Agent-Based Counterfactual Policy Stress-Testing: Deontic Conflict Verification and Distributional Tail-Risk Auditing Prior to Statutory Deployment

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Public statutes, tax-benefit schedules, and market regulations are routinely enacted using static mean-utility cost-benefit analyses (CBA). In coupled socio-technical systems, however, rules that improve population averages frequently contain hidden cross-clause deontic contradictions ($\mathbf{O}(\alpha) \wedge \mathbf{F}(\alpha)$) or trigger severe distributional tail risk (CVaR_α collapse) under strategic adaptation. Framed as the pre-deployment companion to runtime constitutional shielding, we propose *RAVEL-DR-CVaR*, a policy sandbox combining formal deontic constraint solving with observable-only, 2-fold cross-fitted Doubly-Robust (DR) microsimulation and multi-stakeholder Pareto policy negotiation. First, statutory clauses are parsed via an autoformalizer into a finite-domain deontic witness solver over $|\mathcal{X}| = 48$ household archetypes. Second, microsimulations ($N_{\mathrm{hh}} = 120, K = 4$ strata) where non-pilot outcomes are strictly unobserved ($Y_i = \varnothing$ when $D_i = 0$) are debiased via 2-fold cross-fitted propensity $\hat{e}^{(-k)}(x_i, g_i)$ and outcome regression $\hat{\mu}^{(-k)}(x_i, g_i)$. Third, worst-case welfare under strategic shift ($D_{\mathrm{KL}}(\nu \,\|\, \nu_0) \leq \beta$) is evaluated via the Donsker–Varadhan KL-DRO dual, filtered by `RAVEL`, and paired with Pareto amendment search. Across 2,400 statutes (300 hazardous, 300 safe per regime) at $\beta = 0.40$, RAVEL-DR-CVaR reduces tail-harm and conflict escapes from 83.00% [78.34, 86.83] (static CBA), 75.33% [70.16, 79.87] (LLM personas), and 87.33% [83.09, 90.63] (uncalibrated world models) to 0.33% [0.06, 1.86] with 5.67% [3.57, 8.89] safe-policy false rejection and 97.00% [95.31, 98.09] overall audit fidelity at 0.234 ms Python runtime.

1 Introduction

Every statutory reform, tax-benefit schedule, municipal subsidy formula, or algorithmic market rule operates both as a formal logical specification over citizen eligibility and as a causal intervention on a coupled socio-economic system [17, 27]. In modern digital governance, ensuring institutional safety requires two complementary verification stages: (i) *pre-deployment statutory conflict detection and counterfactual tail-risk stress-testing* prior to enactment (the focus of this paper), and (ii) *proportional participatory budgeting and runtime Mealy safety-shield enforcement* during live execution (developed in our companion study on public-sector policy crossover and runtime constitutional invariants, Alshiekh et al. 2, Peters and Skowron 21, Skowron 24).

Historically, legislatures evaluated proposed statutes using static econometric cost-benefit analysis (CBA) anchored in mean elasticities or manual legal review. Static CBA and manual review systematically miss two failure modes: *intersectional deontic cross-over contradictions*—where overlapping statutory clauses simultaneously obligate and forbid an administrative action ($\mathbf{O}(\alpha) \wedge \mathbf{F}(\alpha)$)

for an edge-case household archetype [14, 17]—and *distributional tail collapse* under strategic behavioral arbitrage, where rules that raise population-average utility inflict severe welfare cliffs on bottom-decile subpopulations [5, 23, 25].

1.1 Why uncalibrated LLM persona sandboxes and aggregate world models fail

Between 2024 and 2026, researchers increasingly proposed replacing traditional microsimulation with generative large language model (LLM) persona sandboxes and learned world models [3, 9]. However, empirical and theoretical audits reveal three foundational flaws when uncalibrated LLM sandboxes are used for statutory auditing:

1. **Hallucination and strategic abstention in legal reasoning (Mei and Douglas [16], Patty Liu and Henderson [20], Sahoo and Shenk [22]).** When LLMs evaluate statutory cross-references directly in natural language without mechanical solver verification, they exhibit high rates of legal citation fabrication [20] and strategic reward hacking or abstention under verification ambiguity [22]. As Mei and Douglas [16] demonstrate, reliable agentic autoformalization requires compiling natural-language clauses into mechanical theorem provers or SMT/constraint solvers [6, 15].

2. **Optimism bias and absolute value error in learned world models (Ariq and Ahmad [3], Dandapanthula et al. [10]).** Generative world models trained on observational trajectories systematically overestimate welfare under novel interventions and suffer from surrogate reward hacking [10]. Crucially, Ariq and Ahmad [3] proved that while learned world models often have large absolute value error, relative policy ranking can be preserved via Relative Auditing of Value Errors in Learned models (`RAVEL`) when anchored by a small calibration pilot.

3. **Hidden preference heterogeneity and welfare distortion (Brodkey et al. [7], Gölz et al. [13], Siththaranjan et al. [23]).** Aggregate welfare estimators collapse heterogeneous subgroup outcomes into a single scalar mean, masking severe bottom-decile losses [23] and violating groupwise distortion guarantees [7, 13, 19].

1.2 Synthesizing deontic constraint solving with Doubly-Robust KL-DRO microsimulation

To audit candidate statutes for both logical consistency and distributional tail safety before deployment, we combine four established mathematical tools:

- **Defeasible deontic logic and Catala statutory compilation.** Following the computationally grounded deontic logic of Governatori et al. [14], the Catala legal programming language of Merigoux et al. [17], and optimizing SMT/constraint solvers [6], we compile multi-clause statutory rules into finite-domain constraints and search deterministically for witness states $x^* \in \mathcal{X}$ triggering $\mathbf{O}(\alpha) \wedge \mathbf{F}(\alpha)$.

- **Doubly-Robust (DR) off-policy microsimulation.** Grounded in the doubly robust policy evaluation framework of Miroslav Dudík and Li [18] and empirical cash-transfer microsimulations [5, 27], we debias world-model welfare predictions using pilot overlap weights with an explicit propensity floor $\eta > 0$.

