

Environment	Optimal welfare	Risk only	Uncertainty first
Baseline	8.379	6.711	6.138
Equal recovery	10.989	9.742	10.196
Expensive automation	6.277	4.306	3.659
Expensive human	6.443	3.831	3.258
Gamma recovery	4.708	3.600	3.175
High dispersion	17.520	13.838	12.847
High false burden	6.515	4.411	3.507
High fiscal premium	21.131	17.994	16.796
High prevalence	28.785	24.494	23.460
High recovery scale	29.448	25.509	23.937
Independent errors	8.455	6.724	6.285
Low dispersion	3.870	3.305	2.978
Low prevalence	1.894	1.261	1.163
Low recovery scale	0.852	-0.046	-0.324
Maximally dependent errors	8.365	6.730	5.861
Overconfident score	8.145	6.513	5.979
Pareto recovery	9.710	8.463	8.002
Strong AI	13.152	10.957	10.575
Strong human	9.243	7.083	6.877
Underconfident score	8.029	6.417	5.844
Uniform prior	19.417	16.298	15.297
Weak AI	5.542	3.907	3.652
Weak human	8.193	6.590	5.650
Zero fiscal premium	0.000	-0.800	-0.800

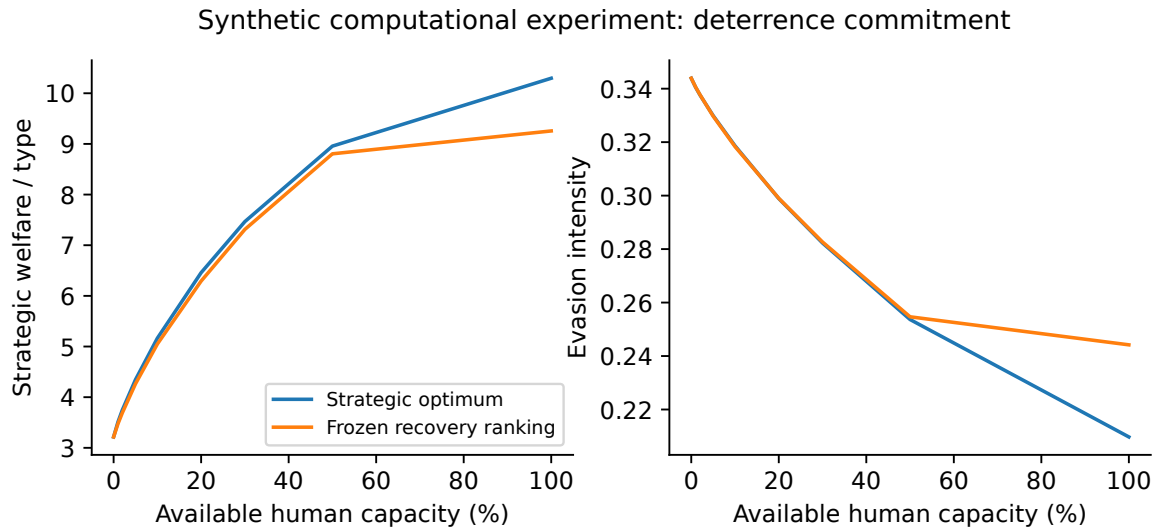


Figure 9: Separate deterrence commitment model in a synthetic computational experiment with 1,000 equally weighted types. The optimized allocation is continuous; the frozen-recovery comparator selects types on a binary basis. The numerical difference includes this intensive-margin distinction.

that make either pattern possible and shows that their coexistence can produce more complicated allocation geometry.

This distinction also sharpens the relationship to existing research. Conditional-loss ranking under limited expert capacity is already part of algorithmic triage (Okati et al., 2021), and cost-sensitive allocation with workload constraints is already studied directly (Alves et al., 2024). Expected human information value appears in Wilder et al. (2020). Information-sensitive auditing and voluntary reporting are central to Macho-Stadler and Pérez-Castrillo (2002); Kuchumova (2017) studies the allocation of enforcement resources between information and audits, including circumstances in which less precise information calls for more auditing. Economic and distributional objectives are already incorporated in tax-audit algorithms (Black et al., 2022). The present contribution is a joint, explicitly limited treatment of additional conditional evidence, procedural recovery, tax-transfer accounting, and incentive-adjusted capacity allocation. The article does not claim priority for any of those ingredients separately.

The decomposition explains a potentially misleading interpretation of correlated errors. If a human repeats the information already contained in a machine output, reviewing the case cannot improve the decision through evidence alone. But redundancy of judgments does not imply redundancy of investigative procedures. Inspectors may obtain documents, trace transactions, or reconcile an amount that an automated route can only partially recover. The model represents such differences through recovery fractions, not through an unexplained assumption of uniformly superior human accuracy. Whether those fractions differ in an institution is an empirical question. A procurement evaluation that records only classifier agreement cannot resolve it.

The unequal-accuracy counterexample gives a related lesson about information measurement. A reviewer can be less accurate overall while providing a signal whose disagreement with the observed machine output is informative. Conversely, a highly accurate reviewer can add little when both judgments rely on the same evidence. The coupling experiment keeps accuracies fixed and changes dependence precisely to separate these margins. The nonmonotone example additionally shows why a single scalar error-correlation statistic is not a general ordering of evidence systems. Conditional likelihoods and the authority’s ability to interpret disagreement determine decision value.

