OPTIMIZING THE INEFFABLE: GENERATIVE POLICY LEARNING FOR HUMAN-CENTERED DECISION-MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Algorithmic decision-making is widely adopted in high-stakes applications affecting our daily lives but often requires human decision-makers to exercise their discretion within the process to ensure alignment. Explicitly modeling human values and preferences is challenging when tacit knowledge is difficult to formalize, as Michael Polanyi observed, "We can know more than we can tell." To address this challenge, we propose generative near-optimal policy learning (`GenNOP`). Our framework leverages a conditional generative model to reliably produce diverse, near-optimal, and potentially high-dimensional stochastic policies. Our approach involves a re-weighting scheme for training generative models according to the estimated probability that each training sample is near-optimal. Under our framework, decision-making algorithms focus on a primary, measurable objective, while human decision-makers apply their tacit knowledge to evaluate the generated decisions, rather than developing explicit specifications for the ineffable, human-centered objective. Through extensive synthetic and real-world experiments, we demonstrate the effectiveness of our method.

1 INTRODUCTION

We have witnessed a pragmatic shift in automated decision-making systems from a heavily first-principles approach (*e.g.*, the DENDRAL (Buchanan & Feigenbaum, 1981), the MYCIN (Van Melle et al., 1984), and the INTERNIST-1 (Miller et al., 1986) expert systems from the 1960–80s) towards a mostly empirically-grounded one (*e.g.*, IBM Watson for Oncology (Strickland, 2019), automated insulin delivery (AID) systems (Sherr et al., 2022), and the COMPAS assessment for recidivism risk (Dressel & Farid, 2018) from the 2000–20s), as empirical data availability and modeling capacity grow by orders of magnitude. The apparent success of data-driven systems is typically evidenced by their superior predictive accuracy and calibration on held-out evaluations. This aligns with classic results showing that statistical (actuarial) aggregation outperforms unaided clinical judgment for quantifiable information (Meehl, 1954; Dawes et al., 1989; Grove et al., 2000). Yet this success masks a crucial oversight, as decision-makers often conflate the *positive* capabilities of their empirical toolkits (what can be predicted or optimized for) with the inherently *normative* nature of decision-making (what ought to be done). This conflation has produced many unintended consequences: decisions that are optimal in a statistical sense but misaligned with human values.

Consider a critical care physician attending to a patient just admitted to the intensive care unit (ICU). The physician can observe the patient's conditions, gather her demographic information and medical history, and order a series of tests. Suppose the physician has access to a database containing patient characteristics (X), actions taken by critical care physicians (A), and clinical outcomes (Y), as well as an algorithm (f) derived from this database that can provide accurate and well-calibrated predictions of Y given X, A. Should the physician simply adopt the solution to the optimization problem $\arg\max_a f(X = x, A = a)$?

The algorithm provides a positive statement: "patients with characteristics x can expect a y-day reduction in length of stay at the ICU if dosage a is administered"; however, adopting the $\arg\max$ implies a normative statement: "reducing the length of stay is the sole objective of the patient's care". This implication holds for any one or combination of quantifiable clinical outcomes. Instead, the critical care physician's true normative statement is: "we should treat Ms Wang with a dosage of medications because we believe this is the best course of action for her care", stressing the importance

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

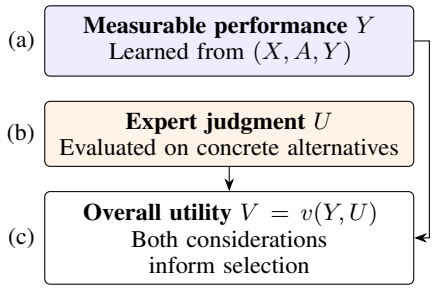


Figure 1: The learner observes Y; the expert evaluates U. Overall utility depends on both.

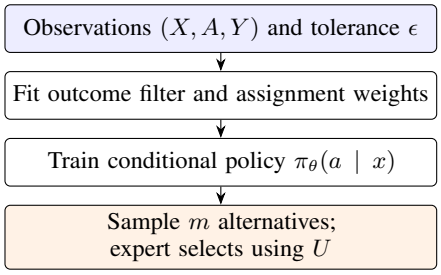


Figure 2: `GenNOP`: learn a conditional distribution of alternatives, then let the expert select.

of the individual (Tonelli, 1998). Any care derived from the true normative statement necessitates the clinical judgment by the physician to reflect the *unquantifiable* characteristics and welfare of the patient and to strike a balance among quantifiable clinical outcomes, the patient's agency for their own care, and medical ethics and best practices. We term the class of decision-making tasks where the importance of the individual requires human judgment *human-centered decision-making* problems.

To model the effect of human judgment and to formalize normative statements, we introduce two implicit quantities: (1) a human-centered objective U that can be evaluated by human judgment but never measurable; and (2) an overall utility V whose order determines true normative preferences but is *ineffable* to both humans and algorithms. Figure 1 illustrates the relationship among the measurable objective Y, the human-centered objective U, and the overall utility V, for a given action space. Specifically, as shown in Figure 1(a), Y can be measured and optimized by algorithms through explicit knowledge $f(\cdot \mid x)$ and/or by learning the contextual mapping $\hat{f}(\cdot \mid x)$ from empirical data. In contrast, as shown in Figure 1(b), U can be evaluated by human judgment but never measurable; human evaluators determine the relative preferences through their tacit knowledge $h_x(\cdot)$, which cannot be formalized as $\hat{h}(\cdot \mid x)$. Consequently, Figure 1(c) shows that we effectively face a bi-objective optimization problem with trade-offs between the two objectives, where the overall utility V is a function of both. See Section 3 for more details. Several examples of human-centered decision-making problems in various domains and their corresponding A, Y, U, V are included in Appendix B.

A natural solution to injecting human values into automated decision-making systems is to place humans-in-the-loop (HITL). However, most current HITL systems either:

1. attempt to create a proxy for human judgment either by directly eliciting human tacit knowledge (Polanyi, 1966) or through preference-based learning methods such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). However, human judgment is too complex to capture comprehensively under all contexts, observable and latent; once human judgment is formalized, it can no longer adapt to individualities, distributional shifts, and subtle changes in context, thereby losing the flexibility that made it valuable in the first place; or
2. reduce human involvement to simply accepting or rejecting singular algorithmic recommendations. An oversimplified human role can lead to both algorithmic aversion and over-reliance (Dietvorst et al., 2015; Banker & Khetani, 2019). When humans reject singular algorithmic recommendations, they resort to either making local perturbations to the recommended decisions or making *de novo* decisions on their own, leaving valuable algorithmic power untapped.

Neither approach captures the strengths of both humans and algorithms. A more promising strategy is to design for complementarity (McLaughlin & Spiess, 2024; Hemmer et al., 2024): clearly delineating roles so that humans and algorithms each operate where they excel. We adopt such a strategy and propose a framework termed "generative near-optimal policy learning" (`GenNOP`). Under our framework: (1) Human experts define a measurable objective Y and set ϵ, the acceptable gap from optimal Y-value. (2) A generative model, trained on empirical (X, A, Y) data, produces

a distribution π over actions whose expected Y-values are within ϵ of the optimum. (3) Human experts sample candidate actions from π and select the one that maximizes their judgment of U. With an appropriate choice of ϵ, the candidate set may include actions with higher overall utility V than the Y-maximizing action, giving a human evaluator a meaningful choice. Figure 2 illustrates this workflow. Appendix C compares it with other approaches.

Our Contributions Our framework offers a natural way of allocating roles to human experts and algorithms, without inducing substantial performativity (Perdomo et al., 2020), as human experts are not asked to consider balancing the measurable objective with the human-centered objective—an unnatural task that depends on the algorithmic output—except when they determine the hyperparameter ϵ. We formally define the human-centered decision-making problems in Section 3. In Section 4, we introduce our framework `GenNOP` aimed at solving these problems along with an implementation using max-stable process regression and diffusion models. In Section 5, we showcase and evaluate our framework and implementation using synthetic and real datasets.

2 Related Work

Algorithmic Decision-Making Under Ambiguity in Human Factors Many (Gabriel, 2020; Nick, 2014; Klingefjord et al., 2024; Truong & Koyejo, 2024) have highlighted the challenges in formalizing the multifaceted human values (Schwartz, 1992). Tackling these challenges by surrogate objectives has many pitfalls and carries much risk (Zhuang & Hadfield-Menell, 2020). Another stream of approaches such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and contrastive preference learning (CPL) (Hejna et al., 2024) assume that human evaluators are capable of expressing their values through preferences. Alur et al. (2024) introduce a framework leveraging "algorithmic indistinguishability" to identify specific instances where human judgment can improve algorithmic predictions. Multi-objective and set-valued approaches also preserve alternatives (Zhou et al., 2024; Lin et al., 2024). Generative curation (Li & Zhu, 2024) optimizes a portfolio using a model of residual utility dependence. `GenNOP` learns a covariate-specific near-optimal action distribution from observational assignments and outcomes without fitting such a model for U. Our framework takes a hybrid approach: following the division of the HITL decision-maker into two personas: a strategic decision-maker ("human designer")—one that determines the goals of the machine—and a practical decision-maker ("human practitioner")—one that oversees the recommended actions by the machine—in Tschiatschek et al. (2024), we put less emphasis on the human designer faithfully specifying their goal and more on the human practitioner correctly judging decisions given by the machine.

Optimization with Multiple Solutions Our work views machines through optimization. The idea of "fighting uncertainty with uncertainty" (Kashyap, 2016) dates back to the Anscombe–Aumann framework (Anscombe et al., 1963), where lotteries address ambiguity in decision-making (Hoxby & Rockoff, 2004; Wouters et al., 2018; Chan, 2013). Stochastic policies can outperform deterministic ones (Delage et al., 2019). Quality-diversity (QD) algorithms (Mouret & Clune, 2015; Cully & Demiris, 2017) relate to our work but have limitations: they need random variations in action spaces, rely on model-based metrics with performance issues (Maragno et al., 2023), and require intensive computation. Other approaches include simulated annealing (Van Laarhoven et al., 1987; Bertsimas & Tsitsiklis, 1993), large-scale neighborhood search (Ahuja et al., 2002; Pisinger & Ropke, 2019), and multi-objective optimization (Deb et al., 2016). These iterative methods face computational costs and stability-convergence tradeoffs (Chen et al., 2018). Our approach leverages generative models, eliminating iterative search requirements. Learning to defer routes decisions between a model and a human (Madras et al., 2018); conformal decision support supplies prediction sets with coverage guarantees (Straitouri et al., 2023; De Toni et al., 2024; Hullman et al., 2025; Ruggieri & Pugnana, 2025). `GenNOP` instead learns a distribution of actions subject to a measurable-outcome tolerance, from which a human selects.

Generative Models for Optimization Recent generative model advances have been applied to optimization problems. Generative model-based optimization (GMO) methods (Nguyen et al., 2016; Lu et al., 2018; Guo et al., 2022) enable optimization in high-dimensional spaces by using Bayesian optimization and evolutionary algorithms in latent spaces. However, these risk mode collapse due to over-reliance on latent space structure. Our method, while still assuming low-dimensional latent

spaces, directly generates action space samples rather than projecting latent-space interpolations. Another approach uses generative models for black-box optimization (BBO) (Krishnamoorthy et al., 2022; Li et al., 2024), applying conditional generative models with outcomes as conditions.

3 HUMAN-CENTERED DECISION-MAKING

Problem Setup We assume access to n i.i.d. observations $\{(\mathbf{a}_i, \mathbf{x}_i, y_i)\}_{i=1}^n \sim \mathcal{D}$ from an offline dataset with covariates $\mathbf{x}_i \in \mathcal{X}$ that characterize individual i, action $\mathbf{a}_i \in \mathcal{A}$ taken by individual i, and measurable objective value $y_i \in \mathbb{R}$. Note that $\mathbf{x}_i, \mathbf{a}_i$ are vectors, as `GenNOP` admits multi-dimensional covariates and actions. For notational convenience, we use x_i, a_i thereafter in place of vector notations. We adopt the potential outcomes framework (Rubin, 1974; Imbens & Rubin, 2015).

Formally, our hybrid decision-maker solves a utility-maximization problem with regard to the chosen decision (or *action*) a:

$$\max_a V_a, \tag{1}$$

where $V_a = v(Y_a, U_a)$ is the overall utility (or *value*) of an action, $Y_a = y(a)$ the measurable objective, and $U_a = u(a)$ the human-centered objective. The decision-maker does not have full knowledge about the shape of $v(\cdot, \cdot)$ but can impose mild assumptions about it.

Unlike the conventional *human-agnostic* setup which aims at optimizing for Y, under the *human-centered* setup, the goal (as in Equation (1)) is to optimize for V, the overall utility of an action as a function of the measurable objective Y and the human-centered objective U, by incorporating both the observed dataset $\mathcal{D}$ and a human evaluator in the loop. Key notations in this paper are listed in Table A.

ϵ-Optimality vs Quantitative Optimality If there is some $a^\star$ such that $a^\star = \arg\max_a Y_a = \arg\max_a U_a$, intuitively, we should never sacrifice quantitative optimality in Y. In practice, however, Y_a and U_a can trade off near their maxima. For example, a movie can gain box-office revenue at a cost to its artistic value. Appendix D gives further examples. To formalize such cases, we assume a strictly concave Pareto frontier and an overall utility that increases in both objectives and has a unique interior optimum; the precise assumptions are in Appendix E.1.