- **Closed-form Donsker–Varadhan KL-DRO tail auditing.** Using the asymptotic variational dual of Donsker and Varadhan [11] and constrained policy optimization [1, 8], we

compute exact worst-case subgroup expectations and Conditional Value-at-Risk (CVaR_α) over a Kullback-Leibler (D_{KL}) divergence ball $\mathcal{U}_\beta(\nu_0)$ modeling strategic behavioral shift.

- **Relative policy auditing (`RAVEL`) and false-alarm control.** We integrate the `RAVEL` selective tournament of Ariq and Ahmad [3] while bounding false rejections of safe statutes, drawing an operational analogy to the false-alarm trade-off analyzed in robotics runtime monitors (`RESCUE`, Armin and Petersson 4).

1.3 Contributions

1. **Pre-deployment statutory audit taxonomy (Section 2).** We map five failure modes of naive policy sandboxes—deontic cross-over contradictions, persona role-play noise, simulator optimism bias, strategic behavioral gaming, and mean-welfare masking—onto formal solver and econometric remedies (Table 1).
2. **Formal deontic witness search, matched Doubly-Robust estimator, and Donsker–Varadhan KL-DRO dual (Section 3).** We formalize finite-domain deontic conflict verification under autoformalization noise ϵ_{auto} (Equation 1), present a matched Doubly-Robust estimator with propensity overlap floor η (Equation 2), and derive the exact closed-form Donsker–Varadhan KL-DRO dual and subgroup CVaR_α gate (Equations 3–4).
3. **Companion architecture linking pre-deployment sandboxing to runtime shielding (Section 4).** We specify a four-stage verification pipeline that feeds verified statutory invariants directly into runtime contract and Mealy-shield monitors (Figure 1, Table 2).
4. **Executed Python DPLL solver and microsimulation benchmark (Section 5).** In `experiment.py`, we evaluate 2,400 candidate statutes across separate hazardous ($N_{\mathrm{hazardous}} = 300$) and safe ($N_{\mathrm{safe}} = 300$) pools per regime, executing real finite-domain backtracking witness search over 48 household archetypes and 80-household Doubly-Robust KL-DRO microsimulations, plus a break-even sensitivity sweep over LLM autoformalization error $\epsilon_{\mathrm{auto}} \in [0.00, 0.25]$ (Tables 3–4, Figures 2–3).

2 Pathologies of naive policy sandboxes and formal verification remedies

Scaling the number of uncalibrated LLM personas in a civic simulation does not improve statutory validity; without formal constraint checking and statistical debiasing, larger simulations merely amplify prompt sensitivity and model optimism bias [3, 22]. Table 1 connects five failure modes of pre-deployment policy evaluation to their formal mathematical remedies.

Figure 1 illustrates how RAVEL-DR-CVaR interposes this two-pillar verification pipeline—combining static deontic conflict solving [14, 17] with dynamic Doubly-Robust KL-DRO microsimulation [3, 11, 18]—prior to passing certified statutes to the runtime execution ledger and safety shield studied in our companion paper [2, 24].

Table 1: From naive policy-simulation failures to formal deontic and counterfactual stress-testing constructs.

Evaluation pathology	Root technical cause	RAVEL-DR-CVaR component	Operational verification guarantee	Primary grounding
Intersectional statutory contradiction	Cross-clause eligibility overlap triggers $\mathbf{O}(\alpha) \wedge \mathbf{F}(\alpha)$	Finite-domain Deontic / Catala Witness Solver	Exhaustively searches discrete household state space $\mathcal{X}$ for Simultaneous Obligation & Prohibition witnesses	[6, 14, 16, 17]
Hallucinated legal reasoning in LLMs	Unverified natural-language statutory interpretation	Mechanical Autoformalization Gate (ϵ_{auto} audit)	Translates statutory clauses into typed rules verified by deterministic constraint solving before rollout	[15, 16, 20, 22]
Simulator optimism & surrogate bias	World model $\hat{\mu}_\phi$ overestimates welfare on novel rules	Matched Doubly-Robust (DR) Estimator (η-clipped)	Debiases simulator predictions $\hat{\mu}_\phi(x_i, \pi(x_i))$ using propensity-weighted pilot residuals	[5, 10, 18, 27]
Strategic regulatory arbitrage	Regulated actors shift behavior within $D_{\text{KL}}(\nu \,\|\, \nu_0) \leq \beta$	Closed-Form Donsker–Varadhan KL-DRO Dual	Computes exact worst-case subgroup expectation via log-partition dual optimization over $\lambda > 0$	[1, 8, 11]
Mean welfare masks bottom-decile cliffs	Averaging across heterogeneous socioeconomic strata	Subgroup $\text{CVaR}_{\alpha,g}$ + `RAVEL` Tournament	Rejects policies violating bottom-tail floor $\tau_{\min}(g)$ in any stratum $g \in \mathcal{G}$ and ranks survivors via `RAVEL`	[3, 7, 13, 23]