4.2 Tax administration, staffing, and technology choices

Within the model, inspection staffing has a well-defined marginal object: the value of the next feasible slot. It measures a case’s gain after accounting for the route that it displaces. It is therefore more informative for a capacity decision than average recovery per investigated case. Average recovery can remain high when the next case is scarcely worth investigating, and can include tax that the automated route would also recover. Similarly, a positive fiscal return need not imply positive welfare once private burdens and the financing of administrative inputs are included. These distinctions connect to the welfare analysis of audit spending (Keen and Slemrod, 2017; Boning et al., 2025) without substituting synthetic shadow values for empirical ones.

The staffing result requires caution in moving from slots to employees. Real investigators differ in hours, skill, legal authority, and specialization; returns have different expected durations. One additional employee does not necessarily create a constant number of exchangeable case slots. With heterogeneous resource requirements, the allocation becomes a knapsack or scheduling problem, and sorting by raw value can fail. An operational extension would need time distributions, case queues, and perhaps multiple skill capacities. The one-slot assumption is useful for isolating opportunity cost, but it is not a description of a personnel budget. The shadow value reported here is explicitly a slot value.

Comparing investment in AI with expansion of human capacity also requires prices. A numerical

improvement in an accuracy parameter is not a monetary investment. It may require retraining, data acquisition, integration, legal review, and ongoing monitoring. The analysis identifies the marginal benefit side of the decision: changes in optimized value from an information improvement and changes from an extra slot. A cost-effective investment comparison would divide those benefits by the corresponding marginal resource requirements. No such requirements are estimated here. Consequently, the article cannot rank an AI purchase against hiring an inspector, even when both numerical derivatives are available.

The simulations nonetheless identify why that comparison may be nonlinear. Improved screening can release cases from the review queue, while concentrating available slots on cases with substantial procedural advantage. Positive review demand and review productivity can therefore move in opposite directions. Under a fixed staffing cap, the observed number of human investigations may conceal this substitution. Measuring both unconstrained positive demand and the gain distribution avoids confusing a binding institutional limit with the technologically desired use of humans. The distinction is consistent with complementarity frameworks (Donahue et al., 2022; Vaccaro et al., 2024), but the parameters here concern an idealized investigation process rather than observed teamwork.

4.3 Identifying human value from real operational data

The optimizer assumes that the authority knows the state probabilities, conditional inspector channel, recovery technologies, and burdens. Estimating those quantities is substantially harder than sorting their implied values. Administrative audit outcomes are usually observed only for selected cases, and intensive inspection can itself change what is observed. A detected discrepancy is therefore not a random label of latent noncompliance. The selection problem is central to tax-targeting research (Battaglini et al., 2025) and to deferral systems that lack representative expert predictions (Mozannar and Sontag, 2020; Charusaie et al., 2022; Narasimhan et al., 2022). The synthetic generator removes these problems by design and should not be presented as solving them.

A useful empirical design would distinguish the assignment effect from the processing-technology effect. Randomly escalating some cases near an operational margin could reveal the human route’s incremental recovery and burden relative to the nonhuman route. To learn conditional evidence value, it would also be necessary to record the information available before escalation, what the inspector learned, and what decision would have followed without that new information. Recording only final revenue cannot separate evidence from procedure. Random exploration in a small part of the portfolio may be necessary to obtain information on cases normally released or automated, but the appropriate exploration rate is outside this model.

Burden measurement is equally consequential. A false-positive action in the computations is a verification request directed at a truly compliant synthetic case, not a wrongful criminal conviction. Its social cost depends on document preparation, disruption, uncertainty, legal assistance, and the eventual response process. Those costs may vary systematically with size, recordkeeping ability, or access to representation. The baseline treats them as constants, making the objective auditable but omitting distributional heterogeneity. An empirical welfare application should estimate these burdens and specify welfare weights. A risk-ranking advantage cannot by itself establish equitable treatment.

Calibration is another input rather than an automatic property of a machine score. The robustness exercise changes the belief used for routing while retaining the true DGP for evaluation. This makes the loss from belief misspecification visible, but does not model every dimension of score misuse. Human-aligned calibration concerns whether probabilities support the decisions a person actually makes (Benz and Rodriguez, 2023, 2025); accurate probabilities alone do not

guarantee appropriate interpretation. Studies of explanations and human–AI teaming similarly caution against equating model output quality with complementary decisions (Bansal et al., 2021). In the present benchmark, both the authority and the reviewer correctly use the specified likelihoods. Misinterpretation, automation bias, and organizational incentives could reduce the achievable gain.

Several measurements would be required before a quantitative application. An authority would need a common information snapshot before routing, route-specific staff time, records of requests sent to compliant cases, recoveries attributable to the intensive procedure rather than to ordinary reconciliation, and a way to observe how missing outcomes are selected. The human evidence channel should be evaluated conditional on the exact machine information available to the decision maker. Measuring inspector accuracy on cases without that conditioning can overstate the new evidence an investigation supplies. Where an inspector sees the score itself, the joint law may reflect reliance on the machine as well as access to independent evidence.