Proposition 3.1 (Maxima non-coincidence)**.** *Let $\{(Y^{(n)}, U^{(n)})\}_{n=1}^N$ be i.i.d. draws satisfying Assumptions E.1 — E.3. Define:*

$$M_Y \in \arg\max_{1\le n\le N} Y^{(n)}, \qquad M_V \in \arg\max_{1\le n\le N} v\left(Y^{(n)}, U^{(n)}\right),$$

with arbitrary tie-breaking. Then:

$$\lim_{N\to\infty} \mathbb{P}(M_Y = M_V) \;=\; 0.$$

Proposition 3.2 (Local V-order dominance)**.** *Under Assumptions E.1–E.5, for any finite sample with distinct U-values there exists a sample-dependent $\bar\epsilon \in (0, \epsilon]$ such that for any two sampled points with*

$$Y^{(i)}, Y^{(j)} \in [y^\star - \bar\epsilon,\, y^\star],$$

we have, writing $V^{(k)} := v\big(Y^{(k)}, U^{(k)}\big)$,

$$U^{(i)} \;\ge\; U^{(j)} \quad\Longrightarrow\quad V^{(i)} \;\ge\; V^{(j)}.$$

Consequently, within any finite sample restricted to $y \in [y^\star - \bar\epsilon,\, y^\star]$, an $\arg\max U$ is also an $\arg\max V$.

See Appendix E.2 for proofs. We formally define ϵ-optimality of actions for an individual characterized by covariates x as follows:

Definition 3.3 (ϵ-optimality)**.** For some $\epsilon \ge 0$, an action $a \in \mathcal{A}$ is considered ϵ-optimal if

$$|y^\star(x) - \mathbb{E}[Y_a \mid X = x]| \le \epsilon,$$

where $y^\star(x) = \max_{a\in\mathcal{A}} \mathbb{E}[Y_a \mid X = x]$ is the optimal Y-value for an individual with covariates x.

Under this definition, ϵ-optimality can vary from individual to individual. For example, treatments that are considered near-optimal for an otherwise healthy individual are intuitively different from those considered near-optimal for an individual with many comorbidities, even when the treatments are intended for the same condition.

Given a tolerance ϵ, we can generate candidate actions whose expected Y-values are close to $y^\star(x)$ and ask a human evaluator to select among them using U. The finite-sample ordering result above requires a sample-dependent tolerance and does not by itself guarantee global V-optimality. In practice, ϵ is an acceptable loss expressed in the units of Y, specified using domain knowledge without requiring a numerical model of v. Figure 5 examines the resulting trade-off across tolerances.

4 Generative Near-Optimal Policy Learning (GenNOP)

Our proposed framework, GenNOP, aims at using a conditional generative model to learn a stochastic policy that is ϵ-optimal in Y-value. Compared to a deterministic policy that is quantitatively optimal in Y-value, an ϵ-optimal stochastic policy offers human evaluators choices from which a higher V-value (overall utility) is attained. The hyperparameter ϵ represents the acceptable performance trade-off and is calibrated heuristically in practice by domain experts based on their tolerance for deviation from the quantitative optimality. We precisely define our target policy of interest, ϵ-optimal stochastic policies, as follows:

Definition 4.1 (ϵ-optimal policies)**.** An ϵ-optimal policy $\pi_\epsilon^\star$ is a stochastic policy that maps covariates x to the uniform distribution of all ϵ-optimal actions:

$$\pi_\epsilon^\star(a|x) = \frac{1}{\nu(\Omega_\epsilon^\star(x))} \mathbf{1}\left\{a \in \Omega_\epsilon^\star(x)\right\}, \tag{2}$$

where $\Omega_\epsilon^\star(x) = \{a : |y^\star(x) - \mathbb{E}[Y_a \mid X = x]| \leq \epsilon\}$ and $\mathbf{1}\{\cdot\}$ denotes the indicator function. Here ν is counting measure for discrete actions or Lebesgue measure on the chosen continuous action coordinates, with $0 < \nu(\Omega_\epsilon^\star(x)) < \infty$. A lower-dimensional feasible support requires a reference measure on that support. Densities and uniformity are defined relative to this choice.

Figure 3 illustrates ϵ-optimality and ϵ-optimal policy. Whereas the density of the absolute-optimal policy $p(\pi^\star(x))$ represents a point mass in the action space, that of the ϵ-optimal policy $\Omega_\epsilon^\star(x)$ represents a richer, more diverse, and potentially multi-modal density over the action space, at the cost of ϵ in the Y-space, which is desirable for practitioners under various settings.

4.1 Conditional Generative Model as $\pi_\epsilon^\star$

The goal of GenNOP is to train a generative model parametrized by θ whose generative distribution π_θ approximates $\pi_\epsilon^\star$. To this end, we define our learning objective as minimizing the Kullback-Leibler (KL) divergence between the target policy $\pi_\epsilon^\star$ and the generative policy π_θ:

$$\min_\theta L(\theta) = \mathbb{E}_{x\sim p(X)}[D_{\mathrm{KL}}[\pi_\epsilon^\star(\cdot|x) \,||\, \pi_\theta(\cdot|x)]].$$

If we have a dataset $\mathcal{D}^\star = \{(x_i, a_i)\}_{i=1}^n$ where x_i is drawn from $p(X)$ and a_i from the true ϵ-optimal policy $\pi_\epsilon^\star$, our learning objective above can be conveniently expressed in distribution as minimizing (see Appendix H.1 for derivation):

$$L(\theta) \overset{d}{=} \mathbb{E}_{(x,a)\sim\mathcal{D}^\star}[-\log \pi_\theta(a|x)]. \tag{3}$$

However, we cannot directly obtain action samples drawn from the target policy $\pi_\epsilon^\star$. Instead, we perform a *re-weighting* step to the observational dataset $\mathcal{D}$, so that the distributions of action samples drawn from the re-weighted dataset approximate the distributions of those drawn from $\pi_\epsilon^\star$. We rely on two quantities:

1. $g_\epsilon(y, x) \in [0, 1]$ is a soft outcome filter based on the fitted CDF of local block maxima, evaluated at $y + \epsilon$. Its estimation is discussed in Section 4.2.
2. $p(a|x)$ is the generalized propensity score (GPS) of action a given covariates x. It is used to transform the filtered *observational* distribution into a *counterfactual* distribution via inverse

probability weighting (IPW). Its estimation is discussed in Appendix I.1 by adopting the strategy by Zou et al. (2020).

We define the weight function of an observation given ϵ as $w(x, a, y; \epsilon) = g_\epsilon(y, x)/p(a|x)$, which is identifiable from observational data under standard causal assumptions in Appendix F. Adopting the conditional diffusion model with classifier-free guidance parametrized by θ as the generative policy π_θ, we re-weight its loss function (5) and have the following learning objective (see Appendix H.2 for derivation): For a fixed filter and exact propensity density, define

$$h_\epsilon(a, x) = \mathbb{E}[g_\epsilon(Y, x) \mid A = a, X = x], \qquad q_\epsilon(a \mid x) = \frac{h_\epsilon(a, x)}{\int h_\epsilon(a', x)\, d\nu(a')}.$$

Under overlap and a finite positive denominator, weighting induces q_ϵ. It equals $\pi^\star_\epsilon$ when outcomes are deterministic given (A, X) and the filter is the exact near-optimality indicator. With noisy outcomes, the induced density instead depends on expected filter acceptance; Appendix H.2 gives the derivation.

$$L(\theta) = \mathbb{E}_{t,x,a,y,\varepsilon}\left[w(x, a, y; \epsilon) \cdot \|\varepsilon - \varepsilon_\theta\left(a_t, t, x\right)\|^2\right]. \tag{4}$$

Overall, our strategy for learning π_θ can be viewed as a two-stage learning process: (1) weight construction via training neural networks that parametrize generalized extreme value (GEV) distributions and estimating the GPS via a variational autoencoder (VAE); and (2) conditional diffusion model training with re-weighted objective. In the first stage, the filter emphasizes high outcomes and propensity weighting corrects the observational assignment distribution; their induced target is q_ϵ above. The training procedure is summarized in Appendix Algorithm 1.

4.2 Estimating Conditional Optimality via Max-Stable Process Regression

We estimate an outcome filter from local block maxima. If $F_{b,x}$ is the fitted CDF of a maximum over b locally comparable observations, set $g_\epsilon(y, x) = F_{b,x}(y + \epsilon)$. This gives higher weights to observations whose outcomes, plus the tolerance, compare favorably with local maxima.

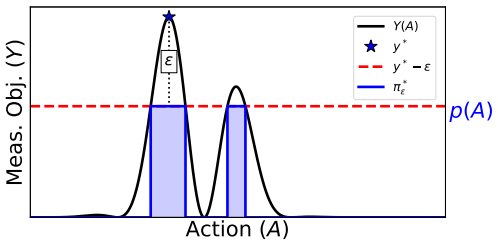


Figure 3: ϵ-Optimal Policy

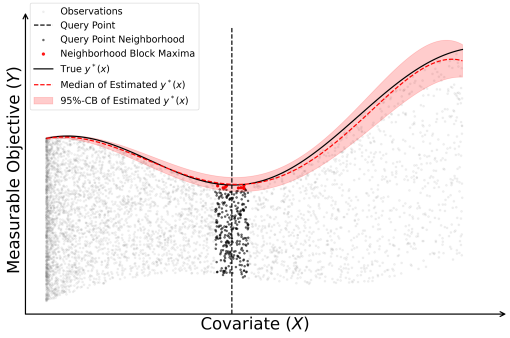


Figure 4: Max-Stable Process Regression

For deterministic outcomes with a finite upper endpoint and positive sampling probability arbitrarily close to the optimum, block maxima approach the optimal outcome as block size grows. With outcome noise, their distribution also reflects the noise tail; sparse coverage of good actions affects the approximation. We therefore use the fitted CDF as a soft filter rather than a posterior distribution for $y^\star(x)$.

Let $\mathcal{X}$ be a metric space with distance $d(\cdot, \cdot)$. For each x_i in our sample, we treat its kb nearest neighbors as locally comparable to x_i, randomly partition them into k blocks of size b, and take the maximum within each block: $\{y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(k)}\}$. We repeat this for $i \in \{1, \ldots, n\}$. Under appropriate tail regularity, extreme-value theory motivates a GEV approximation to normalized block maxima. Local pooling further requires the outcome distribution to vary slowly within neighborhoods. We fit covariate-dependent GEV marginals; because neighborhoods overlap, summing their log-likelihoods gives a composite marginal objective.

To allow μ, σ, and ξ to vary with x, we write: $\mu(x;\psi_1),\ \sigma(x;\psi_2),\ \xi(x;\psi_3)$, where $\psi = \{\psi_1, \psi_2, \psi_3\}$ are parameters of neural networks. The GEV log-likelihood for block maxima at location x_i is:

$$\ell\Big(\{y_i^{(j)}\}; \mu(x_i;\psi_1), \sigma(x_i;\psi_2), \xi(x_i;\psi_3)\Big),$$

subject to the positivity constraint $1 + \xi(x_i;\psi_3)\,\frac{y_i^{(j)} - \mu(x_i;\psi_1)}{\sigma(x_i;\psi_2)} > 0$ for all $j = 1, \ldots, k$. Maximizing the sum of the log-likelihoods over $i = 1, \ldots, n$ with respect to ψ yields the fitted parameters $\hat{\psi}$. The resulting filter is:

$$g_\epsilon(y, x) = F_{\mathrm{GEV}}\big(y + \epsilon; \mu(x;\hat{\psi}_1), \sigma(x;\hat{\psi}_2), \xi(x;\hat{\psi}_3)\big).$$

The CDF provides a smooth transition between low and high weights, avoiding a hard cutoff. Figure 4 illustrates the local observations, block maxima, and fitted model against the known synthetic optimum. The displayed GEV range describes the fitted maxima distribution, rather than a confidence interval for the optimal expected outcome.

4.3 Ablation Study of GEV Parameter Estimation

We vary the number of blocks k, block size b, and GEV regularization, fitting each setting from five random initializations. Table 1 reports a representative subset of negative log-likelihoods (mean and standard deviation; lower is better). At $k = 20, b = 30$, regularization strength 1 yields $-2.17\,(0.17)$ in 1D and $-1.22\,(0.12)$ in 2D, compared with $-0.15\,(3.27)$ and $-0.56\,(1.02)$ without regularization. The 2D result at strength 10 is worse at this setting, so stronger regularization does not uniformly improve fit. Block size and block count also matter; the complete grid and fitted examples are in Appendix J.

Table 1: GEV parameter ablation: negative log-likelihood, mean (standard deviation) over five initializations. Lower is better.

Dimension	Regularization	$k = 10, b = 30$	$k = 20, b = 30$	$k = 50, b = 30$
1D	0	$0.52\,(2.90)$	$-0.15\,(3.27)$	$-1.24\,(0.90)$
1D	1	$-2.06\,(0.17)$	$-2.17\,(0.17)$	$-2.05\,(0.07)$
1D	10	$-1.84\,(0.31)$	$-1.72\,(0.55)$	$-1.82\,(0.07)$
2D	0	$3.09\,(4.58)$	$-0.56\,(1.02)$	$0.29\,(3.11)$
2D	1	$-0.95\,(0.65)$	$-1.22\,(0.12)$	$-0.29\,(1.56)$
2D	10	$-1.25\,(0.13)$	$-0.72\,(0.42)$	$-0.36\,(0.47)$

5 Experiments

Synthetic Results We created the following synthetic datasets to compare `GenNOP` with the baseline methods: (1) A fully-synthetic dataset in which covariates, actions, and measurable objective values are all 1-dimensional and bounded on $(0, 1)$; (2) A semi-synthetic dataset in which actions are represented by images drawn from the `Fashion-MNIST` dataset to showcase the capability of handling high-dimensional action spaces; covariates in this dataset are also multi-dimensional. Details can be found in Appendix K.