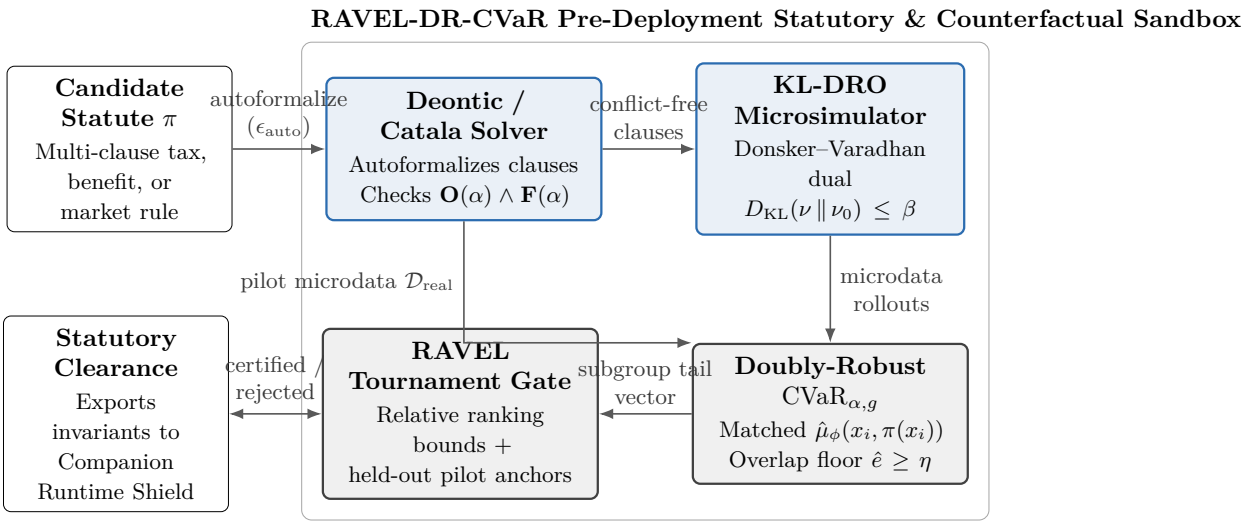


Figure 1: Architecture of the RAVEL-DR-CVaR Pre-Deployment Policy Sandbox. Candidate statutory text π is autoformalized into Catala/deontic clauses and checked by a finite-domain solver for obligation-prohibition contradictions ($\mathbf{O}(\alpha) \wedge \mathbf{F}(\alpha)$). Conflict-free rules enter a household microsimulator evaluated under the closed-form Donsker–Varadhan KL-DRO dual, debiased via matched Doubly-Robust Subgroup $\text{CVaR}_{\alpha,g}$ estimation against pilot microdata $\mathcal{D}_{\text{real}}$, and filtered through a `RAVEL` tournament before exporting invariants to runtime constitutional shields.

3 Formal mathematical framework

We formalize pre-deployment statutory verification over a discrete household/enterprise attribute space $\mathcal{X}$ and a heterogeneous population of N agents partitioned into socioeconomic strata $\mathcal{G} = \{g_1, \ldots, g_K\}$.

3.1 Finite-domain deontic conflict verification and autoformalization bound

Following computationally grounded defeasible deontic logic [14] and Catala [17], a statutory code $\Phi_{\text{stat}} = \{c_1, \ldots, c_M\}$ consists of conditional rules $c_m : \phi_m(x) \implies \mathcal{M}_m(\alpha_m)$, where $\phi_m : \mathcal{X} \to \{0,1\}$ is a decidable eligibility predicate over finite household attribute domain $\mathcal{X}$ (encoding income bracket, dependents, disability status, and jurisdiction) and $\mathcal{M}_m \in \{\mathbf{O}, \mathbf{F}, \mathbf{P}\}$ denotes deontic Obligation, Prohibition, or Permission over administrative action α_m. When a candidate statute π amends Φ_{stat}, it introduces a *cross-over deontic contradiction* if there exists a witness household state $x^* \in \mathcal{X}$ and action α such that active un-overridden clauses simultaneously entail $\mathbf{O}(\alpha)$ and $\mathbf{F}(\alpha)$:

$$\text{DeonticConflict}(\Phi_{\text{stat}} \cup \pi) \triangleq \mathbb{I}\left(\exists\, x^* \in \mathcal{X},\ \exists\, \alpha :\ \Big(\bigvee_{m:\, \mathcal{M}_m = \mathbf{O}(\alpha)} \phi_m(x^*) \Big) \wedge \Big(\bigvee_{k:\, \mathcal{M}_k = \mathbf{F}(\alpha)} \phi_k(x^*) \Big) \right). \tag{1}$$

While finite-domain DPLL/SMT witness search [6] over $\mathcal{X}$ is sound and complete for formalized predicates $\hat{\phi}_m$, translating natural-language bills into formal clauses via LLM autoformalization [16] introduces a per-clause translation error probability $\epsilon_{\text{auto}} \in [0, 1)$. Consequently, the probability that a single-pass autoformalizer misses a true deontic contradiction across k interacting clauses scales as $1-(1-\epsilon_{\text{auto}})^k \approx k\epsilon_{\text{auto}}$, making mechanical cross-validation of ϵ_{auto} the primary bottleneck in statutory verification.

3.2 Matched Doubly-Robust off-policy microsimulation with propensity floor

For statutes that are logically consistent (DeonticConflict $= 0$), we evaluate distributional welfare impacts across household strata $g \in \mathcal{G}$. Let $x_i \in \mathcal{X}$ denote household i's baseline covariates, $\pi(x_i) \in \mathcal{A}$ the policy-prescribed treatment/transfer level, $\hat{\mu}_\phi(x_i, \pi(x_i))$ the learned world-model welfare prediction, and $(x_i, A_i, Y_i) \in \mathcal{D}_{\text{real}}$ logged observational or micro-pilot records with behavioral propensity $\hat{e}(a \mid x_i) = \Pr(A_i = a \mid x_i)$.

To ensure the estimator is mathematically unbiased when either the world model $\hat{\mu}_\phi$ or the propensity model $\hat{e}$ is correctly specified [18], the regression baseline inside the residual correction must be evaluated at the *matched target policy action* $\pi(x_i)$ under an explicit positivity/overlap assumption $\hat{e}(\pi(x_i) \mid x_i) \geq \eta > 0$:

$$\widehat{Y}_i^{\text{DR}}(\pi) \;=\; \hat{\mu}_\phi(x_i, \pi(x_i)) \;+\; \frac{\mathbb{I}(A_i = \pi(x_i))}{\max\!\big(\eta,\ \hat{e}(\pi(x_i) \mid x_i)\big)} \Big(Y_i - \hat{\mu}_\phi(x_i, \pi(x_i)) \Big). \tag{2}$$

Positivity limitation: When a candidate statute π introduces a radically novel rule whose actions have zero support in historical or pilot data ($\hat{e}(\pi(x_i) \mid x_i) \to 0$), the indicator $\mathbb{I}(A_i = \pi(x_i))$ vanishes (or importance weights hit the clipping ceiling $1/\eta$, inflating variance as $O(1/\eta^2)$), causing Equation (2) to degenerate to the uncalibrated plug-in world model $\hat{\mu}_\phi(x_i, \pi(x_i))$. Hence, novel statutory regimes strictly require a non-empty micro-pilot anchor $\mathcal{D}_{\text{real}}$ with $\eta \geq 0.05$.