It would also be necessary to distinguish uncertain knowledge from genuine heterogeneity. An authority can be uncertain about an inspector’s ability because it has few validation observations; a case can be uncertain because its calibrated state probability is near an action boundary. These are different objects. The model varies the human evidence channel as a primitive and uses posterior ambiguity for case uncertainty. It does not treat the amount of available validation data as another observed feature. Research on limited expert samples and configurable expert benchmarks provides designs for studying that missing estimation layer (Charusaie et al., 2022; Alves et al., 2025). The formal allocation calculation can then serve as a benchmark for assessing the costs of using an estimated channel.

This perspective suggests evaluating simple economic rules before assuming that every modeled distinction must be estimated finely. Recovery-based ranking performs close to the oracle in the baseline, while channel and recovery misspecification can erase the oracle’s advantage. The appropriate comparison in a real application is therefore between complete systems, including the cost and uncertainty of learning their inputs. A more elaborate value model can be useful for explaining choices without necessarily producing a better feasible policy from a limited operational dataset. The synthetic findings provide examples of that tradeoff and a reproducible environment in which alternative estimators could be tested; they do not resolve it empirically.

4.4 Deterrence, commitment, and institutional implementation

The commitment extension supplies a fiscal mechanism absent from static triage: the review allocation changes concealment and voluntary payments. It thereby distinguishes direct recovery from the full tax-receipt response. This accounting is consistent with the broad lesson that future reporting effects can be economically important (Advani et al., 2023). The extension does not borrow an empirical deterrence magnitude from that literature or treat a quadratic concealment response as established behavior. Its two-type example proves that a ranking based on existing recovery can be wrong when response elasticities and saved real resources differ across types.

Commitment itself is a substantive assumption. The model presumes that types understand the announced review probabilities and that the authority implements them. Actual case selection may be overridden, delayed, or only partially executed. Bachas et al. (2025) studies the interaction between algorithmic selection and officials’ implementation, while Barwahwala et al. (2024) emphasizes why field evaluation must consider operational performance. Their findings motivate measuring implementation rather than inserting an exogenous “human oversight” step and presuming success. The computed optimum gives an implementable mathematical allocation under its assumptions, not evidence that an organization will follow it.

The extension also makes clear what remains unsolved. Endogenous concealment does not

feed back into the static machine posterior, inspector error law, or recovery fraction. A unified equilibrium would need a reporting or manipulation model, a retraining rule, and the government’s commitment or timing structure. It could generate exploration incentives, pooling, or changes in the information used to target cases. Those features could invalidate fixed-gain nesting and the simple capacity formula. We preserve the separation between the two models rather than suggesting that a one-period type response establishes a general dynamic enforcement equilibrium.

4.5 Limitations and a reproducible research agenda

The numerical exercise is informative about mechanisms but deliberately limited about magnitudes. Prior probabilities and recovery amounts are independent in the baseline; human skill falls with complexity; procedural advantage rises with complexity. Those are chosen functions. They are varied and ablated, but a broad parameter scan cannot establish that any configuration resembles a specific administration. Heavy recovery tails can concentrate benefits in a small number of synthetic cases, making early capacity increments appear unusually valuable. The saved manifest, fixed seeds, and separate expected and realized calculations support reproduction, not external validity.

The human signal is binary and state symmetric, and positive dependence is interpolated within a particular feasible coupling. Real evidence may be continuous, asymmetric, or correlated through unobserved case features. With different joint laws, the same marginal accuracies and an aggregate correlation can imply different conditional review values. The clone theorem is robust to a stated garbling relation, but the unrestricted dependence claim is deliberately rejected. An empirical application should estimate or bound the joint conditional information structure rather than treat correlation as an independently selectable universal parameter.

The model excludes appeals, legal constraints on automation, strategic inspectors, heterogeneous audit durations, and spillovers across related taxpayers. It does not determine the legal threshold for initiating an investigation or a penalty. Cross-domain work on reject options, medical second opinions, and expert-assisted classification offers alternative architectures (Raghu et al., 2019; De et al., 2021; Verma and Nalisnick, 2022; Mozannar et al., 2023; Cortes et al., 2026), but their evaluation data do not validate tax procedures. The algebra can accommodate different action sets once their values and feasibility are specified; changing the action set can change the optimum materially.

Finally, reproducibility is distinct from scientific author validation. The code records a transparent DGP and executes the optimization and numerical checks, while the proofs state their restrictions and exhibit counterexamples. Internal AI reviewer perspectives helped identify mistakes, including omitted voluntary tax receipts and unjustified correlation monotonicity. Such diagnostics do not replace an accountable human author’s review or external peer review. The article’s appropriate use is as an auditable model-based research contribution and a starting point for institution-specific measurement, with its theoretical restrictions and synthetic status retained in any later application.

5 Conclusion

Scarce inspection capacity should be allocated using the incremental value of human intervention relative to the best feasible nonhuman route. This principle follows from a standard allocation argument; its tax interpretation depends on how evidence, recoverability, transfers, burdens, and incentives enter the payoff. The model separates information from procedural recovery advantage, identifies conditions for intermediate-posterior review, and gives counterexamples to universal threshold and correlation intuitions. Its separate commitment extension shows how concealment responses and voluntary payments alter marginal review value.