We compare with `DRPolicyForest`, sampling from its action probabilities; `GP-UCB`, optimizing a Gaussian-process upper confidence bound; and `DDOM`, conditioning generation on covariates and outcomes. Precision measures the fraction of generated actions in the target set, recall measures coverage under the discretization or classes used in each experiment, and FID compares generated and reference images in feature space. Appendix K gives the protocols.

Table 2 reports means over covariates, with fifth-percentile results in parentheses and variability over ten generated samples after $\pm$; 0.00 denotes values below 0.005. `GenNOP` performs strongly on both datasets, particularly in recall, though its ablation without propensity weighting has higher semi-synthetic precision. The fifth-percentile columns indicate variation across individuals that an overall mean would hide.

Table 2: Evaluation Metrics

Method	Fully-synthetic dataset		Semi-synthetic dataset		
	Precision ↑	Recall ↑	Precision ↑	Recall ↑	FID ↓
`GenNOP`	0.85 ± 0.00 (0.47 ± 0.00)	0.97 ± 0.00 (0.80 ± 0.00)	0.83 ± 0.03 (0.38 ± 0.06)	0.69 ± 0.03 (0.27 ± 0.03)	5.0 ± 1.8
`GenNOP` w/o p	0.83 ± 0.00 (0.37 ± 0.01)	0.94 ± 0.00 (0.80 ± 0.00)	0.87 ± 0.03 (0.44 ± 0.08)	0.67 ± 0.02 (0.25 ± 0.03)	6.5 ± 3.4
`GenNOP` w/o g_ϵ	0.01 ± 0.00 (0.02 ± 0.00)	0.45 ± 0.00 (0.10 ± 0.00)	0.38 ± 0.02 (0.00 ± 0.00)	0.45 ± 0.02 (0.00 ± 0.00)	14.3 ± 3.8
`DRPolicyForest` (Athey & Wager, 2021)	0.37 ± 0.00 (0.00 ± 0.00)	0.06 ± 0.00 (0.00 ± 0.00)	0.44 ± 0.02 (0.00 ± 0.00)	0.33 ± 0.03 (0.00 ± 0.00)	22.0 ± 3.3
`DDOM` (Krishnamoorthy et al., 2023)	0.29 ± 0.00 (0.00 ± 0.00)	0.11 ± 0.00 (0.00 ± 0.00)	0.66 ± 0.04 (0.04 ± 0.08)	0.54 ± 0.03 (0.04 ± 0.09)	34.7 ± 12.4
`GP-UCB` (Srinivas et al., 2010)	0.13 ± 0.00 (0.00 ± 0.00)	0.27 ± 0.00 (0.00 ± 0.00)	0.40 ± 0.03 (0.00 ± 0.00)	0.43 ± 0.05 (0.00 ± 0.00)	19.1 ± 7.3

Removing the near-optimality filter g_ϵ substantially reduces precision and recall. Omitting propensity weighting has a smaller, mixed effect across the reported metrics, consistent with the observational action distribution differing from the target uniform distribution.

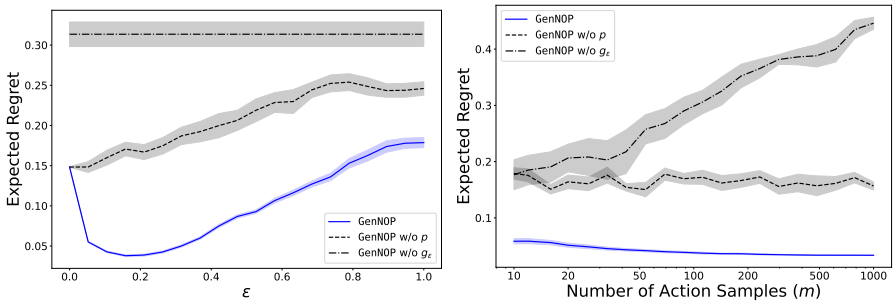


Figure 5: Effect of Hyperparameters on Expected Regret

To bring ϵ-optimal policy into the context of human-centered decision-making, we created an end-to-end synthetic dataset by including U, V in its data-generating process. We define the regret of an action a for an individual with covariates x as the difference between the optimal overall utility value (V) for that individual and the overall utility value attained by that action: $\text{Regret}(a, x) = v^\star(x) - V_a | X = x$, where $v^\star(x) = \max_a V_a | X = x$; additionally, we define the regret of a stochastic policy $\pi(x)$ given the number of draws m as the difference between $v^\star(x)$ and the U-maximizing action among the m draws: $\text{Regret}(\pi, x; m) = v^\star(x) - V_{a^\star} \mid X = x$, where $a^\star = \arg\max_{a \in \{a_i\}_{i=1}^m} U_a \mid X = x$ and $a_i \sim \pi(\cdot \mid x)$. We examine how the choices of ϵ and m influence expected regret in Figure 5. Regret falls as ϵ increases from zero to an intermediate value, then rises; with a suitable ϵ, a few candidate actions suffice to approach the minimum regret. This is encouraging because experts can evaluate only a small candidate set in practice.

Variation in Human Evaluation We vary the decision-maker's preference between the measurable and human-centered objectives, setting $\lambda_Y = 1 - \lambda_U$ and $\lambda_V = 0$ in the selection score $Q = \lambda_Y Y + \lambda_U U$. Figure 6 reports regret from five repeated synthetic evaluations with 100 candidate actions per choice.

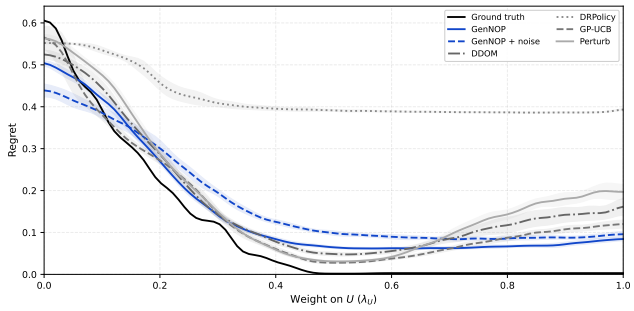


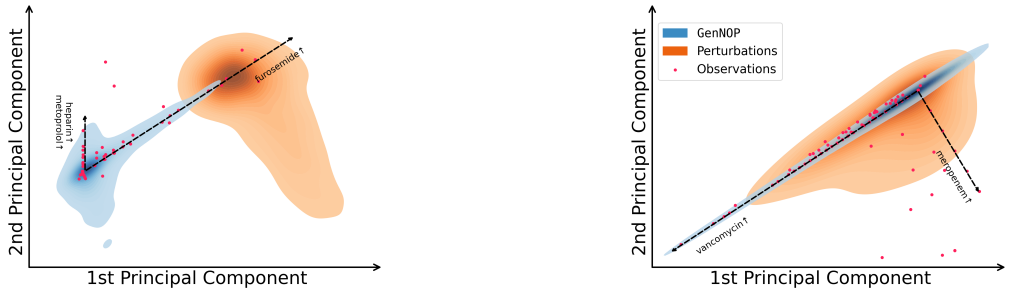
Figure 6: Regret along the Y–U preference axis ($\lambda_V = 0$). Curves are smoothed means over five runs; bands show one across-run standard deviation. Each run uses 100 sets of 100 candidate actions per preference weight. Noise scales each candidate's perceived outcomes by a factor in $[0.8, 1.2]$.

As λ_U approaches one, GenNOP has lower regret than the evaluated baselines: at $\lambda_U = 1$, its mean regret is 0.085 ± 0.004, versus 0.168 ± 0.020 for DDOM. The advantage is not uniform: near $\lambda_U = 0.5$, DDOM, GP−UCB, and local perturbation have lower regret than GenNOP. Applying a shared random multiplier in $[0.8, 1.2]$ to each candidate's perceived outcomes increases GenNOP's regret at $\lambda_U = 1$ to 0.096 ± 0.007. These are simulations of human selection, not human-rater observations. Appendix K.4 gives the evaluation protocol and broader preference maps.

The preference-axis simulation uses the oracle tolerance $\epsilon = \max_a Y_a - Y_{a^\star_V}$, where $a^\star_V = \arg\max_a V_a$, available from its synthetic ground truth. This choice isolates variation in selection preferences; it is not a calibration procedure for applications.

Real Datasets We apply our framework to the dosing problem mentioned in the Introduction, which is a good example of a human-algorithm hybrid system solving a human-centered decision-making problem. To this end, we extracted 2 datasets from the Medical Information Mart for Intensive Care (MIMIC)-IV (Johnson et al., 2023) dataset: (1) mimic-icu-cardio and (2) mimic-icu-sepsis, which contain dosages of sets of medications (A), patient characteristics (X), and measurable objectives (Y) of patients admitted to ICUs for cardiovascular and sepsis diagnoses. Details can be found in Appendix K.5.

Figure 7 displays a two-dimensional VAE embedding of observed and generated actions. Generation uses the full four-dimensional cardiovascular and five-dimensional sepsis medication vectors. We compare GenNOP with Gaussian perturbations of recorded short-stay actions. The generated distribution covers more of the observed action range in the cardiovascular cohort and better follows its main axis in the sepsis cohort. These descriptive plots assess distributional alignment with recorded clinician actions; they do not establish that generated doses are clinically optimal or safe.



(a) mimic-icu-cardio (b) mimic-icu-sepsis

Figure 7: 2-D VAE representations of observed actions and generated action distributions; the axes labeled PC1/PC2 denote VAE coordinates.

6 Discussion

GenNOP generates individualized sets of actions that trade a bounded amount of measurable performance for a broader range of choices. Its intended benefit depends on the human evaluator's ability to identify important qualities that the model cannot encode.

The synthetic evaluations assume a simplified model of how people choose among candidates; actual preferences may depend on the displayed set. In MIMIC-IV, recorded clinician actions are a comparison distribution rather than verified optimal decisions, and the cohort is selected from observational records. Clinical deployment would require expert assessment of proposed doses, safety constraints, and prospective evaluation. Learning also depends on sufficient action coverage in every patient subgroup. We leave these evaluations and safeguards to future work.

Reproducibility Statement

Open Access to Code Repository Access to our implementation of `GenNOP` is available at `https://anonymous.4open.science/r/GenNOP/`.

Compute Resources All experiments, with the exception of those on the semi-synthetic dataset, should not require significant compute resources beyond the CPU capabilities of a typical recent personal computer. Specifically, they were conducted using a laptop computer with an Apple M3 Pro chip with 18 GB of memory. Each experiment took an insignificant amount of time (on the order of a few minutes).

The experiments on the semi-synthetic dataset were conducted using a virtual machine from Google Cloud Compute Engine with the following configuration: `n1-standard-8 (8 vCPUs, 30 GB Memory), 1 x NVIDIA Tesla P100, 100 GB Storage`. These experiments (including the repetitions needed to establish statistical significance) took about 8 hours.

Ethics Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, among which we highlight the following:

- The use of automation in decision-making can restrict human agency and autonomy. Suggesting a set of potential actions instead of a single action to human decision-makers may increase the risk of creating an illusion of agency, thereby undermining any potential mitigating efforts that are otherwise taken.
- Our method relies on the availability of past actions to learn a generative model. Our method may perform better on groups of people with more available past actions.
- Our method may be prone to adversarial injection of a small number of false observations into the training data, which can cause the generative model to perform poorly for certain groups of people.

AI Use Statement

We used LLMs to reorganize and polish writing; identify related literature; propose and assess research ideas and methodological revisions, including analysis of existing experiments; and help formulate and check mathematical claims and proof steps. We manually reviewed the cited literature. The authors take responsibility for the final content.

References

Ravindra K Ahuja, Özlem Ergun, James B Orlin, and Abraham P Punnen. A survey of very large-scale neighborhood search techniques. *Discrete Applied Mathematics*, 123(1-3):75–102, 2002.

Rohan Alur, Manish Raghavan, and Devavrat Shah. Human expertise in algorithmic prediction. *Advances in Neural Information Processing Systems*, 37:138088–138129, 2024.

Francis J Anscombe, Robert J Aumann, et al. A definition of subjective probability. *Annals of mathematical statistics*, 34(1):199–205, 1963.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. doi: 10.3982/ECTA15732.

Sachin Banker and Salil Khetani. Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. *Journal of Public Policy & Marketing*, 38(4):500–515, 2019.

Dimitris Bertsimas and John Tsitsiklis. Simulated annealing. *Statistical science*, 8(1):10–15, 1993.

Bruce G Buchanan and Edward A Feigenbaum. Dendral and meta-dendral: Their applications dimension. In *Readings in artificial intelligence*, pp. 313–322. Elsevier, 1981.

Ernie Chan. *Algorithmic trading: winning strategies and their rationale*. John Wiley & Sons, 2013.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. *arXiv preprint arXiv:1804.01619*, 2018.

Antoine Cully and Yiannis Demiris. Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation*, 22(2):245–259, 2017.

Robyn M Dawes, David Faust, and Paul E Meehl. Clinical versus actuarial judgment. *Science*, 243 (4899):1668–1674, 1989.

Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Rodriguez. Towards human-ai complementarity with prediction sets. *Advances in Neural Information Processing Systems*, 37:31380–31409, 2024.

Kalyanmoy Deb, Karthik Sindhya, and Jussi Hakanen. Multi-objective optimization. In *Decision sciences*, pp. 161–200. CRC Press, 2016.

Erick Delage, Daniel Kuhn, and Wolfram Wiesemann. "dice"-sion–making under uncertainty: when can a random decision reduce risk? *Management Science*, 65(7):3282–3301, 2019.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1):114, 2015.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Robert Gauer, Damon Forbes, and Nathan Boyer. Sepsis: Diagnosis and management. *Am. Fam. Physician*, 101(7):409–418, April 2020.

William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12(1):19, 2000.

Zhendong Guo, Haitao Liu, Yew-Soon Ong, Xinghua Qu, Yuzhe Zhang, and Jianmin Zheng. Generative multiform bayesian optimization. *IEEE Transactions on Cybernetics*, 53(7):4347–4360, 2022.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. Contrastive preference learning: Learning from human feedback without rl, 2024. URL https://arxiv.org/abs/2310.13639.

Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. Complementarity in human-ai collaboration: Concept, sources, and evidence. *arXiv preprint arXiv:2404.00029*, 2024.

Keisuke Hirano and Guido W. Imbens. *The Propensity Score with Continuous Treatments*, chapter 7, pp. 73–84. John Wiley & Sons, Ltd, 2004. ISBN 9780470090459. doi: https://doi.org/10.1002/0470090456.ch7. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0470090456.ch7.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022.

Caroline Minter Hoxby and Jonah E Rockoff. *The impact of charter schools on student achievement*. Department of Economics, Harvard University Cambridge, MA, 2004.

Jessica Hullman, Yifan Wu, Dawei Xie, Ziyang Guo, and Andrew Gelman. Conformal prediction and human decision making. *arXiv preprint arXiv:2503.11709*, 2025.

Guido W. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000. ISSN 00063444. URL http://www.jstor.org/stable/2673642.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL http://dx.doi.org/10.1038/s41597-022-01899-x.

Ravi Kashyap. Fighting uncertainty with uncertainty: A baby step. *arXiv preprint arXiv:1601.04043*, 2016.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align ai to them?, 2024. URL https://arxiv.org/abs/2404.10636.

Siddarth Krishnamoorthy, Satvik Mehul Mashkaria, and Aditya Grover. Generative pretraining for black-box optimization. *arXiv preprint arXiv:2206.10786*, 2022.

Siddarth Krishnamoorthy, Satvik Mehul Mashkaria, and Aditya Grover. Diffusion models for black-box optimization. *arXiv preprint arXiv:2306.07180*, 2023.

Michael Lingzhi Li and Shixiang Zhu. Balancing optimality and diversity: Human-centered decision making through generative curation, 2024. URL https://arxiv.org/abs/2409.11535.

Zihao Li, Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Yinyu Ye, Minshuo Chen, and Mengdi Wang. Diffusion model for data-driven black-box optimization. *arXiv preprint arXiv:2403.13219*, 2024.

Bo Lin, Erick Delage, and Timothy CY Chan. Conformal inverse optimization. *arXiv preprint arXiv:2402.01489*, 2024.

Xiaoyu Lu, Javier Gonzalez, Zhenwen Dai, and Neil D Lawrence. Structured variationally auto-encoded optimization. In *International conference on machine learning*, pp. 3267–3275. PMLR, 2018.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

Donato Maragno, Holly Wiberg, Dimitris Bertsimas, Ş İlker Birbil, Dick den Hertog, and Adejuyigbe O Fajemisin. Mixed-integer optimization with constraint learning. *Operations Research*, 2023.

Bryce McLaughlin and Jann Spiess. Designing algorithmic recommendations to achieve human-ai complementarity. *arXiv preprint arXiv:2405.01484*, 2024.

Paul E. Meehl. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press, Minneapolis, MN, 1954. doi: 10.1037/11281-000.

Randolph A Miller, Melissa A McNeil, Sue M Challinor, Fred E Masarie Jr, and Jack D Myers. The internist-1/quick medical reference project—status report. *Western Journal of Medicine*, 145(6): 816, 1986.

Michael Moor, Nicolas Bennett, Drago Plecko, Max Horn, Bastian Rieck, Nicolai Meinshausen, Peter Bühlmann, and Karsten M. Borgwardt. Predicting sepsis in multi-site, multi-national intensive care cohorts using deep learning. *CoRR*, abs/2107.05230, 2021. URL https://arxiv.org/abs/2107.05230.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, 29, 2016.

Bostrom Nick. *Superintelligence: Paths, dangers, strategies*. Oxford University Press, Oxford, 2014.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

David Pisinger and Stefan Ropke. Large neighborhood search. *Handbook of metaheuristics*, pp. 99–127, 2019.

Owens PL, Miller MA, Barrett ML, and Hensche M. Overview of outcomes for inpatient stays involving sepsis, 2016–2021. HCUP Statistical Brief Statistical Brief #306, Agency for Healthcare Research and Quality, Rockville, MD, April 2024. URL https://hcup-us.ahrq.gov/reports/statbriefs/sb306-overview-sepsis-2016-2021.pdf.

Michael Polanyi. *The Tacit Dimension*. University of Chicago Press, 1966.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment - a deep reinforcement learning approach. *CoRR*, abs/1705.08422, 2017. URL http://arxiv.org/abs/1705.08422.

Eamon P Raith, Andrew A Udy, Michael Bailey, Steven McGloughlin, Christopher MacIsaac, Rinaldo Bellomo, David V Pilcher, et al. Prognostic accuracy of the sofa score, sirs criteria, and qsofa score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit. *Jama*, 317(3):290–300, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Salvatore Ruggieri and Andrea Pugnana. Things machine learning models know that they don't know. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28684–28693, 2025.

Shalom H Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology/Academic Press*, 1992.

Jennifer L Sherr, Lutz Heinemann, G Alexander Fleming, Richard M Bergenstal, Daniela Bruttomesso, Hélène Hanaire, Reinhard W Holl, John R Petrie, Anne L Peters, and Mark Evans. Automated insulin delivery: benefits, challenges, and recommendations. a consensus report of the joint diabetes technology working group of the european association for the study of diabetes and the american diabetes association. *Diabetes Care*, 45(12):3058–3074, 2022.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias W Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2010.

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, pp. 32633–32653. PMLR, 2023.

Eliza Strickland. Ibm watson, heal thyself: How ibm overpromised and underdelivered on ai health care. *IEEE spectrum*, 56(4):24–31, 2019.

Mark Tonelli. The philosophical limits of evidence-based medicine. *Academic Medicine*, 1998.

Sang Truong and Sanmi Koyejo. *Machine Learning from Human Preferences*. Stanford University, 2024.

Sebastian Tschiatschek, Eugenia Stamboliev, Timothée Schmude, Mark Coeckelbergh, and Laura Koesten. Challenging the human-in-the-loop in algorithmic decision-making, 2024. URL https://arxiv.org/abs/2405.10706.

Peter JM Van Laarhoven, Emile HL Aarts, Peter JM van Laarhoven, and Emile HL Aarts. *Simulated annealing*. Springer, 1987.

William Van Melle, Edward H Shortliffe, and Bruce G Buchanan. Emycin: A knowledge engineer's tool for constructing rule-based expert systems. *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project*, pp. 302–313, 1984.

Anouk Wouters, Gerda Croiset, and Rashmi A Kusurkar. Selection and lottery in medical school admissions: who gains and who loses? *MedEdPublish*, 7, 2018.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 10586–10613, 2024.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.

Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. Counterfactual prediction for bundle treatment. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19705–19715. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e430ad64df3de73e6be33bcb7f6d0dac-Paper.pdf.

A KEY NOTATIONS

Table 3: Key Notations

Symbol	Space	Note	Accessible?	
			Human	Algorithm
$\mathbf{x}$	$\mathcal{X} \subset \mathbb{R}^p$	Covariates; individual characteristics	Comprehensive[1]	Yes
$\mathbf{a}$	$\mathcal{A} \subset \mathbb{R}^q$	Actions	Yes	Yes
$Y(\mathbf{x}, \mathbf{a})$	$\mathbb{R}$	Measurable outcome of an individual with covariates $\mathbf{x}$ having taken action $\mathbf{a}$	No[2]	Yes
$U(\mathbf{x}, \mathbf{a})$	$\mathbb{R}$	Human-centered objective of an individual with covariates $\mathbf{x}$ having taken action $\mathbf{a}$	Yes	No
$V(\mathbf{x}, \mathbf{a})$	$\mathbb{R}$	Overall utility of an individual with covariates $\mathbf{x}$ having taken action $\mathbf{a}$	No	No
$v(\cdot, \cdot)$	$\mathbb{R}^2 \mapsto \mathbb{R}$	Utility function mapping $(Y(\mathbf{x}, \mathbf{a}), U(\mathbf{x}, \mathbf{a}))$ to $V(\mathbf{x}, \mathbf{a})$	Partial[3]	No
ϵ	$\mathbb{R}$	Amount of measurable objective sacrificed	Yes	Yes
$\pi(\mathbf{a} \mid \mathbf{x})$	$[0, 1]$	Probability of individual with covariates $\mathbf{x}$ taking action $\mathbf{a}$ under stochastic policy π	Yes	Yes

Note. 1. Comprehensive accessibility of covariates by humans: in addition to the quantifiable covariates represented by $\mathbf{x}$, humans can access the unquantifiable part of their individualities. 2. No accessibility of $Y(\mathbf{x}, \mathbf{a})$ by humans: under our setting, humans focus on the human-centered objective and do not access the measurable outcome, which is offloaded to algorithms. 3. Partial accessibility of $v(\cdot, \cdot)$ by humans: although its exact shape is not accessible, humans can make mild assumptions about it by specifying ϵ, with larger values indicating higher relative importance of the human-centered objective to the overall utility.

B EXAMPLES OF HUMAN-CENTERED DECISION-MAKING PROBLEMS

Table 4: Examples of Human-Centered Decision-Making Problems

Domain	**Key Decision** (A)	**Measurable Objective** (Y)	**Human-Centered Objective** (U)	**Overall Utility** (V)	**Illustrative Trade-Off**
Corporate Hiring	Selecting a new employee.	Candidate's quantifiable metrics (years of experience, test scores, keyword match).	Candidate's unmeasurable qualities (cultural fit, potential, team chemistry).	A productive, collaborative, long-term team member.	Rejecting a high-potential candidate (high U) because their resume lacks a specific keyword (low Y).
Urban Planning	Approving a new development project.	Hard metrics (housing density, tax revenue, traffic flow).	Qualitative experience (neighborhood character, sense of community, aesthetics).	A vibrant, equitable, and livable city.	Maximizing housing density (high Y) at the cost of destroying a beloved historic district (low U).
Product Design	Designing a software feature.	Business KPIs (conversion rate, daily active users, click-through rate).	User's subjective experience (delight, trust, intuitive flow).	A successful product with deep user loyalty.	Using a "dark pattern" to boost sign-ups (high Y) while eroding user trust (low U).
University Admissions	Selecting an incoming class.	Standardized metrics (GPA, SAT/ACT scores, class rank).	Applicant's ineffable potential (leadership, creativity, resilience).	A diverse and dynamic student body that succeeds post-graduation.	Admitting a "test-taker" with perfect scores (high Y) over a creative leader with a unique story (high U).
Financial Investing	Constructing an investment portfolio.	Financial performance indicators (ROI, alpha, Sharpe ratio).	Ethical and moral alignment (ESG principles, social impact, personal values).	A portfolio that generates wealth and provides peace of mind.	Forgoing a highly profitable but unethical investment (high Y, low U).
Film Curation	Acquiring a film for distribution.	Commercial performance (gross box office revenue).	Artistic merit and critical acclaim (story quality, creative vision).	A film that is both a commercial and cultural success.	Green-lighting a formulaic sequel (high Y potential) over an innovative indie film (high U potential).
Personal Well-being	Choosing a meal on a diet.	Nutritional data (low sugar content, low calories).	Subjective experience (perceived sweetness, palatability, satisfaction).	A healthy and satisfying food choice that supports long-term adherence.	Eating a nutritionally perfect but tasteless meal (high Y, low U), leading to later cravings and diet failure.