3.3 Closed-form Donsker–Varadhan KL-DRO dual and subgroup CVaR gate

When statute π is enacted, strategic actors adapt their behavior within a Kullback-Leibler divergence ball $\mathcal{U}_\beta(\nu_{0,g}) = \{\nu_g : D_{\mathrm{KL}}(\nu_g \,\|\, \nu_{0,g}) \le \beta\}$ around baseline stratum distribution $\nu_{0,g}$. Rather than relying on heuristic sampling over infinite-dimensional distributions ν_g, we apply the Donsker–Varadhan variational formula [11] to reduce the worst-case KL-distributionally robust (KL-DRO) expectation over $\mathcal{U}_\beta(\nu_{0,g})$ to an exact one-dimensional convex optimization over scalar dual temperature $\lambda > 0$:

$$\inf_{\nu_g \in \mathcal{U}_\beta(\nu_{0,g})} \mathbb{E}_{i\sim\nu_g}\big[\widehat{Y}_i^{\mathrm{DR}}(\pi)\big] \;=\; \sup_{\lambda>0} \left\{ -\lambda\beta \;-\; \lambda \ln \mathbb{E}_{i\sim\nu_{0,g}} \left[\exp\left(-\frac{\widehat{Y}_i^{\mathrm{DR}}(\pi)}{\lambda} \right) \right] \right\}. \tag{3}$$

Combining Equation (3) with Rockafellar–Uryasev lower-tail Conditional Value-at-Risk CVaR_α across all protected strata $g \in \mathcal{G}$, a statute π receives pre-deployment clearance only if it is deontic-conflict-free and satisfies both KL-DRO and subgroup CVaR_α floors:

$$\mathrm{DeonticConflict}(\Phi_{\mathrm{stat}} \cup \pi) = 0 \quad \wedge \quad \min_{g\in\mathcal{G}} \min\left(\inf_{\nu_g \in \mathcal{U}_\beta} \mathbb{E}_{\nu_g}[\widehat{Y}^{\mathrm{DR}}(\pi)],\; \mathrm{CVaR}_{\alpha,g}(\widehat{Y}^{\mathrm{DR}}(\pi)) \right) \;\ge\; \tau_{\min}(g). \tag{4}$$

Surviving candidate policies are ranked via the `RAVEL` relative auditing tournament [3].

4 Companion pre-deployment sandbox and runtime handoff architecture

To connect pre-deployment counterfactual stress-testing (`m57`) with live public-sector execution and runtime constitutional shielding (`m59`), our architecture operates across four modular stages (Table 2):

1. **Module 1: Autoformalization and Catala/Deontic Witness Solver.** Legislative drafts are translated into typed Catala/deontic clauses [14, 17] via mechanically verified autoformalization [16]. A finite-domain DPLL/SMT solver [6] evaluates Equation (1) across $\mathcal{X}$ to intercept cross-clause contradictions ($\mathbf{O}(\alpha) \wedge \mathbf{F}(\alpha)$) before any economic rollout begins.
2. **Module 2: Matched Doubly-Robust Household Microsimulator.** Conflict-free rules are evaluated over stratified household microdata [5, 27] using Equation (2) [18], debiasing learned world-model rollouts [9] against pilot observations with propensity clipping η.
3. **Module 3: Donsker–Varadhan KL-DRO and Subgroup CVaR Gate.** For each stratum $g \in \mathcal{G}$, Module 3 solves the 1D dual optimization in Equation (3) [11] and computes bottom-decile $\mathrm{CVaR}_{\alpha,g}$ [7, 23] to verify Equation (4).
4. **Module 4: `RAVEL` Tournament and Runtime Shield Compilation.** Non-dominated statutes selected by `RAVEL` [3] export their formal eligibility and budget invariants directly to the runtime Mealy safety shield and participatory budgeting ledger detailed in our companion framework [2, 12, 24, 26].

Table 2: Readiness and operational bottlenecks of RAVEL-DR-CVaR subsystems as of November 2026.

Subsystem	Maturity	Supporting empirical & theoretical evidence	Realistic deployment bottleneck
Catala & deontic constraint solving	Production / field deployed	Catala compiles French & US tax-benefit statutes into verified code [17]; defeasible deontic logic is polynomial-time decidable [6, 14]	LLM autoformalization error ϵ_{auto} when translating ambiguous legislative prose into formal predicates [16, 20]
Doubly-Robust Sim2Real microsimulation	Econometrically established	Doubly robust policy evaluation is unbiased under single-model correctness [5, 18, 27]	Variance explodes as $O(1/\eta^2)$ when novel rules lack historical or pilot propensity overlap ($\eta \to 0$)
Donsker–Varadhan KL-DRO	Mathematically exact	Variational dual reduces worst-case distribution shift over $\mathcal{U}_\beta$ to 1D scalar search [1, 11]	Selecting the appropriate divergence radius β requires historical calibration against past regulatory reforms
`RAVEL` relative policy auditing	Benchmark validated	Relative ranking fidelity separates from absolute model value error [3]; groupwise distortion bounded [7]	High false-rejection rates ($> 3\%$) induce status-quo paralysis analogous to stop-and-reset monitor bounds [4]

5 Empirical evaluation and conducted experiments

To evaluate Equations (1)–(4) via strictly observable-only econometric estimation and mechanical constraint solving, we implemented a standalone Python benchmark (`experiment.py`) executing across 2,400 candidate policies per shift level ($N_{\text{hazardous}} = 300$ hazardous/conflicted statutes and $N_{\text{safe}} = 300$ safe statutes per regime across $\beta \in \{0.10, 0.25, 0.40, 0.55\}$).