The synthetic computational experiments evaluate these mechanisms through policy comparisons, repeated random seeds, dependence and capacity sweeps, and explicit ablations. They report economic outcomes and numerical uncertainty without administrative data or empirical calibration. Known-primitives optimization is an oracle benchmark; a useful operational system must estimate its inputs and respect institutional constraints. The results support a research question about where human intervention adds economic value, rather than a general claim that combining humans and AI improves tax enforcement in every setting.

Declarations

Data and code availability. No confidential, proprietary, or administrative taxpayer data were used. All computational results are generated from synthetic cases produced by the code in the accompanying computational supplement. The synthetic baseline dataset can be regenerated, and the supplement includes its generator, parameter configuration, machine-readable results, and reproduction instructions. No public repository or dataset DOI is claimed.

Ethics and consent. No human participants, identifiable personal data, or animal subjects were used. No institutional ethics approval or waiver is claimed. Consent to participate and consent for publication of participant information are not applicable.

Competing interests. The author declares no competing interests.

Funding. No project-specific funding has been declared for this manuscript.

Author contribution. Kun Huang is the sole named human author; scientific accountability rests with the human author. The detailed scope of AI assistance appears in Methods; AI is not an author.

References

- Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark. *Econometrica* 79(3):651–692, 2011. URL <https://doi.org/10.3982/ECTA9113>.
- Arun Advani, William Elming, and Jonathan Shaw. The Dynamic Effects of Tax Audits. *Review of Economics and Statistics* 105(3):545–561, 2023. URL <https://direct.mit.edu/rest/article/105/3/545/107397/The-Dynamic-Effects-of-Tax-Audits>.
- Michael G. Allingham and Agnar Sandmo. Income tax evasion: a theoretical analysis. *Journal of Public Economics* 1(3–4):323–338, 1972. URL [https://doi.org/10.1016/0047-2727\(72\)90010-2](https://doi.org/10.1016/0047-2727(72)90010-2).
- Jean V. Alves, Diogo Leitão, Sérgio Jesus, Marco O. P. Sampaio, Javier Liébana, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. Cost-Sensitive Learning to Defer to Multiple Experts with Workload Constraints. *Transactions on Machine Learning Research* (July 2024), 2024. URL <https://openreview.net/forum?id=TAvGZm2Rqb>.

- Jean V. Alves, Diogo Leitão, Sérgio Jesus, Marco O. P. Sampaio, Javier Liébana, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. A benchmarking framework and dataset for learning to defer in human-AI decision-making. *Scientific Data* 12:506, 2025. URL <https://www.nature.com/articles/s41597-025-04664-y>.
- Pierre Bachas, Anne Brockmeyer, Alipio Ferreira, and Bassirou Sarr. Algorithms and Bureaucrats: Evidence from Tax Audit Selection in Senegal. *World Bank Policy Research Working Paper* 11205, 2025. URL <https://documents1.worldbank.org/curated/en/099318309052518864/pdf/IDU-c1c5baa2-34d1-4425-b1ae-5cb02d5b9795.pdf>.
- Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *CHI Conference on Human Factors in Computing Systems (CHI 2021)*, 2021. URL <https://aiweb.cs.washington.edu/ai/pubs/bansal-chi21.pdf>.
- Taha Barwahwala, Aprajit Mahajan, Shekhar Mittal, and Ofir Reich. Is Model Accuracy Enough? A Field Evaluation of a Machine Learning Model to Catch Bogus Firms. *ACM Journal on Computing and Sustainable Societies* 2(3), Article 30:1–43, 2024. URL <https://www.nber.org/papers/w32705>.
- Marco Battaglini, Luigi Guiso, Chiara Lacava, Douglas L. Miller, and Eleonora Patacchini. Refining public policies with machine learning: The case of tax auditing. *Journal of Econometrics* 249:105847, 2025. URL <https://www.sciencedirect.com/science/article/pii/S0304407624001921>.
- Nina L. Corvelo Benz and Manuel Gomez Rodriguez. Human-Aligned Calibration for AI-Assisted Decision Making. *Advances in Neural Information Processing Systems* 36:14609–14636, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2f1d1196426ba84f47d115cac3dcb9d8-Abstract.html.
- Nina L. Corvelo Benz and Manuel Gomez Rodriguez. Human-alignment influences the utility of AI-assisted decision making. *Scientific Reports* 15:29154, 2025. URL <https://www.nature.com/articles/s41598-025-12205-1>.
- Emily Black, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho. Algorithmic Fairness and Vertical Equity: Income Fairness with IRS Tax Audit Models. *Proceedings of FAccT 2022*:1479–1503, 2022. URL <https://arxiv.org/abs/2206.09875>.
- William C Boning, Nathaniel Hendren, Ben Sprung-Keyser, and Ellen Stuart. A Welfare Analysis of Tax Audits Across the Income Distribution. *Quarterly Journal of Economics* 140(1):63–112, 2025. URL <https://academic.oup.com/qje/article/140/1/63/7888907>.
- Kim C. Border and Joel Sobel. Samurai Accountant: A Theory of Auditing and Plunder. *Review of Economic Studies* 54(4):525–540, 1987. URL <https://academic.oup.com/restud/article-abstract/54/4/525/1587783>.
- Paul Carrillo, Dina Pomeranz, and Monica Singhal. Dodging the Taxman: Firm Misreporting and Limits to Tax Enforcement. *American Economic Journal: Applied Economics* 9(2):144–164, 2017. URL <https://www.aeaweb.org/articles?id=10.1257/app.20140495>.
- Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample Efficient Learning of Predictors that Complement Humans. *ICML; Proceedings of Machine Learning*