C Different Paradigms of Solving Human-Centered Decision-Making Problems

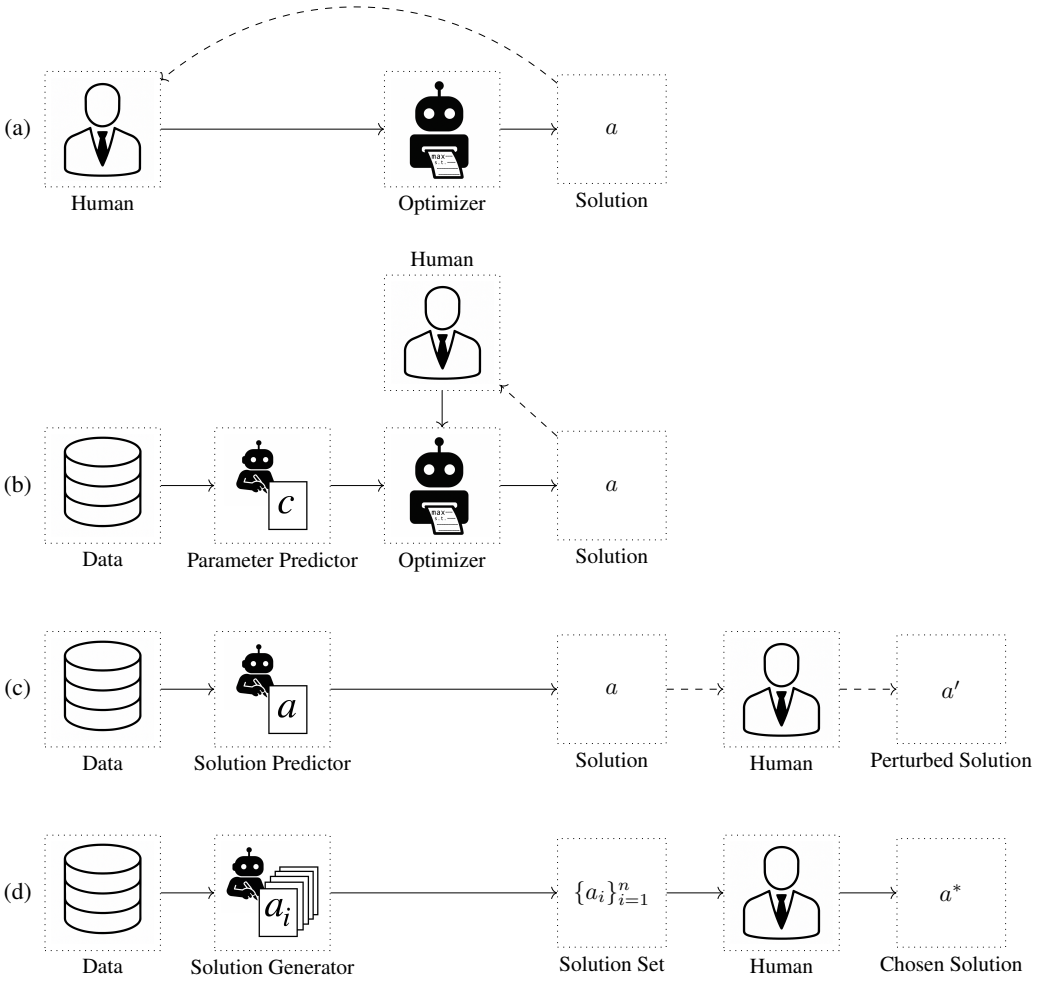


Figure 8: Illustration of Different Paradigms of Solving Human-Centered Decision-Making Problems

(a) Human-programmed optimization; (b) Data-driven optimization with human input; (c) Direct learning; (d) Generative near-optimal policy learning (`GenNOP`). Solid arrows indicate primary procedures; dashed arrows indicate secondary procedures if the solution from a primary procedure is not accepted by a human evaluator: in (a) and (b), human programmers have to readjust the optimization parameters, a hard task involving human prediction of optimizer behavior; in (c), human evaluators make localized perturbations to the solution as they have no control over predictor behavior. `GenNOP` is the only paradigm that does not involve any secondary procedure as it explicitly allows human evaluators to express their tacit knowledge by presenting them multiple solutions.

D REAL-WORLD EXAMPLES OF THE $Y_a - U_a$ RELATIONSHIP

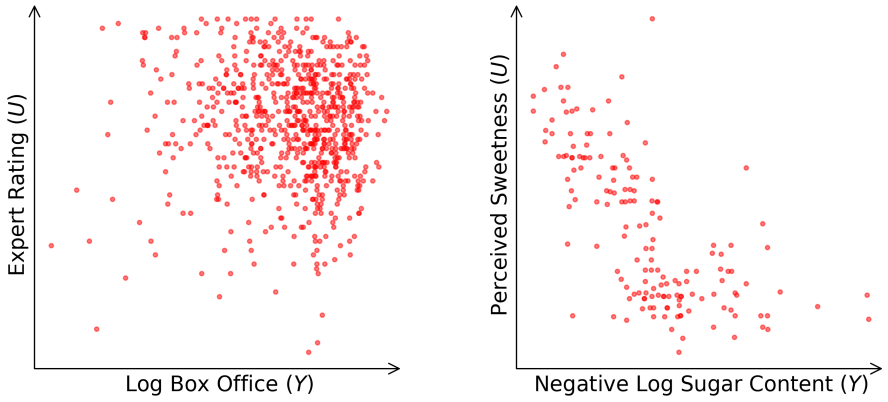


Figure 9: Real-World Examples of the $Y_a - U_a$ Relationship

(a) The left figure illustrates a scenario where an expert movie curator recommends a movie title to a general consumer: the curator believes both the commercial success and the artistic value of a movie positively contribute to the general consumer's utility. The measurable objective (Y_a) is represented by the log of the gross box office; the human-centered objective (U_a) is represented by the IMDb Metascore, a proxy for what the curator in our scenario would evaluate the artistic value of the movies. Here Y_a and U_a have a weak, positive correlation. The movie with the highest box office is not the one with the highest expert rating. (b) The right figure illustrates a scenario where a sweet-tooth consumer on a low-calorie diet makes a food choice: the consumer's utility depends on both the lack of sugar content (the measurable objective, Y_a) and the perceived sweetness (the human-centered objective, U_a) of the foods. Here Y_a and U_a have a negative correlation.

E ASSUMPTIONS AND PROOFS OF SECTION 3

E.1 ASSUMPTIONS

Assumption E.1 (Bounded support with strictly concave frontier)**.** There exist $y_- < y_+$ and a continuous, strictly concave $g : [y_-, y_+] \to \mathbb{R}$ such that

$$\mathbb{P}\big((Y,U) \in S\big) = 1, \qquad S := \{(y,u) : y \in [y_-, y_+],\ u \le g(y)\},$$

and the Pareto frontier is $\{(y, g(y)) : y \in [y_-, y_+]\}$.

Assumption E.2 (Thick near-frontier band)**.** There exist $\eta > 0$ and $c > 0$ such that (Y, U) admits a density f with

$$f(y,u) \ \ge\ c \quad \text{for all } y \in [y_-, y_+] \text{ and } g(y) - \eta \le u \le g(y).$$

Assumption E.3 (Monotone aggregator with interior V-maximizer)**.** The aggregator $v(\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}$ is continuous and strictly increasing in each argument, and with $V(y,u) = v(y,u)$ there is a unique maximizer

$$(y^\star, u^\star) \in \arg\max_{(y,u)\in S} V(y,u) \quad \text{with } u^\star = g(y^\star),\ y^\star \in (y_-, y_+).$$

Assumption E.4 (Local tradeoff bounds for $v(\cdot,\cdot)$ near $(y^\star, g(y^\star))$)**.** There exist $\epsilon_0 > 0$, $\eta_0 > 0$ and constants $L_Y, m_U > 0$ such that for all

$$y, y' \in [y^\star - \epsilon_0,\ y^\star], \quad u, u' \in [g(y) - \eta_0,\ g(y)],$$
$$v(y', u) - v(y, u) \ \le\ L_Y\,|y' - y|, \qquad v(y, u') - v(y, u) \ \ge\ m_U\,(u' - u).$$

Assumption E.5 (Local sampling restriction)**.** For some $\epsilon \in (0, \epsilon_0]$ and $\eta \in (0, \eta_0]$, we draw i.i.d. candidates $(Y^{(k)}, U^{(k)})$ satisfying

$$\mathbb{P}\Big((Y^{(k)}, U^{(k)}) \in \mathcal{N}_{\epsilon,\eta}\Big) = 1, \qquad \mathcal{N}_{\epsilon,\eta} := \{(y,u) : y \in [y^\star - \epsilon,\ y^\star],\ g(y) - \eta \le u \le g(y)\}.$$

E.2 PROOFS

Proof of Proposition 3.1. By Assumption E.2, the marginal of Y has positive density on a right–endpoint neighborhood of y_+. Hence, for any $\varepsilon > 0$,

$$\mathbb{P}\left(Y \in [y_+ - \varepsilon,\, y_+]\right) \;\geq\; c\,\varepsilon > 0,$$

and the sample maximum satisfies $Y^{(M_Y)} \to y_+$ almost surely as $N \to \infty$.

By Assumptions E.1 and E.3, the unique population maximizer of $V(y,u) = v(y,u)$ over the compact set S is $(y^\star, g(y^\star))$ with $y^\star \in (y_-, y_+)$. Uniqueness and continuity yield: for every $\rho > 0$ there exists $\kappa(\rho) > 0$ such that

$$\sup_{\substack{(y,u)\in S \\ \|(y,u)-(y^\star,g(y^\star))\|\geq\rho}} V(y,u) \;\leq\; V(y^\star, g(y^\star)) - \kappa(\rho).$$

Since, by Assumption E.2, any neighborhood of $(y^\star, g(y^\star))$ inside S has positive probability, the empirical maximizer satisfies

$$\left(Y^{(M_V)}, U^{(M_V)}\right) \;\xrightarrow{p}\; (y^\star, g(y^\star))\,.$$

Fix $\varepsilon \in \left(0, \frac{y_+ - y^\star}{3}\right)$. With probability tending to one,

$$Y^{(M_Y)} \geq y_+ - \varepsilon \qquad \text{and} \qquad Y^{(M_V)} \leq y^\star + \varepsilon < y_+ - 2\varepsilon,$$

so $Y^{(M_V)} < Y^{(M_Y)}$ and hence $M_V \neq M_Y$. Therefore $\mathbb{P}(M_Y = M_V) \to 0$. □

Proof of Proposition 3.2. Work on the event (which has probability one under absolutely continuous sampling) that there are no exact ties in the U-coordinates among the finitely many sampled points. Let i, j be two sampled indices with

$$Y^{(i)}, Y^{(j)} \in [y^\star - \epsilon,\, y^\star] \quad \text{and} \quad U^{(i)} \geq U^{(j)}.$$

By Assumption E.4, for any such pair we can write

$$\begin{aligned} V^{(i)} - V^{(j)} &= \left[v\left(Y^{(i)}, U^{(i)}\right) - v\left(Y^{(j)}, U^{(i)}\right)\right] + \left[v\left(Y^{(j)}, U^{(i)}\right) - v\left(Y^{(j)}, U^{(j)}\right)\right] \\ &\geq -\,L_Y \left|Y^{(i)} - Y^{(j)}\right| \;+\; m_U \left(U^{(i)} - U^{(j)}\right). \end{aligned}$$

Let $\Delta_U^{\min} := \min\left\{\, U^{(i)} - U^{(j)} :\; U^{(i)} > U^{(j)} \,\right\}$ over the (finite) sample; on the no-tie event, $\Delta_U^{\min} > 0$. Choose

$$\bar{\epsilon} \;:=\; \min\left\{\epsilon,\; \frac{m_U}{L_Y}\,\Delta_U^{\min}\right\}.$$

Then for any i, j with $Y^{(i)}, Y^{(j)} \in [y^\star - \bar{\epsilon},\, y^\star]$ and $U^{(i)} > U^{(j)}$ we have

$$V^{(i)} - V^{(j)} \;\geq\; -L_Y\,\bar{\epsilon} + m_U\,\Delta_U^{\min} \;\geq\; 0.$$

If $U^{(i)} = U^{(j)}$ (a null event under absolute continuity), the conclusion $V^{(i)} \geq V^{(j)}$ holds whenever $Y^{(i)} \geq Y^{(j)}$ by monotonicity of $v(\cdot,\cdot)$ in its first argument. Hence, almost surely, the stated implication holds for all pairs, and in particular any $\arg\max U$ over $\left\{k :\; Y^{(k)} \in [y^\star - \bar{\epsilon},\, y^\star]\right\}$ is also an $\arg\max V$. □

F ASSUMPTIONS IN SECTION 4

We adopt the potential outcomes framework (Rubin, 1974; Imbens & Rubin, 2015). We write Y_a as the potential outcome for the measurable objective of an individual taken action a and make the following standard assumptions:

1. *Consistency*: Provided the action is a, then Y_a is the potential outcome under action $\mathbf{a}$. Formally, $A = a$ implies $Y_a = Y$.
2. *Unconfoundedness*: $Y_a \perp\!\!\!\perp A \mid X$ for all a. This assumption implies action selection is as good as randomized given covariates.
3. *Positivity*: Every individual has a non-zero chance of taking any action in $\mathcal{A}$, namely, $p(a|x) > 0$ for all a. $p(a|x)$ denotes the GPS (Imbens, 2000; Hirano & Imbens, 2004).

G PRELIMINARIES

Diffusion Models Originally introduced by Ho et al. (2020), diffusion models have proven effective in generating high-dimensional data such as images and videos (Rombach et al., 2022; Ho et al., 2022). Diffusion models progressively corrupt data with noise and then learn to reverse this process. In our context, a neural network parametrized by θ is trained to predict the noise added to an action, similar to how these models generate images.

To generate a new action, we begin with a sample drawn from a standard Gaussian distribution. Then, using the learned reverse process, we iteratively update the noisy action. At each reverse step t, the network computes a noise prediction $\tilde{\varepsilon}_\theta(a_t, t, x)$ and updates the action as $a_{t-1} = \frac{1}{\sqrt{1-\gamma_t}}\left[a_t - \gamma_t\, \tilde{\varepsilon}_\theta(a_t, t, x)\right] + \sqrt{\gamma_t}\, z_t$, where $z_t \sim \mathcal{N}(0, I)$. After reversing all steps down to $t = 0$, the resulting a_0 is the generated policy sample, which can be conditioned on x if desired.