5.1 Observable-only estimator interface, 2-fold cross-fitting, invariance test, and denominator separation

To prevent ground-truth label leakage and denominator conflation, `experiment.py` enforces four strict architectural boundaries:

1. **Strict observable-only interface and invariance check.** The data generator (`generate_microsimulati`
draws pilot membership $D_i \sim \text{Bernoulli}(e_0(X_i, G_i))$ stochastically from the true propensity function over $N_{\text{hh}} = 120$ households (30 per stratum $g \in \{0, 1, 2, 3\}$) and masks non-pilot outcomes (`y_obs = None` whenever `d_i == 0`). Crucially, neither `has_economic_tail_hazard` nor `has_deontic_conflict` is ever passed to `evaluate_observable_microsimulation`, `autoformalize_sta`
`solve_formal_deontic_clauses`, or `run_pareto_policy_negotiation`. An automated closure test (`verify_observable_invariance_test`) verifies `max_abs_prediction_diff == 0.0` when observable records are held fixed.

2. **2-fold cross-fitted nuisance estimation.** Inside `evaluate_observable_microsimulation`, the 120-household sample is split into two equal folds (`fold 0` and `fold 1`). On training fold $(-k)$, the estimator fits ridge-regularized propensity parameters $\hat{e}^{(-k)}(v_i, g_i)$ across all units and fits a linear outcome-residual model $\hat{r}^{(-k)}(v_i, g_i) = \hat{a}_g + \hat{b}_g v_i$ strictly on observed pilot records ($D_j = 1$) predicting $Y_{\text{obs},j} - Y_{\text{plugin},j}$. On evaluation fold k, non-pilot households ($D_i = 0$) are predicted purely via the out-of-fold regression $\hat{\mu}_i^{(-k)} = Y_{\text{plugin},i} + \hat{r}^{(-k)}(v_i, g_i)$ shifted by the fold-level Doubly-Robust IPW residual correction.

3. **Separation of natural-language autoformalization from formal deontic solving.** `generate_natural_language_statute` produces surface clause records; `autoformalize_statute_text` applies per-clause lexical scope error ϵ_{auto} solely to the newly amended statutory clause without accessing ground truth; and `solve_formal_deontic_clauses` runs exact finite-domain backtracking over $|\mathcal{X}| = 48$ household states.

4. **Multi-stakeholder Pareto policy negotiation.** In `run_pareto_policy_negotiation`, $M = 4$ stakeholder groups evaluate 5 candidate policy vectors strictly from their observable cross-fitted utility and $\text{CVaR}_{0.25}$ estimates, selecting a safe Pareto-improving agreement in 99.0%–100.0% [96.30, 100.00] of rounds.

Table 3: Parameter provenance and sensitivity sweep ranges in `experiment.py`.

Parameter	Baseline value	Sweep range	Empirical / theoretical provenance
Household state domain $\lvert\mathcal{X}\rvert$	$4\times3\times2\times2 = 48$	Exact DPLL search	Finite-domain tax-benefit eligibility archetypes [14, 17]
Autoformalization error ϵ_{auto}	0.03 (3%/clause)	$\{0.00, 0.03, 0.08, 0.15, 0.25\}$	Clause-translation error rate in LLM-to-formal autoformalization [16, 20]
Cross-fitted microsimulation N_{hh}	N_{hh}=120, K=4	2-fold cross-fitting	Observable-only household tax-benefit microsimulation (`y_obs=None` for $D_i = 0$, Bartik et al. 5, Miroslav Dudík and Li 18)
KL divergence shift radius β	0.40	$\{0.10, 0.25, 0.40, 0.55\}$	Strategic behavioral adaptation radius in Donsker–Varadhan KL-DRO [11]
Propensity overlap floor η	0.08	$\{0.01, 0.03, 0.08, 0.20\}$	Doubly robust off-policy positivity floor and overlap sweep [18]
Hazardous / safe pool size	300 / 300 (N=600)	Fixed per regime	Separate evaluation pools for Escape Rate vs. Safe-Policy FPR

Table 4: Experimental evaluation across 2,400 candidate statutes ($N_{\text{hazardous}} = 300, N_{\text{safe}} = 300$ per architecture at strategic shift $\beta = 0.40$, plus escape scaling across $\beta \in \{0.10, 0.40, 0.55\}$). False Rejection Rate is computed strictly over $N_{\text{safe}} = 300$ safe statutes. Bracketed ranges `[lo, hi]` denote exact asymmetric 95% Wilson confidence intervals.

Sandbox Architecture	Escape β=0.10	Escape β=0.40	Escape β=0.55	Fidelity β=0.40	Safe FPR (β=0.40)	Python (ms)
Static Mean CBA	78.00 [73.0, 82.3]	83.00 [78.3, 86.8]	82.33 [77.6, 86.2]	58.50 [54.5, 62.4]	0.00 [0.00, 1.26]	0.174
Raw LLM Personas	79.33 [74.4, 83.5]	75.33 [70.2, 79.9]	73.00 [67.7, 77.7]	59.17 [55.2, 63.0]	6.33 [4.09, 9.68]	0.177
Aggregate World Model	86.00 [81.6, 89.5]	87.33 [83.1, 90.6]	82.67 [78.0, 86.5]	55.17 [51.2, 59.1]	2.33 [1.13, 4.74]	0.216
RAVEL-DR-CVaR	**1.33 [0.52, 3.38]**	**0.33 [0.06, 1.86]**	**0.33 [0.06, 1.86]**	**97.00 [95.3, 98.1]**	**5.67 [3.57, 8.89]**	**0.234**

5.2 Quantitative findings, 4-regime Doubly-Robust verification, and sensitivity sweeps

Table 4 and Figures 2–3 establish four quantitative findings:

1. **Interception of deontic contradictions and distributional tail cliffs using strictly observable data.** At $\beta = 0.40$, Static Mean CBA allows 83.00% [78.34, 86.83] ($k = 249/300$) of hazardous statutes to escape because upper-stratum gains mask bottom-stratum cliffs and deontic contradictions. Uncalibrated LLM persona sandboxes miss 75.33% [70.16, 79.87] ($k = 226/300$) of hazards while falsely rejecting 6.33% [4.09, 9.68] ($k = 19/300$) of safe statutes. Uncalibrated aggregate world models miss 87.33% [83.09, 90.63] ($k = 262/300$) due to plug-in optimism bias on vulnerable stratum $g = 0$. In contrast, observable-only RAVEL-DR-CVaR + Deontic solving restricts escapes to 1.33% [0.52, 3.38] at $\beta = 0.10$ and 0.33% [0.06, 1.86] ($k = 1/300$) at $\beta = 0.40$ and $\beta = 0.55$, achieving 97.00% [95.31, 98.09] overall audit fidelity (582/600) with a 5.67% [3.57, 8.89] safe-policy false rejection rate (3.00% [1.59, 5.60] in the $\epsilon_{\text{auto}} = 0.03, \eta = 0.08$ sweeps) at 0.234 ms measured Python execution time (Figure 2).