- Research 162:2972-3005, 2022. URL <https://proceedings.mlr.press/v162/charusaie22a.html>.
- Corinna Cortes, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Optimized Deferral for Imbalanced Settings. ICML; Proceedings of Machine Learning Research 306:21467-21492, 2026. URL <https://proceedings.mlr.press/v306/cortes26b.html>.
- Sanmay Das, Fang-Yi Yu, and Yuang Zhang. Optimally Auditing Adversarial Agents. Proceedings of the AAAI Conference on Artificial Intelligence 40(20):16787–16794, 2026. URL <https://ojs.aaai.org/index.php/AAAI/article/view/38722>.
- Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez Rodriguez. Classification Under Human Assistance. Proceedings of the AAAI Conference on Artificial Intelligence 35(7):5905-5913, 2021. URL <https://ojs.aaai.org/index.php/AAAI/article/view/16738>.
- Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. ACM Conference on Fairness, Accountability, and Transparency (FAccT):1639-1656, 2022. URL https://facctconference.org/static/pdfs_2022/facct22-3533221.pdf.
- Michael Keen and Joel Slemrod. Optimal tax administration. Journal of Public Economics 152:133–142, 2017. URL <https://www.nber.org/papers/w22408>.
- Yulia (Paramonova) Kuchumova. The Optimal Deterrence of Tax Evasion: The Trade-off Between Information Reporting and Audits. Journal of Public Economics 145:56–66, 2017. URL <https://www.sciencedirect.com/science/article/pii/S0047272716301633>.
- Inés Macho-Stadler and J. David Pérez-Castrillo. Auditing with Signals. Economica 69(273):1–20, 2002. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0335.00268>.
- David Madras, Toniann Pitassi, and Richard Zemel. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. Advances in Neural Information Processing Systems 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.
- Ivan P. Malashin, Igor S. Masich, Vadim S. Tynchenko, Andrei P. Gantimurov, Vladimir A. Nelyub, and Aleksei S. Borodulin. Confidence-Aware Rejection Class Minimization in Indirect Tax Audit Classification. IEEE Access 12:192554–192571, 2024. URL <https://ieeexplore.ieee.org/document/10804786>.
- Dilip Mookherjee and Ivan Png. Optimal Auditing, Insurance, and Redistribution. Quarterly Journal of Economics 104(2):399–415, 1989. URL <https://academic.oup.com/qje/article-abstract/104/2/399/1854660>.
- Hussein Mozannar and David Sontag. Consistent Estimators for Learning to Defer to an Expert. ICML; Proceedings of Machine Learning Research 119:7076-7087, 2020. URL <https://proceedings.mlr.press/v119/mozannar20b.html>.
- Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who Should Predict? Exact Algorithms For Learning to Defer to Humans. AISTATS; Proceedings of Machine Learning Research 206:10520-10545, 2023. URL <https://proceedings.mlr.press/v206/mozannar23a.html>.

- Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. *Advances in Neural Information Processing Systems* 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/bc8f76d9caadd48f77025b1c889d2e2d-Abstract.html.
- Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable Learning Under Triage. *Advances in Neural Information Processing Systems* 34:9140-9151, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/4c4c937b67cc8d785cea1e42ccea185c-Abstract.html.
- Dina Pomeranz. No Taxation without Information: Deterrence and Self-Enforcement in the Value Added Tax. *American Economic Review* 105(8):2539–2569, 2015. URL <https://www.aeaweb.org/articles?id=10.1257/aer.20130393>.
- Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct Uncertainty Prediction for Medical Second Opinions. *ICML; Proceedings of Machine Learning Research* 97:5281-5290, 2019. URL <https://proceedings.mlr.press/v97/raghu19a.html>.
- Jennifer F. Reinganum and Louis L. Wilde. Income tax compliance in a principal-agent framework. *Journal of Public Economics* 26(1):1–18, 1985. URL <https://www.sciencedirect.com/science/article/pii/0047272785900350>.
- Joel Slemrod. Tax Compliance and Enforcement. *Journal of Economic Literature* 57(4):904-954, 2019. URL <https://www.aeaweb.org/articles?id=10.1257/jel.20181437>.
- Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8:2293-2303, 2024. URL <https://www.nature.com/articles/s41562-024-02024-1>.
- Rajeev Verma and Eric Nalisnick. Calibrated Learning to Defer with One-vs-All Classifiers. *ICML; Proceedings of Machine Learning Research* 162:22184-22202, 2022. URL <https://proceedings.mlr.press/v162/verma22c.html>.
- Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to Complement Humans. *IJCAI* 2020:1526-1533, 2020. URL <https://www.ijcai.org/proceedings/2020/212>.