Training is performed by randomly selecting a time step t and corrupting the original action a_0 with Gaussian noise $a_t = \sqrt{\bar{\lambda}_t}\, a_0 + \sqrt{1-\bar{\lambda}_t}\, \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I)$, where $\bar{\lambda}_t = \prod_{s=1}^{t}(1-\gamma_s)$. The network parameters θ are optimized by minimizing the mean squared error between the actual noise ε and the predicted noise $\varepsilon_\theta(a_t, t, x)$:

$$L(\theta) = \mathbb{E}_{t, a_0, \varepsilon} \left\| \varepsilon - \varepsilon_\theta(a_t, t, x) \right\|^2 . \tag{5}$$

Max-Stable Processes Let $Y_x^{(1)}, \ldots, Y_x^{(k)}$ be a sequence of k independent copies of a stochastic process $\{Y_x : x \in \mathcal{X}\}$. Define the rescaled pointwise maximum with functions c_k and d_k as:

$$Y_x^\star = \left[\max_{i=1,\ldots,k} Y_x^{(i)} - d_k(x)\right] \Big/ c_k(x), \quad x \in \mathcal{X},$$

If there are sequences of functions $c_k(x) > 0$ and $d_k(x) \in \mathbb{R}$ such that for all $k \in \mathbb{N}$, $\{Y_x^\star\}_{x\in\mathcal{X}} \overset{d}{=} \{Y_x\}_{x\in\mathcal{X}}$, then $\{Y_x\}_{x\in\mathcal{X}}$ is a *max-stable* process. Its marginal distributions are generalized extreme-value (GEV) distributions, which can be expressed parametrically. The log-likelihood of observing $\{y_j\}$ under the GEV distribution parametrized by μ, σ, ξ is given by:

$$\ell\left(\{y^{(j)}\}; \mu, \sigma, \xi\right) = -\, k \log \sigma(l) - (1 + 1/\xi) \sum_{j=1}^{k} \log\left[1 + \xi\left(\frac{y^{(j)} - \mu}{\sigma}\right)\right] - \sum_{j=1}^{k} \left[1 + \xi\left(\frac{y^{(j)} - \mu}{\sigma}\right)\right]^{-1/\xi},$$

provided that $1 + \xi\left(\frac{y^{(j)}-\mu}{\sigma}\right) > 0, \qquad \text{for } j = 1, \ldots, k.$

H GENNOP LEARNING OBJECTIVE DERIVATION

H.1 DATA-BASED LEARNING OBJECTIVE

To derive Equation (3) from Equation (4.1), we expand the KL divergence:

$$\begin{aligned}
\min_\theta L(\theta) &= \mathbb{E}_{x\sim p(X)}[D_{\mathrm{KL}}[\pi_\epsilon^\star(\cdot|x) \,||\, \pi_\theta(\cdot|x)]] \\
&= \mathbb{E}_{x\sim p(X)}\left[\mathbb{E}_{a\sim\pi_\epsilon^\star(\cdot|x)}\left[\log \frac{\pi_\epsilon^\star(a|x)}{\pi_\theta(a|x)}\right]\right] \\
&= \mathbb{E}_{x\sim p(X)}\left[\mathbb{E}_{a\sim\pi_\epsilon^\star(\cdot|x)}\left[\log \pi_\epsilon^\star(a|x) - \log \pi_\theta(a|x)\right]\right] \\
&= \mathbb{E}_{x\sim p(X), a\sim\pi_\epsilon^\star(\cdot|x)}\left[\log \pi_\epsilon^\star(a|x) - \log \pi_\theta(a|x)\right] \\
&\overset{d}{=} \mathbb{E}_{(x,a)\sim\mathcal{D}^\star}\left[\log \pi_\epsilon^\star(a|x) - \log \pi_\theta(a|x)\right]
\end{aligned}$$

Since $\log \pi_\epsilon^\star(a|x)$ does not depend on θ, the above can be equivalently expressed as:

$$\min_\theta \mathbb{E}_{(x,a)\sim\mathcal{D}^\star}\left[-\log \pi_\theta(a|x)\right].$$

Algorithm 1 `Generative Near-Optimal Policy Learning (GenNOP)`

Require: Dataset $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^n$, neighbors k, block size b, metric d, threshold ϵ
Ensure: ϵ-optimal actions $\{a_1^*, \ldots, a_m^*\}$ for given x

Initialize parameters: GEV (ψ), VAE (ϕ, φ), diffusion model (θ).
for each $x_i \in \mathcal{D}$ **do**
  Find $k \cdot b$ nearest neighbors; partition into k blocks; compute block maxima $\{y_i^{(j)}\}$.
end for
Train the GEV model via MLE to obtain $\hat{\psi}$.
for each $(x_i, a_i, y_i) \in \mathcal{D}$ **do**
  Compute $g_\epsilon(y_i, x_i) = F_{\mathrm{GEV}}(y_i + \epsilon; \hat{\psi}(x_i))$.
  Estimate $p(a_i|x_i)$; if using the VAE ratio, recover the density as described in Appendix I.1.
end for
while not converged **do**
  Sample mini-batch $(x, a, y) \sim \mathcal{D}$, timestep t, and noise ε.
  Update θ by minimizing

$$L(\theta) = \mathbb{E}\Big[w(x, a, y; \epsilon) \cdot \|\varepsilon - \varepsilon_\theta(a_t, t, x)\|^2\Big].$$

end while
Return m actions sampled from $\pi_\theta(\cdot|x)$.

H.2 Re-Weighted Diffusion Model Learning Objective

Treat the fitted filter as fixed, and let $\ell_\theta(a, x)$ be an integrable action loss. With the same action reference measure ν as in the target policy, overlap on its support, and $0 < Z_\epsilon(x) := \int h_\epsilon(a, x)\, d\nu(a) < \infty$,

$$\begin{aligned}\mathbb{E}[w(X, A, Y; \epsilon)\ell_\theta(A, X) \mid X = x] &= \int\!\!\int p(a \mid x)p(y \mid a, x)\frac{g_\epsilon(y, x)}{p(a \mid x)}\ell_\theta(a, x)\, dy\, d\nu(a) \\ &= \int h_\epsilon(a, x)\ell_\theta(a, x)\, d\nu(a) \\ &= Z_\epsilon(x)\mathbb{E}_{A\sim q_\epsilon(\cdot|x)}[\ell_\theta(A, x)].\end{aligned}$$

For diffusion training, substitute $\ell_\theta(a, x) = \mathbb{E}_{t,\varepsilon}\|\varepsilon - \varepsilon_\theta(a_t, t, x)\|^2$. This identifies the weighted training distribution; it does not equate denoising loss with negative log-likelihood. Unnormalized weights also weight covariates by $Z_\epsilon(x)$; dividing by $Z_\epsilon(x)$ preserves the original covariate distribution.

If $Y = \mathbb{E}[Y \mid A, X]$ almost surely and $g_\epsilon(y, x) = \mathbf{1}\{y \geq y^\star(x) - \epsilon\}$, consistency and unconfoundedness give $h_\epsilon(a, x) = \mathbf{1}\{a \in \Omega_\epsilon^\star(x)\}$ and $q_\epsilon = \pi_\epsilon^\star$. With noisy outcomes, the same hard filter instead gives $h_\epsilon(a, x) = \mathbb{P}(Y_a \geq y^\star(x) - \epsilon \mid X = x)$. A fitted soft filter additionally introduces estimation and smoothing error.

I Details in Section 4

I.1 Estimating Generalized Propensity Scores via Variational Sample Weight Learning

Because the policies of our consideration are potentially high-dimensional, traditional approaches to estimating GPS in the denominator will fail. To circumvent this issue, we assume that the policies have a latent low-dimensional representation, denoted as Z, and adopt the strategy by Zou et al. (2020) that learns Z via variational autoencoder (VAE) (Kingma, 2013). With encoder and decoder networks parametrized by ϕ and φ, respectively, we maximize the evidence lower bound (ELBO): $L_{\mathrm{ELBO}} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{z\sim q_\phi(z|a_i)}\left[\ell(a_i, z; \varphi, \phi)\right]$, where $\ell(a_i, z; \varphi, \phi) = \log p_\varphi(a_i|z) + \log p(z) - \log q_\phi(z|a_i)$.

Under the latent conditional-independence assumption $A \perp\!\!\!\perp X \mid Z$, Bayes' rule gives

$$\frac{p(a \mid x)}{p(a)} = \int p(z \mid a)\frac{p(z \mid x)}{p(z)}\, dz \approx \mathbb{E}_{z\sim q_\phi(z|a)}\left[\frac{p(z \mid x)}{p(z)}\right].$$

The reciprocal estimates the stabilized ratio $p(a)/p(a \mid x)$; recovering $1/p(a \mid x)$ also requires $p(a)$. With stabilized weights $p(a)g_\epsilon(y, x)/p(a \mid x)$, the induced conditional target is proportional to $p(a)h_\epsilon(a, x)$, rather than $h_\epsilon(a, x)$. A common positive rescaling of all weights preserves the minimizer; action-dependent clipping generally changes the target. The derivation assumes integrable weights and overlap on the target support.

J Ablation Study of GEV Parameter Estimation Details

Table 5: Ablation Study of GEV Parameter Estimation (Full).

Dimensionality	Regularization	$k = 10, b = 10$	$k = 10, b = 20$	$k = 10, b = 30$	$k = 20, b = 10$	$k = 20, b = 20$	$k = 20, b = 30$	$k = 30, b = 10$	$k = 30, b = 20$	$k = 30, b = 30$	$k = 50, b = 10$	$k = 50, b = 20$	$k = 50, b = 30$
1D	0	1.10 (4.63)	13.98 (16.49)	0.52 (2.90)	−1.34 (0.16)	−0.56 (1.40)	−0.15 (3.27)	0.35 (2.81)	−1.38 (0.97)	−1.31 (0.66)	−1.17 (0.48)	−1.81 (0.17)	−1.24 (0.90)
1D	0.05	−0.84 (0.58)	1.58 (4.18)	−0.22 (1.38)	−1.48 (0.28)	3.30 (10.05)	−1.46 (0.62)	−1.48 (0.15)	−1.31 (0.30)	−1.11 (0.97)	−1.24 (0.29)	−1.54 (0.49)	−2.01 (0.17)
1D	0.1	−0.11 (2.15)	−1.62 (0.42)	2.35 (7.56)	−0.56 (1.51)	−1.45 (0.65)	−0.39 (3.04)	−1.39 (0.19)	−1.39 (0.56)	−0.51 (1.52)	−0.38 (1.88)	−1.90 (0.07)	−1.91 (0.35)
1D	0.5	−0.64 (1.17)	0.81 (5.39)	0.61 (4.43)	−1.59 (0.07)	−0.95 (1.88)	−2.10 (0.28)	−1.61 (0.06)	−2.05 (0.07)	−1.81 (0.49)	−1.47 (0.11)	−1.33 (0.77)	−2.08 (0.07)
1D	1	−1.48 (0.06)	−1.88 (0.22)	−2.06 (0.17)	−1.45 (0.07)	−1.99 (0.13)	−2.17 (0.17)	−1.40 (0.13)	−1.95 (0.09)	−2.02 (0.21)	−1.38 (0.07)	−1.77 (0.19)	−2.05 (0.07)
1D	10	−1.13 (0.13)	−1.33 (0.57)	−1.84 (0.31)	−1.19 (0.06)	−1.53 (0.17)	−1.72 (0.55)	−0.85 (0.23)	−1.63 (0.26)	−1.49 (0.65)	−0.66 (0.35)	−1.26 (0.40)	−1.82 (0.07)
2D	0	0.36 (1.61)	−0.61 (0.53)	3.09 (4.58)	0.60 (1.41)	1.07 (2.62)	−0.56 (1.02)	−0.52 (0.48)	−0.94 (0.57)	−0.74 (0.88)	−0.37 (0.61)	−0.59 (0.33)	0.29 (3.11)
2D	0.05	0.16 (1.24)	−0.92 (0.18)	1.56 (5.28)	1.47 (2.62)	−0.85 (0.45)	−0.66 (0.68)	−0.15 (1.13)	−0.87 (0.24)	−1.19 (0.18)	−0.56 (0.39)	−0.75 (0.39)	−1.13 (0.25)
2D	0.1	0.79 (1.34)	0.21 (1.24)	3.85 (6.90)	−0.69 (0.16)	−0.71 (0.51)	0.53 (2.94)	−0.25 (0.40)	−0.95 (0.32)	−0.94 (0.50)	0.54 (1.23)	−0.86 (0.34)	−0.77 (0.84)
2D	0.5	−0.53 (0.50)	−0.44 (1.47)	−1.23 (0.24)	−0.29 (0.95)	−0.26 (1.26)	−1.22 (0.23)	−0.48 (0.38)	−0.92 (0.17)	−1.19 (0.09)	1.26 (3.35)	−0.81 (0.24)	−1.01 (0.22)
2D	1	−0.64 (0.12)	−1.00 (0.26)	−0.95 (0.65)	−0.36 (0.24)	−0.66 (0.44)	−1.22 (0.12)	−0.15 (0.26)	−1.08 (0.04)	−1.05 (0.34)	10.13 (16.32)	3.06 (6.88)	−0.29 (1.56)
2D	10	−0.16 (0.34)	−0.93 (0.09)	−1.25 (0.13)	−0.33 (0.06)	−0.22 (0.24)	−0.72 (0.42)	8.66 (17.81)	−0.64 (0.32)	−0.68 (0.49)	12.41 (18.23)	4.04 (6.56)	−0.36 (0.47)