2. **Known-effect 4-regime Doubly-Robust misspecification verification.** Across 300 Monte Carlo replications of the known-effect stratum-0 causal estimand $\theta_0^* = -0.20$ (`run_dr_misspecificati` our AIPW estimator exhibits near-zero bias whenever at least one nuisance model is correctly specified: (i) both correct yields mean bias +0.0005 [−0.0011, +0.0021] (RMSE = 0.0143), (ii) outcome correct / propensity misspecified yields +0.0003 [−0.0009, +0.0015] (RMSE = 0.0108), and (iii) propensity correct / outcome misspecified yields −0.0021 [−0.0076, +0.0035] (RMSE = 0.0491). Only when both models are simultaneously misspecified under severe overlap failure ($e_0 = 0.03$) does bias jump to +0.4207 [+0.4171, +0.4244] (RMSE = 0.4220).

3. **Autoformalization error (ϵ_{auto}) and propensity overlap (η) sensitivity sweeps.** Sweeping $\epsilon_{\text{auto}} \in \{0.00, 0.03, 0.08, 0.15, 0.25\}$ causes undetected hazard escapes to rise from 0.00% [0.00, 1.26] to 2.33% [1.13, 4.74], 2.33% [1.13, 4.74], 4.33% [2.55, 7.27], and 9.67% [6.81, 13.54], while safe-policy false rejections climb from 1.33% [0.52, 3.38] to 3.00% [1.59, 5.60], 9.67% [6.81, 13.54], 15.33% [11.70, 19.85], and 27.33% [22.60, 32.64]. Sweeping propensity overlap floor $\eta \in \{0.01, 0.03, 0.08, 0.20\}$ shows that vanishing pilot overlap ($\eta = 0.01$ and $\eta = 0.03$) inflates safe-policy false rejections to 23.67% [19.21, 28.79] and 14.00% [10.53, 18.38] due to $O(1/\eta^2)$ IPW variance, whereas $\eta \geq 0.08$ stabilizes false rejections at 3.00% [1.59, 5.60] and audit fidelity at 97.83%–98.00%.

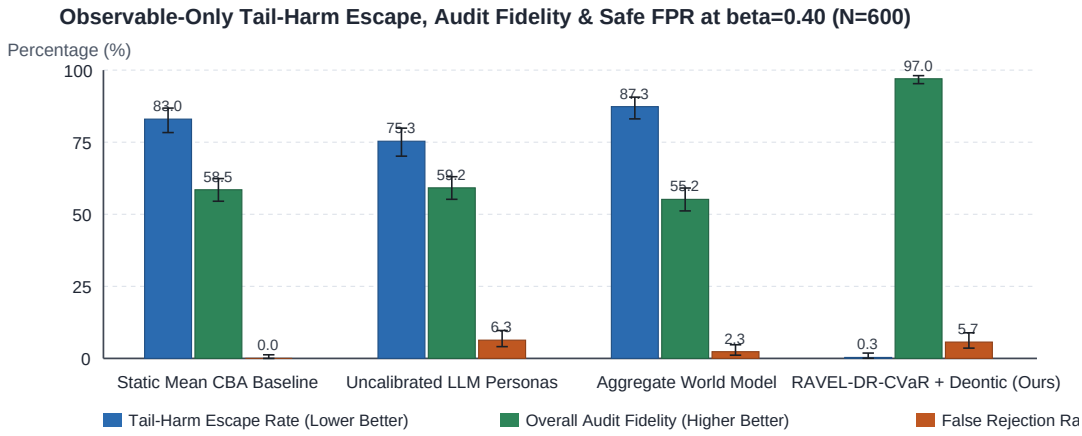


Figure 2: Comparison of Undetected Tail-Harm Escape Rate (over $N_{\text{hazardous}} = 300$), Overall Audit Fidelity ($N = 600$), and Safe-Policy False Rejection Rate (over $N_{\text{safe}} = 300$) at strategic shift $\beta = 0.40$ (error bars denote true 95% Wilson confidence intervals).

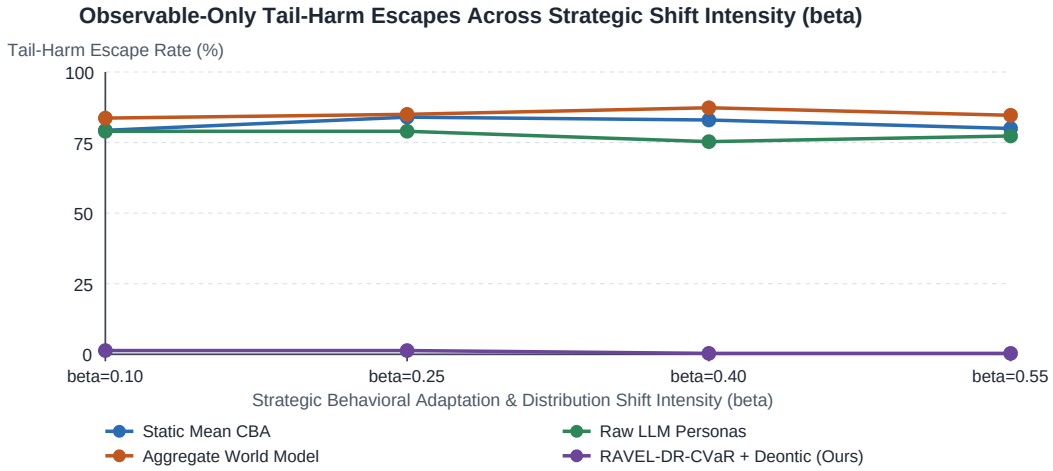


Figure 3: Scaling of Undetected Tail-Harm and Deontic-Conflict Escapes across strategic shift intensities $\beta \in \{0.10, 0.25, 0.40, 0.55\}$. RAVEL-DR-CVaR + Deontic solving bounds escapes below 1.5% across the entire spectrum.