A Mathematical appendix

A.1 Allocation, existence, and capacity value

For any selected set S , optimal nonhuman decisions contribute $\sum_i B_{Ai}$. Review adds $\sum_{i \in S} v_i$ and consumes $|S|$ slots. Removing a negative increment strictly improves the objective. If an unselected case has a larger gain than a selected case, their exchange preserves capacity and strictly improves value. Repeated exchanges yield the largest positive gains, proving Theorem 1. Equal gains permit multiple deterministic optima. Neither perfect information nor universally superior human performance is needed.

For the fractional relaxation, the dual objective is

$$\min_{\lambda \geq 0} \left\{ \sum_i B_{Ai} + \lambda H + \sum_i (v_i - \lambda)_+ \right\}.$$

At a binding integer cutoff without ties, any λ between the next unselected positive gain and the smallest selected gain supports the allocation. At slack capacity the multiplier is zero. The reported next-slot value corresponds to the lower endpoint of this interval. The continuous atomless problem has the analogous dual with an integral. Integrable gains ensure finite value; choosing all gains above a quantile threshold and randomizing on a positive-mass tie implements any binding capacity. The primal and dual values agree by the explicit threshold construction.

The value with H integer slots is the baseline plus the sum of the first H nonnegative order statistics. Its forward difference is the next order statistic, proving Proposition 2. A fixed tie convention gives a nested sequence. Fractional interpolation is concave because its slopes are ordered. A deterministic value function evaluated at the floor of a real-valued capacity is a step function; the concavity claim concerns integer differences or the fractional relaxation. If case duration differs, neither the exchange proof nor raw-gain sorting suffices. For example, a value-10 case requiring two slots loses to two value-6 cases requiring one each, despite having the largest individual gain.

A.2 The decomposition and homogeneous review interval

By posterior consistency and iterated expectations, the gross reviewed payoff is $\mathbb{E}B_H(p_Z)$. Adding and subtracting $B_H(p)$ gives equation (4). Convexity of the maximum of affine payoffs implies $\mathbb{E}B_H(p_Z) \geq B_H(p)$; equivalently, the authority can ignore the signal. Because $g_H \geq g_A$ in the maintained technology ordering, $B_H(p) \geq B_A(p)$. These nonnegative components do not imply positive net review value once k_H is charged.

For identical technologies, write the terminal processing value as $pG - (1 - p)L$. The positive and negative human signals switch that decision at

$$p_- = \frac{\beta L}{\alpha G + \beta L}, \quad t = \frac{L}{G + L}, \quad p_+ = \frac{(1 - \beta)L}{(1 - \alpha)G + (1 - \beta)L}.$$

When $G, L > 0$ and $\alpha > \beta$, one has $p_- < t < p_+$, with limiting weak inequalities at boundary channels. Below p_- both signals lead to release, and above p_+ both lead to processing. Gross information value is therefore

$$I(p) = \begin{cases} 0, & p \leq p_-, \\ \alpha G p - \beta L(1 - p), & p_- < p \leq t, \\ (1 - \beta)L(1 - p) - (1 - \alpha)Gp, & t < p < p_+, \\ 0, & p \geq p_+. \end{cases}$$

The middle expressions have positive and negative slopes, respectively, and agree at t , where the value is $(\alpha - \beta)GL/(G + L)$. This proves Proposition 3. If the peak exceeds $k_H + \lambda$, the strict review interval has endpoints

$$\ell = \frac{\beta L + k_H + \lambda}{\alpha G + \beta L}, \quad u = \frac{(1 - \beta)L - k_H - \lambda}{(1 - \alpha)G + (1 - \beta)L}.$$

Below ℓ the nonhuman optimum releases; above u it processes. At equalities either optimal branch can be used consistently with capacity. If $G \leq 0$ and $L \geq 0$, processing is never valuable and this nondegenerate interval characterization does not apply.

For the disconnected example in Results, human gross value is piecewise linear with breakpoints $1/19$ and $9/11$, while the direct-processing baseline changes at $1/2$. On the first active human segment, $v = 1.9p - 0.93$ for $p < 1/2$ and $v = 0.07 - 0.1p$ for $1/2 < p < 9/11$. Above $9/11$, $v = p - 0.83$. Setting each expression greater than zero yields $(0.93/1.9, 0.7) \cup (0.83, 1]$. The exact lower endpoint is $93/190$, and the reported decimal is rounded. This verifies that adding a recovery advantage can produce $N-H-A-H$ even with homogeneous costs and signal probabilities.

A.3 Refinement, costs, and selected sets

If a new human signal can reproduce the old one through state-independent garbling conditional on X , the authority can apply that garbling and follow its old terminal decision. The new signal therefore cannot lower optimal gross value, proving the refinement part of Proposition 4. This is a statement about a specified information order; an increase in an aggregate accuracy statistic alone need not imply it. A common added review cost d changes each gain to $v_i - d$. Pairwise differences are unchanged. With the same tie ordering, selected cases either remain or exit as positive demand contracts. Heterogeneous improvements can instead reorder cases.