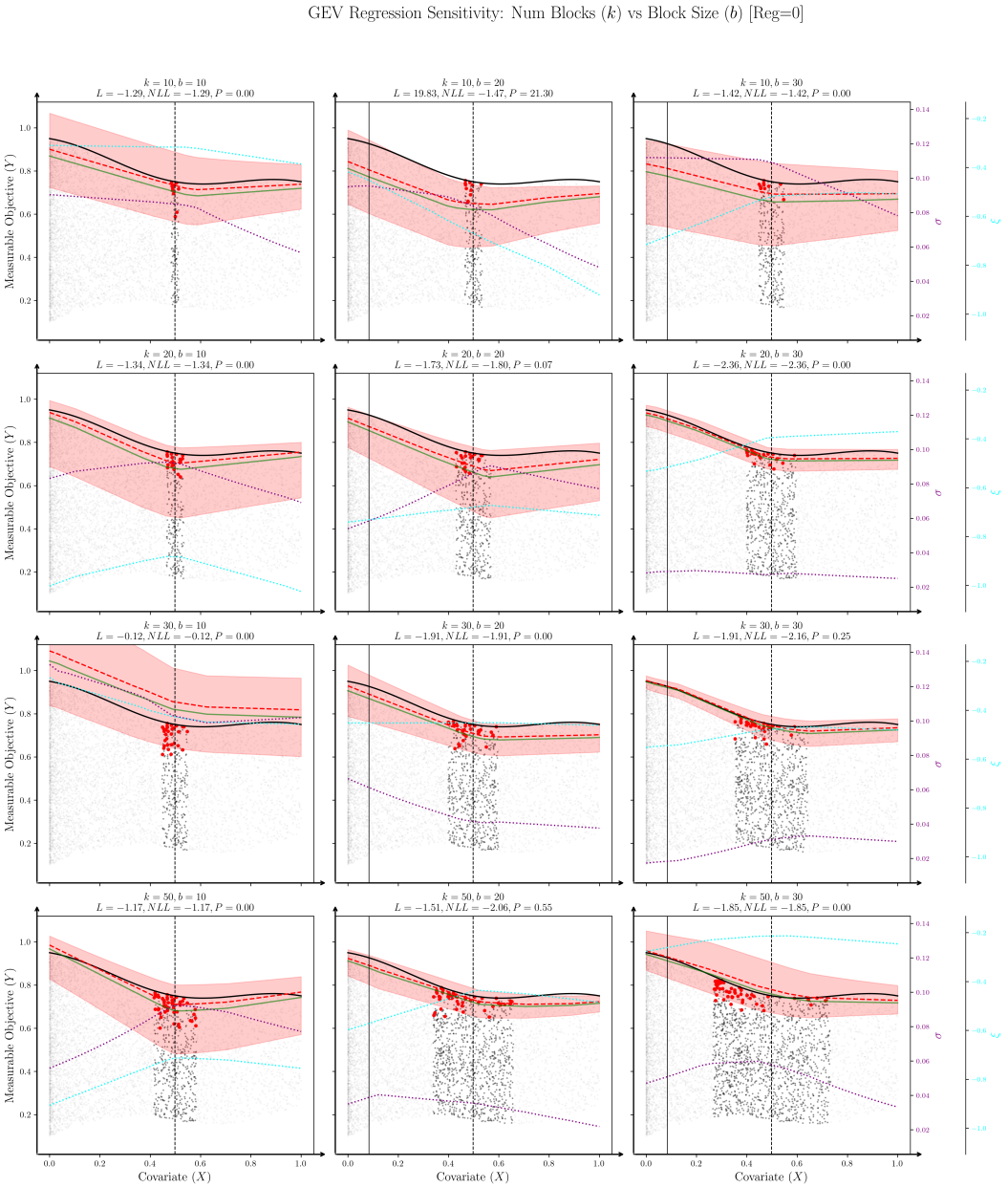


Figure 10: Regularization = 0.

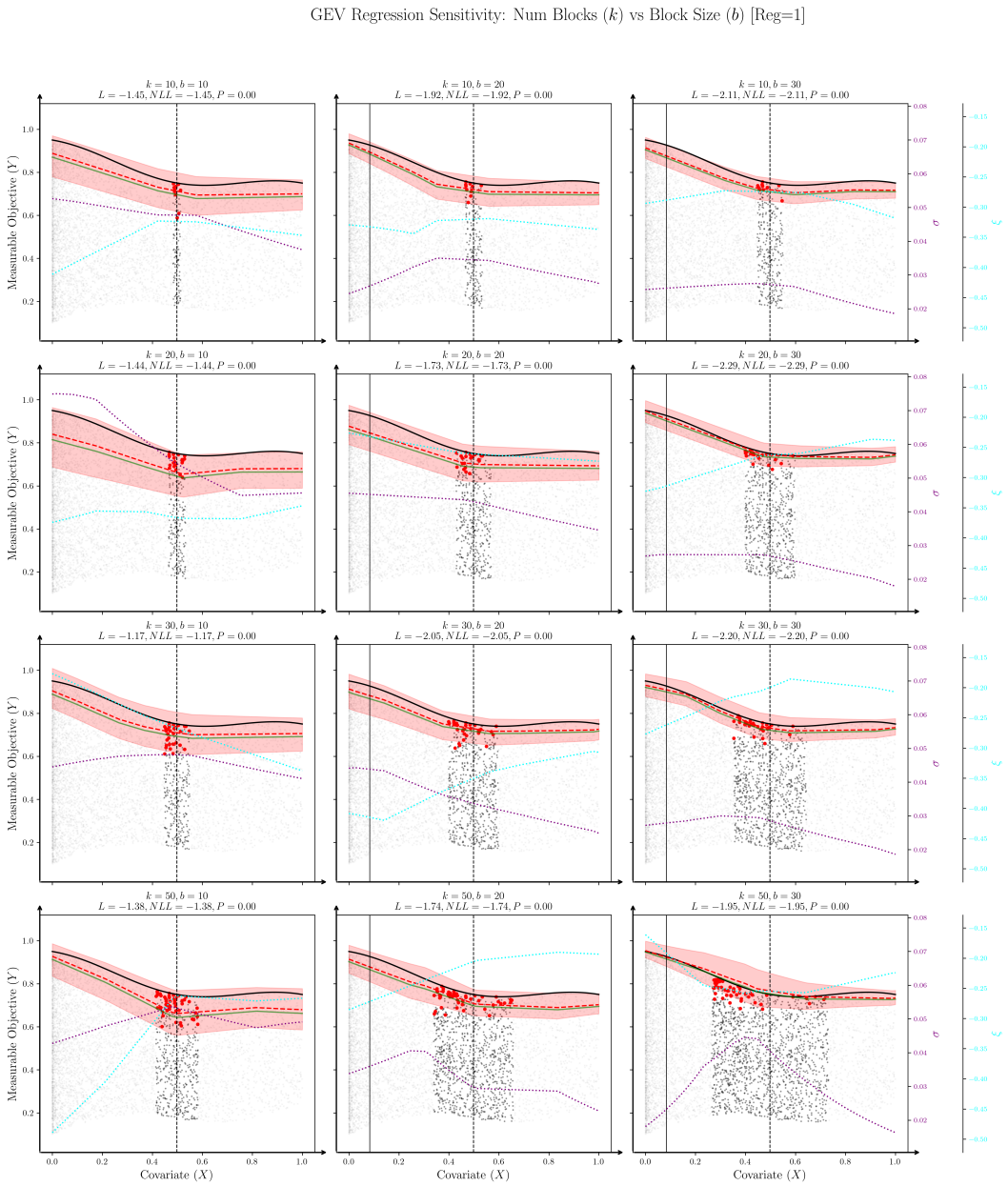


Figure 11: Regularization = 1.

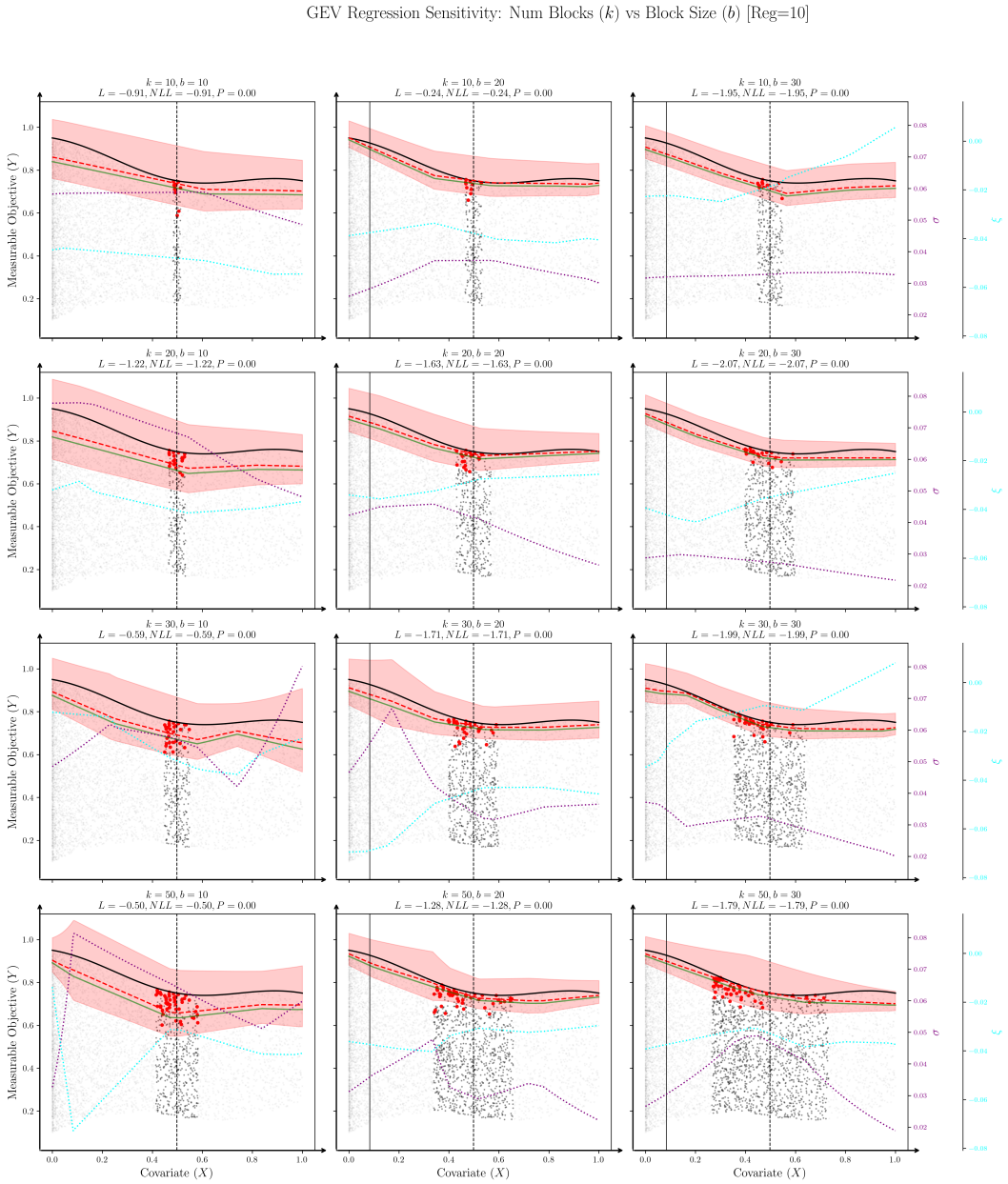


Figure 12: Regularization = 10.

K Experiment Details

K.1 Baseline Methods

Doubly-Robust Policy Learning Methods Conventional methods for policy learning take 2-step approaches: (1) Estimate the counterfactual outcomes for each individual-action pair; and (2) Learn a policy that gives the counterfactual-outcome-maximizing actions for given individuals. Notably among them is `DRPolicyForest`, a method derived from the doubly-robust policy trees (Athey & Wager, 2021), which estimates the probability that an action a is optimal for a unit characterized by covariates x. The original `DRPolicyForest` maps covariates to their most-probable optimal actions; since we aim to map covariates to a distribution of actions, we modify the mapping by introducing randomness: given covariates x, instead of always selecting the most-probable action, we sample actions according to their probabilities of being the optimal action.

Black-Box Optimization (BBO) Methods BBO methods are a class of general-purpose methods aiming at finding the input $x^\star$ that maximizes an unknown function f by observing a set of $(x, f(x))$ pairs. Notable BBO methods include the Gaussian Process Upper Confidence Bound (`GP-UCB`) algorithm (Srinivas et al., 2010) and denoising diffusion optimization models (`DDOM`) (Krishnamoorthy et al., 2023). We adapt the `GP-UCB` algorithm to our setting as we consider the *consistency* assumption and thereby view the covariate ($\mathcal{X}$)-action ($\mathcal{A}$) joint space as the input space to the data generating process of the counterfactual outcome $y : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$. We adapt the `DDOM` method by joining the covariate space $\mathcal{X}$ with the outcome space $\mathbb{R}$ as the new condition space.

`GenNOP` Ablations To validate the components of `GenNOP` empirically, we conduct the following ablation studies:

- Set the numerator of the learning objective weight, $g_\epsilon(y, x)$, to 1. This ablation effectively removes the "filter", thereby allowing the generative model to be trained on suboptimal actions.
- Set the denominator of the learning objective weight, $p(a|x)$, to 1. This ablation removes inverse probability weighting, thereby allowing selection biases to continue to exist in the training data.