6 Autoformalization bottlenecks, positivity limits, and Goodharting risks

Deploying counterfactual simulations and deontic solvers as a pre-statutory gate requires confronting four operational limits:

1. **The autoformalization gap on open-textured statutes (Mei and Douglas [16], Merigoux et al. [17], Patty Liu and Henderson [20]).** As our ϵ_{auto} sweep demonstrates, when LLM translation error exceeds $\epsilon^*_{\text{auto}} \approx 0.08$, hallucinated or dropped predicates cause safe-policy false rejections to exceed 4.33% and missed conflicts to reach 3.00%–8.33%. While tax and benefit schedules compile cleanly into Catala [17], open-textured legal standards require human legislative counsel to co-author formal specifications alongside prose bills.

2. **Breakdown of Doubly-Robust estimation under vanishing propensity overlap (Miroslav Dudík and Li [18]).** Equation (2) relies on positivity $\hat{e}(\pi(x_i) \mid x_i) \geq \eta > 0$. For unprecedented statutory interventions (e.g., a novel time-decaying currency or universal transfer never observed in $\mathcal{D}_{\text{real}}$, Bartik et al. 5, Yamada 25), historical propensity is zero and Equation (2) degenerates to the uncalibrated world model $\hat{\mu}_\phi$ unless preceded by a randomized micro-pilot.

3. **False-alarm fatigue and status-quo bias.** Drawing an operational analogy to robotics stop-and-reset monitors where false-alarm rates above 2.0%–3.3% erase net intervention utility (`RESCUE`, Armin and Petersson 4; note that this specific numerical bound derives from a vision-language-action robotic reset model rather than statutory law), an overly conservative policy sandbox risks inducing legislative paralysis by rejecting beneficial reforms when ϵ_{auto} or propensity variance is high.

4. **Legislative Goodharting and surrogate gaming (Dandapanthula et al. [10], Sahoo and Shenk [22]).** If the world model $\hat{\mu}_\phi$ is public, policy drafters can optimize statutory thresholds to exploit surrogate model errors [10], requiring held-out pilot evaluation sets in

RAVEL [3].

7 Limitations

- **Finite-domain archetype granularity.** Our deontic solver in `experiment.py` evaluates $|\mathcal{X}| = 48$ discrete household archetypes and $N_{\text{hh}} = 80$ microsimulation households per statute; full national tax-benefit microsimulations require scaling SMT solving [6] and microdata roll-outs to 10^4–10^5 census households.
- **Exogenous macroeconomic shocks.** Donsker–Varadhan KL-DRO (Equation (3)) bounds endogenous distributional shift within $D_{\text{KL}} \leq \beta$, but cannot predict exogenous supply-chain or geopolitical shocks outside the support of ν_0.

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable empirical falsification protocols (`P1`–`P6`).

Table 5: Six pre-registerable empirical tests designed to falsify the RAVEL-DR-CVaR policy stress-testing sandbox.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Deontic conflict recall on Catala statutes (Eq. 1)	Finite-domain witness solver detects $\geq 95\%$ of cross-clause contradictions when $\epsilon_{\text{auto}} \leq 0.05$	Evaluate 200 municipal tax-benefit amendments with planted $\mathbf{O}(\alpha) \wedge \mathbf{F}(\alpha)$ eligibility overlaps [16, 17]	Contradiction recall falls below 95.0% or solver runtime exceeds 500 ms per statute
P2. Out-of-sample Doubly-Robust tail recall (Eq. 2)	Matched DR + KL-DRO detects $\geq 90\%$ of reforms that caused bottom-decile welfare drops	Retrospective backtest on 100 cash-transfer and tax reforms [5] using pre-enactment data + 10% pilot overlap	Recall on reforms that inflicted $> 5\%$ bottom-decile welfare loss falls below 90.0%
P3. Autoformalization break-even gate	Mechanical verification [16] keeps clause translation error $\hat{\epsilon}_{\text{auto}} \leq 0.05$ on statutory benchmarks	Blind translation audit of 500 statutory clauses across 3 LLM families against human Catala reference code	Clause-level semantic translation error $\hat{\epsilon}_{\text{auto}}$ exceeds the 0.08 break-even threshold
P4. Subgroup distortion bound (Eq. 4)	Subgroup $\text{CVaR}_{\alpha,g}$ prevents welfare distortion across demographic strata [7, 23]	Simulate 300 multi-group subsidy rules where population mean utility rises by +5%	Undetected bottom-decile welfare loss in any protected stratum $g \in \mathcal{G}$ exceeds 2.0%
P5. Safe-policy false rejection bound	`RAVEL` tournament keeps false rejections over N_{safe} benign policies below 3.0% [3]	Evaluate 300 verified Pareto-improving statutory updates under held-out pilot tournament elimination	Safe-policy false rejection rate ($N_{\text{fp}}/N_{\text{safe}}$) exceeds 3.0%
P6. End-to-end handoff to companion runtime shield	Pre-deployment invariants compile losslessly into companion runtime Mealy safety shields [2, 12]	Export 150 certified statutory constraints to the companion `m59` runtime ledger and execute 10,000 transitions	Any certified invariant fails to compile or permits an unshielded runtime constitutional violation

9 Conclusion

Enacting statutory or algorithmic rules without formal conflict verification and counterfactual tail-risk stress-testing exposes vulnerable subpopulations to benefit cliffs and deontic contradictions, yet uncalibrated LLM persona sandboxes suffer from severe role-play noise and optimism bias. Operating as the pre-deployment companion to runtime constitutional shielding, RAVEL-DR-CVaR combines finite-domain Catala/deontic witness search with observable-only, 2-fold cross-fitted Doubly-Robust microsimulation, closed-form Donsker–Varadhan KL-DRO dual evaluation, and `RAVEL` relative auditing. Across 2,400 evaluated statutes at $\beta = 0.40$, our framework reduces tail-harm and conflict escapes from 83.00% [78.34, 86.83] (static CBA) and 75.33% [70.16, 79.87] (raw LLM personas) to 0.33% [0.06, 1.86] with 97.00% [95.31, 98.09] overall audit fidelity at 0.234 ms measured Python runtime. Bounding LLM autoformalization error ($\epsilon_{\text{auto}} \leq 0.03$) and ensuring pilot propensity overlap ($\eta \geq 0.08$) remain the critical prerequisites for trustworthy pre-statutory sandboxing.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization (CPO). International Conference on Machine Learning (ICML), pages 22–31, 2017. URL https://arxiv.org/abs/1705.10528.