For example, two cases with $p = 1/2$, $G = L = 1$, and review cost 0.1 have symmetric human qualities 0.70 and 0.65 . Net gains are 0.10 and 0.05 . With one slot, the first is reviewed. Improving the second quality to 0.80 gives gain 0.20 and displaces the first, although no individual gain fell. Binding capacity prevents a strict increase in review count, and selected sets are not nested under this heterogeneous refinement.

A.4 Error dependence and the clone restriction

The joint error table in Methods follows from the two marginal errors and their joint intersection c . Conditioning on B gives the stated α, β . The four-cell probabilities must be imposed identically conditional on each state; overall marginal accuracies would not otherwise identify the conditional channel. When $q = h$, the conditional channel at dependence ρ is a mixture: with probability $1 - \rho$ use independent human evidence, and with probability ρ copy the already observed B .

For $\rho_2 > \rho_1$, force the former human output to equal B with probability $(\rho_2 - \rho_1)/(1 - \rho_1)$. This garbling is independent of the state conditional on B and produces precisely the channel at ρ_2 . The previous refinement argument proves monotone information loss. At $\rho = 1$ there is no posterior change, hence $I = 0$ and $v = P - k_H$. The procedure can still add value, proving Proposition 5 in its qualified form.

For the unequal-quality counterexample, equation (5) yields $\alpha = 0.7 + (7/90)\rho$ and $\beta = 0.3 + 0.7\rho$. At $p = 1/2$ and $G = L = 1$, the two branch contributions are opposite, so the sum of their positive parts is $|\alpha - \beta|/2$. Substitution yields the expression in Results. At dependence exceeding $9/14$, $\alpha < \beta$, and acting on the conventionally negative signal is optimal. Both marginal accuracies are unchanged. This lawful counterexample refutes the unrestricted correlation comparative static rather than contradicting the equal-quality theorem.

A.5 AI information, fiscal disagreement, and limiting cases

For Proposition 6, take any optimal policy using the former AI information. Under richer information, ignore the added evidence and use the retained former information, select the same review cases, and make the same terminal decisions. Capacity, physical human evidence, and payoffs are unchanged. The old expected value is feasible, so the new optimum weakly exceeds it. This *ex ante* argument relies on literal retention and the unchanged joint signal environment; it is not a pointwise posterior comparison. Marginal Blackwell dominance of the machine signal alone, with only the state–human marginal law fixed, would not suffice because dependence with human evidence could change.

To show substitution, take prior $1/2$, $G = L = 1$, perfect human evidence, and review cost 0.1. An uninformative machine gives review gain 0.4. A perfect machine gives gain -0.1 , and positive review demand disappears. To show complementarity, take prior 0.95, human symmetric quality 0.8, and the same payoffs and review cost. An uninformative machine gives zero gross information value, because even unfavorable human evidence cannot change the action. A machine with symmetric quality 0.9 generates unfavorable outputs with probability 0.14 and posterior $19/28$. Conditional human gross value there is $17/140$, giving net gain $3/140$. At a capacity below 0.14, the slot value rises from zero to $3/140$. Thus total information improvement does not determine the sign of review demand or shadow value.

For Proposition 7, if $\mu = 0$, then $g_A = g_H = -b_1 \leq 0$, $f = b_0 \geq 0$, and all processing values are nonpositive. The human branch optimally closes after every signal and pays k_H . Release dominates costly intervention. To construct disagreement between objectives, take two cases with perfect human evidence, actual $c_A = 0.1$, $c_H = 0.5$, and $b_0 = b_H = 0$. Case 1 has $(p, R, r_A, r_H, b_1) = (0.9, 100, 0.5, 1, 10)$; case 2 has $(0.5, 40, 0.1, 1, 0)$. Fiscal gains are 44.51 and 17.55, selecting case 1. At $\mu = 0.1$, financed welfare costs are 0.11 and 0.55. Welfare gains are -0.55 and 1.305, selecting case 2. State probabilities and human accuracy are unchanged.

A.6 Deterrence response and optimal commitment

The private objective is strictly concave in e when $\kappa > 0$, and its interior first-order condition gives $e = (T - FD)/\kappa$. The imposed restrictions $0 < T - F(d_A + \delta)$ and $T < \kappa$ ensure $0 < e < 1$ for every $t \in [0, 1]$. Also require $0 \leq d_A \leq d_A + \delta \leq 1$, $F \geq T > 0$, and $\delta \geq 0$. The total private sanction includes recovered tax, while additional penalty receipts have no separate premium in the specified welfare objective.