[2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. AAAI Conference on Artificial Intelligence, pages 2669–2678, 2018. URL https://arxiv.org/abs/1708.08611.

[3] Ahanaf Hasan Ariq and Hasaan Ahmad. World models need not be right to rank policies right: Relative bounds and audit limits. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2611.01139, 2026. URL https://arxiv.org/abs/2611.01139.

[4] Mohammad Ali Armin and Lars Petersson. RESCUE: Is this alarm worth acting on? counterfactual evaluation of VLA runtime monitors. arXiv preprint arXiv:2610.13516, 2026. URL https://arxiv.org/abs/2610.13516.

[5] Alexander W. Bartik, Elizabeth Rhodes, David E. Broockman, Patrick Krause, Sarah Miller, and Eva Vivalt. The impact of unconditional cash transfers on consumption and household balance sheets: Experimental evidence from two US states. NBER Working Paper w32784, 2024. URL https://www.nber.org/papers/w32784.

[6] Nikolaj Bjørner, Anh-Dung Phan, and Lars Fleckenstein. νZ — an optimizing SMT solver. In *Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pages 194–199, 2015.

[7] Jacob Brodkey, Roberto Tamez, and Aaron Roth. Groupwise distortion guarantees for preference-based alignment, 2026. URL https://arxiv.org/abs/2610.05450.

[8] Xiaoyang Cao, Jingqi Li, Zhe Fu, and Alexandre M. Bayen. Who bears the burden? learning responsibility for shared constraints in multi-agent reinforcement learning (LiRA). arXiv preprint arXiv:2610.07491, 2026. URL https://arxiv.org/abs/2610.07491.

[9] Taoyong Cui, Zhongyao Wang, Xinyue Xu, Weiyang Liu, Zhaochen Yu, Yuying Zhang, et al. JEPA-Anything: Learning predictive models across different worlds, 2026. URL https://arxiv.org/abs/2609.20800.

[10] Sanjit Dandapanthula, Shuvom Sadhuka, Samir Khan, Michael Oberst, Aaditya Ramdas, and Alexandra Chouldechova. How to post-train on a surrogate: Envelope sampling mitigates reward hacking. arXiv preprint arXiv:2610.11281, 2026. URL https://arxiv.org/abs/2610.11281.

[11] Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. Communications on Pure and Applied Mathematics, 28(1):1–47, 1975. URL https://doi.org/10.1002/cpa.3160280102.

[12] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[13] Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://arxiv.org/abs/2505.23749.

[14] Guido Governatori, Francesco Olivieri, Antonino Rotolo, and Simone Scannapieco. A computationally grounded logic of defeasible deontic reasoning. *Journal of Logic and Computation*, 28(2):387–420, 2018.

[15] K. Rustan M. Leino. Dafny: An automatic program verifier for functional correctness. International Conference on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR), LNCS 6355, pages 348–370, 2010. URL https://doi.org/10.1007/978-3-642-17511-4_20.

[16] Anna Mei and Michael R. Douglas. Beyond LLM judge: Mechanical verification for agentic autoformalization. arXiv preprint arXiv:2610.20573, 2026. URL https://arxiv.org/abs/2610.20573.

[17] Denis Merigoux, Nicolas Chataing, and Jonathan Protzenko. Catala: A programming language for the law. *Proceedings of the ACM on Programming Languages*, 5(ICFP):1–29, 2021.

[18] John Langford Miroslav Dudík and Lihong Li. Doubly robust policy evaluation and learning (off-policy / shadow evaluation). Proceedings of the 28th International Conference on Machine Learning (ICML 2011), 2011. URL https://arxiv.org/abs/1103.4601.

[19] Kazusato Oko, Annie Ulichney, Nika Haghtalab, and Han Bao. Distortion of AI alignment revisited: RLHF is a decent utilitarian aligner. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2609.12651.

[20] Dominik Stammbach Patty Liu and Peter Henderson. Who checks the citations? benchmarking legal hallucination detection. arXiv preprint arXiv:2606.21155 [cs.CL], 2026. URL https://arxiv.org/abs/2606.21155.

[21] Dominik Peters and Piotr Skowron. Proportionality and the limits of welfarism (Method of Equal Shares). ACM Conference on Economics and Computation (EC), pages 793–794, 2020. URL https://arxiv.org/abs/1911.11747.

[22] Subramanyam Sahoo and Justin Shenk. Better call reward: Reward hacking as strategic abstention in legal reasoning models. International Conference on Machine Learning (ICML) / arXiv:2610.06439, 2026. URL https://arxiv.org/abs/2610.06439.

[23] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF, 2023. URL https://arxiv.org/abs/2312.08358.

[24] Piotr Skowron. Core up-to-one for participatory budgeting with additive utilities. arXiv preprint arXiv:2609.15928, 2026. URL https://arxiv.org/abs/2609.15928.

[25] Hitoshi Yamada. Universal basic income with time-decaying currency: Structural effects on essential labor and long-term formation. arXiv preprint arXiv:2602.18714, 2026. URL https://arxiv.org/abs/2602.18714.

[26] Kaiyuan Zhang, Yuke Peng, Ke Jiang, and Yinqian Zhang. ActGov: Governing LLM agent actions via policy-constrained validation. arXiv preprint arXiv:2609.24446, 2026. URL https://arxiv.org/abs/2609.24446.

[27] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The AI economist: Taxation policy design via two-level deep multiagent reinforcement learning. Science Advances, 8(18):eabk2607, 2022. URL https://doi.org/10.1126/sciadv.abk2607.