Because $D' = \delta$ and $e' = -F\delta/\kappa$, differentiating equation (6) gives

$$W' = Fe\delta + \mu T\{-e' + e'D + e\delta\} - \Delta c = \delta[Fe + \mu T\{e + F(1 - D)/\kappa\}] - \Delta c.$$

Differentiating again yields $W'' = -\delta^2(F^2 + 2\mu TF)/\kappa$. Concavity follows for $\mu \geq 0$. The Lagrangian separates by type and gives $M_j(t_j) = \lambda$ for interior choices, $M_j(0) \leq \lambda$ for unused types, and $M_j(1) \geq \lambda$ for fully reviewed types. With $a_j = M_j(0)$ and $s_j = \delta_j^2(F_j^2 + 2\mu T_j F_j)/\kappa_j$, the solution for $s_j > 0$ is

$$t_j(\lambda) = \min\{1, \max\{0, (a_j - \lambda)/s_j\}\}.$$

If the resulting capacity at $\lambda = 0$ is feasible, it is the optimum. Otherwise a scalar root enforces $\sum_j m_j t_j = K$. Zero detection increments and nonnegative review costs optimally receive no review. This establishes Proposition 8 without relying on simulation.

For the two-type example, the marginal schedules are $M_A(t) = 0.125 - 0.0075t$ and $M_B(t) = 0.260 - 0.060t$. With masses one-half each and $K = 0.25$, setting $t_B = 0.5$ exhausts capacity and gives multiplier 0.23. Since $M_A(0) < 0.23$, type A remains unreviewed. The frozen-response recovery

ranking is reversed. Both concealment choices remain interior throughout the feasible interval, so no omitted clipping or discontinuity drives the example.

B Computational appendix

B.1 Reproduction and parameter records

Install the dependencies listed in the computational supplement. From the project root, execute:

```
python scripts/reproduce.py
```

The script saves the baseline synthetic dataset, a full parameter and environment manifest, CSV results, automatically generated LaTeX tables, and PDF/PNG figures. It does not require network access to regenerate computational results. Literature retrieval is separate from reproduction. Exact floating-point results may vary slightly across package and platform versions; the saved manifest records the environment used for the reported run.

The baseline seed is 20261003. Fifty production replication seeds are 20262003–20262052. The Latin hypercube has its own fixed seed and uses 10,000 cases for each of 200 parameter draws. The main quality grid contains 416 cells, and the dependence experiment contains 63 scenario points. The code logs expected values under the true DGP and separate realized values. Probability arrays used for routing never use the drawn true state or the unobserved human signal. The state and signal realizations enter only the realized-outcome diagnostic after routes have been chosen.

B.2 Accounting and Monte Carlo checks

For each case, let Q_1 and Q_0 be the probability of terminal processing conditional on discrepancy and compliance. The expected recovery is pQ_1r_aR , where the route selects its own recovery fraction. Administrative cost is $c_A[pQ_1 + (1-p)Q_0] + c_H1\{H\}$. Private burden is $b_1pQ_1 + b_0(1-p)Q_0 + b_H1\{H\}$. Fiscal net value subtracts administrative cost from recovery; welfare subtracts $(1 + \mu)$ times administrative cost and private burden from μ times recovery. These identities are tested against the independently constructed route values.

The testing suite verifies capacity, mutually exclusive actions, exhaustive six-case optimization, expected-value accounting, channel probabilities, fixed error marginals, clone redundancy, perfect machine and perfect human boundaries, uninformative machine evidence, zero and prohibitive costs, and the strategic dual solution. Additional proof stress checks use independently generated scalar configurations and symbolic differentiation. Their purpose is to detect implementation errors and inconsistent hypotheses; analytical proofs are supplied separately.

B.3 Interpretation of robustness

Named environments use the baseline seed so their comparisons benefit from common random numbers, but a single sensitivity path is not reported as a Monte Carlo confidence statement. The 50 independent replications apply to the baseline policy comparison. The Latin hypercube explores lawful configurations without estimating their probability in any real population. Correctly specified value optimization must weakly exceed its own feasible benchmarks; that is a constraint check, not an independent substantive discovery. Misspecified belief variants are evaluated under true conditional probabilities and do not inherit this guarantee.

The strategic generator draws $T \in [20, 60]$, $F/T \in [1.05, 1.50]$, $\kappa \in [70, 150]$, $d_A \in [0.03, 0.15]$, $\delta \in [0.15, 0.40]$, welfare-valued incremental review cost in $[0.5, 3]$, and base cost 0.5. These bounds

Table 3: Monte Carlo means and standard errors from a synthetic computational experiment with 50 independent 100,000-case replications. Every monetary entry is an arbitrary value unit per case.

Policy	Welfare	MCSE	Net fiscal	MCSE
No action	0.000	0.000	0.000	0.000
AI only	5.844	0.008	14.832	0.017
Random triage	5.362	0.008	15.441	0.017
Risk only	6.734	0.010	17.505	0.020
Uncertainty first	6.139	0.009	16.674	0.018
Human only	7.259	0.009	15.186	0.016
Risk + value screen	7.440	0.010	18.457	0.020
Uncertainty + value screen	7.100	0.009	17.477	0.017
Optimal welfare	8.397	0.011	19.759	0.022
Optimal fiscal	-4.545	0.015	25.363	0.021
Recovery-complexity hybrid	8.258	0.011	20.091	0.022

imply $F \geq T$, $d_A + \delta \leq 0.55$, and interior concealment over the entire review interval. All 1,000 types are synthetic. The extension uses continuous review intensities and is deliberately not passed through the static binary evidence generator.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [02supplementarymathematicalappendix.pdf](#)
- [03computationalsupplement.zip](#)