K.2 Evaluation Metrics

As compared to a ground-truth ϵ-optimal policy, we would like the actions sampled from the generative policy π_θ to be both precise and comprehensive. To this end, we define the sample-based precision and recall metrics as follows:

- Precision: Expected fraction of the m generated treatments that are in $\Omega^\star_\epsilon(x)$. Generative policies with high precision is desirable because including more suboptimal actions (*i.e.*, actions with counterfactual measurable objective value below $y^\star(x) - \epsilon$) increases the risk that the human decision-maker selects an action leading to a lower V (overall utility) than the Y (measurable objective)-maximizing action.
- Recall: Expected fraction of $\Omega^\star_\epsilon(x)$ that have at least one corresponding generated action. The purpose of sacrificing Y is to afford more opportunity to maximizing the overall utility V. To this end, higher diversity among the generated actions is desirable as higher U (human-centered objective) values are more likely to appear, thereby leading to V values higher than that of the Y-maximizing action.
- Fréchet inception distance (FID), which measures the difference between the generated action distribution and the ground-truth action distribution in the semi-synthetic dataset. See Appendix K.3 for details.

K.3 Synthetic Experiment Details

Fully-Synthetic Dataset We created a synthetic dataset with 1-dimensional $\mathcal{A}, \mathcal{X}$ with the following data generating process:

$$
\begin{aligned}
X_i &\sim \mathrm{Unif}(0,1),\\
\alpha_i &= 10X_i + 1,\\
\beta_i &= 12 - \alpha_i,\\
A_i &\sim \mathrm{Beta}(\alpha_i, \beta_i),\\
Y_i &= \exp(-50(A_i - X_i)^2)\sin(9\pi A_i)
\end{aligned}
$$

Using the above relationship, we sample $10,000$ units. Under this setting, the distribution of π_ϵ is multi-modal for any $x \in X$.

Semi-Synthetic Dataset We created a semi-synthetic dataset derived from the `Fashion-MNIST` dataset. In essence, our task is to learn from the fashion preferences of different demographic profiles so as to provide them targeted recommendations of images of fashion items. Specifically, we sample $100,000$ individuals with 2-dimensional demographic profiles: age and gender. We assume there are 6 archetypes of demographic profiles based on age and gender, each having a mapping from the 10 classes of fashion items to Y values. All items in the same class have the same Y value for the same archetype. The probability that an individual of some archetype choosing one of the 10 classes of fashion items is a function of the Y values of their archetype. This is in part to ensure the *overlap* assumption is met. After the individual chooses the class, they will choose one of the $5,000$ images with equal probability from the *training* set of the `Fashion-MNIST` dataset as their self-selected treatment. For each action an individual takes, we add a small Gaussian noise with mean 0 to the Y value mapped from the archetype of that individual as the observed Y value.

We trained a classifier that maps an image to one of the 10 classes of fashion items, which is used to calculate the precision and recall metrics. In addition to the precision and recall metrics, we also report the FID: For each profile, we randomly draw 5 classes with replacement from its ϵ-optimal classes with equal probability and then draw one of the $1,000$ images from the `Fashion-MNIST` *test* set for each class we draw also with equal probability. Representing the generated and the ground-truth distributions of ϵ-optimal policies empirically both using the 5 independently sampled actions, we calculate the FID between the two empirical distributions.

End-to-End Dataset We created a synthetic dataset with 1-dimensional $\mathcal{A}$, $\mathcal{X}$ with the following data generating process:

$$
\begin{aligned}
X_i &\sim U(0.1, 0.9)\\
y_{\mathrm{width},i} &= 50 + 25\sin(2\pi X_i)\\
u_{\mathrm{width},i} &= 25 + 10\cos(3\pi X_i)\\
y_{\mathrm{center},i} &= X_i\\
u_{\mathrm{center},i} &= X_i + 0.4 + 0.15\sin(4\pi X_i)\\
Y_i(a) &= \exp(-y_{\mathrm{width},i}(a - y_{\mathrm{center},i})^2) + 0.3\exp(-40(a - (y_{\mathrm{center},i} - 0.25))^2)\\
U_i(a) &= \exp(-u_{\mathrm{width},i}(a - u_{\mathrm{center},i})^2) + 0.2\exp(-30(a - (u_{\mathrm{center},i} + 0.2))^2)\\
V_i(a) &= Y_i(a)^{0.6} \cdot U_i(a)^{0.4}
\end{aligned}
$$

K.4 Decision-Maker Perception Preferences and Capabilities

For Figure 6, we evaluate 100 equally spaced weights $\lambda_U \in [0,1]$ with $\lambda_Y = 1 - \lambda_U$ and $\lambda_V = 0$. At each weight and in each of five runs, we resample 100 batches of 100 candidate actions from each policy, choose the action with the highest perceived score in each batch, and compute regret from the median true V of the selected actions. The plotted curve is the mean of those five run-level regrets, and its band is one across-run standard deviation. The `DDOM` baseline uses 80 outcome bins. For the noise condition, a single independent multiplier drawn uniformly from $[0.8, 1.2]$ for each candidate scales its perceived Y, U, and V together when ranking actions; regret is evaluated on the unperturbed V.

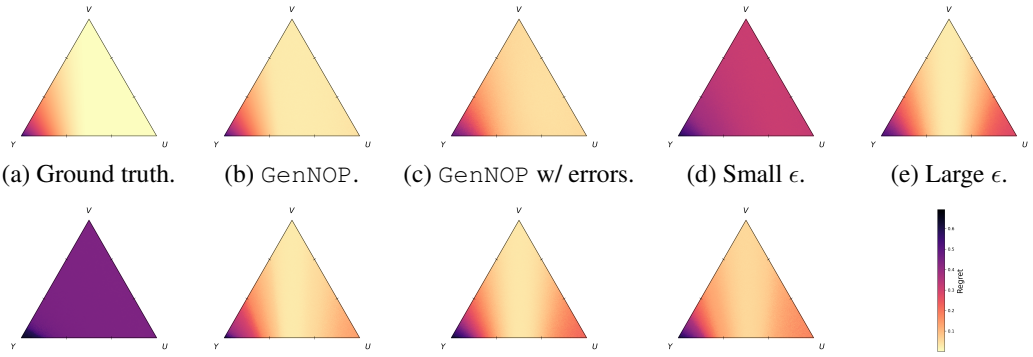
The three-weight maps below are a separate exploratory evaluation that varies λ_V as well. They use a fixed sampling seed and independently perturb perceived Y, U, and V in the noisy panel. They visualize a wider preference space but should not be read as the five-run uncertainty estimates in Figure 6.

We have insofar assumed the decision-makers implementing `GenNOP` are capable of selecting the action that maximizes the human-centered objective (U) from a set of generated actions. This has been a reasonable assumption as we can instruct the decision-makers to delegate their regard for the measurable objective (Y) to `GenNOP` and focus their perception on the human-centered objective by reflecting on their experiences, preferences, and values. However, human decision-makers are trained through both experience and formal education to approach decision-making problems with varying emphasis: for treating the same patient, their loved ones may be heavily biased towards U, a junior clinician towards Y, and a senior clinician — having practiced their specialty for decades and interacted with countless patients — may be capable of selecting the action maximizing the overall utility (V). Since both U and V are unmeasurable, we can only qualitatively assess the decision-maker perception preference. Nevertheless, we would like `GenNOP` to perform well with decision-makers under most practical settings.

To evaluate the decisions made by the human-algorithm hybrid system, we model the human decision-makers as capable of perceiving the quality of any action (Q_a) as a linear combination of their perceived Y_a, U_a, V_a values. Given a policy π, they solve the following optimization problem as their decision:

$$\arg\max_{a\sim\pi} \lambda_Y Y_a(1+\delta_Y) + \lambda_U U_a(1+\delta_U) + \lambda_V V_a(1+\delta_V),$$

where $\lambda_Y + \lambda_U + \lambda_V = 1$ and $\lambda_Y, \lambda_U, \lambda_V \geq 0$; and $\delta_Y, \delta_U, \delta_V \sim \text{Unif}(-\delta, \delta), \delta \geq 0$ are drawn independently to model the limitations in decision-maker perception capabilities. Under this human model, the ideal `GenNOP` decision-maker has $\lambda_Y = \lambda_V = 0, \lambda_U = 1, \delta = 0$, the oracle decision-maker $\lambda_Y = \lambda_U = 0, \lambda_V = 1, \delta = 0$, and the perfect optimization algorithm $\lambda_U = \lambda_V = 0, \lambda_Y = 1, \delta = 0$. Practically, most decision-makers will have $\lambda_Y, \lambda_U, \lambda_V, \delta > 0$.



(a) Ground truth. (b) `GenNOP`. (c) `GenNOP` w/ errors. (d) Small ϵ. (e) Large ϵ.

(f) `DRPF`. (g) `GP-UCB`. (h) `Perturb`. (i) `DDOM`.

Figure 13: **Regrets Under Varying Decision-Maker Perception Preferences and Capabilities.** We model the decision-makers using generated action samples as selecting the best action according to their perceived quality of the actions (Q_a), which is a linear combination of the true Y, U, V values: $Q_a = \lambda_Y Y_a + \lambda_U U_a + \lambda_V V_a$ where $\lambda_Y + \lambda_U + \lambda_V = 1$ and $\lambda_Y, \lambda_U, \lambda_V \geq 0$. We use barycentric coordinates to show the regrets under varying $\lambda_Y, \lambda_U, \lambda_V$ with different policies. For example, a decision-maker with $\lambda_Y = \lambda_V = 0, \lambda_U = 1$ is represented by the lower right vertex (U) of each plot; this decision-maker will select the action with the highest U_a value. Another decision-maker with $\lambda_Y = \lambda_V = \lambda_U = 1/3$ is represented by the barycenter of each plot; this decision-maker will select the action with the highest average value of Y, U, V. (a) shows the ground-truth ϵ-optimal policy with the best ϵ; (b)(d)(e) show the policies learned by `GenNOP` with the best, a small, and a large ϵ, respectively; (c) shows the policy learned by `GenNOP` with the best ϵ with human evaluators having an independent, uniform $\pm 20\%$ random error in their implicitly perceived Y, U, V values; and (f) – (i) show the policies given by `DRPolicyForest`, `GP-UCB`, local perturbation, and `DDOM`. All plots share the same color scale.

We assess the quality of the decisions made by the hybrid system by comparing the overall objective of the chosen decision and that of the oracle decision and calculating the regret. Using barycentric coordinates, we plot the regret against the ternary preferences $(\lambda_Y, \lambda_U, \lambda_V)$ in equilateral triangles in Figure 13. As expected, the ground-truth (a) ϵ-optimal policy with the best ϵ achieves zero regret at $(\lambda_Y, \lambda_U, \lambda_V) = (0, 1, 0)$ and trivially at $(\lambda_Y, \lambda_U, \lambda_V) = (0, 0, 1)$. Conversely, reducing the hybrid system to the algorithm-only system at $(\lambda_Y, \lambda_U, \lambda_V) = (1, 0, 0)$ yields the worst possible regret under $\delta = 0$. Zero regret is also achievable as long as λ_Y is below a threshold determined by the ratio between λ_U and λ_V. Notice that the threshold is higher when the ratio is higher. This indicates that decision-makers implementing `GenNOP` are better off *not* considering the overall objective and focusing on the human-centered objective instead when they have moderate levels of perception preference for the measurable objective that cannot be reduced in practice. We observe that (b) `GenNOP` can no longer achieve zero regret due to its non-zero — albeit diminishing — densities in the region where $Y_a < y^\star - \epsilon$. Depending on the geometry of the true ϵ-optimal region in $\mathcal{A}$, generative models underlying `GenNOP` may make varying levels of errors as they implicitly interpolate and extrapolate. This problem can be ameliorated by introducing classification model $(\mathcal{A} \mapsto \{0, 1\})$ acting as a post-generation filter that further diminishes the densities in the region where $Y_a < y^\star - \epsilon$. `GenNOP` nevertheless enjoys low regret under all but the perception preferences closest to the Y-vertex. We notice that the regret at the U-vertex is slightly higher than that slightly farther away from the U-vertex due to the generative model errors. Fortunately, real-world decision-makers can rarely operate at the U-vertex even when instructed to do so; they likely operate with the ideal perception preferences when generative model errors are considered. In this exploratory map, `GenNOP` has lower regret than `DDOM` near the U-vertex; Figure 6 gives the revised repeated-run comparison along the Y–U axis.

The performance of `GenNOP` is dependent on the choice of ϵ. As $\epsilon \to 0$, `GenNOP` is reduced to an optimization algorithm; and as $\epsilon \to \infty$, `GenNOP` is reduced to an unconditional generative model trained from all observational actions. Since decision-makers do not have access to the value of the best $\epsilon \approx 0.2$, we repeat the analysis for a small $\epsilon = 0.05$ and a large $\epsilon = 0.8$ in (d) and (e) to probe the behavior of `GenNOP` under misspecified ϵ's. Zero regret is not attainable in either case. An ϵ too small deprives the human evaluators of the opportunity of ever seeing actions leading to higher V-values, thereby resulting in higher regret regardless of human evaluators. An ϵ too large, on the other hand, passes much of the burden from algorithms to human evaluators, as the latter operating at or close to the U-vertex no longer significantly improves the V-value beyond that achieved by the global Y-maximizing action. We recommend implementers of `GenNOP` to err on the larger side when unsure about the choice of ϵ to prevent catastrophic regret across the board in cases similar to (d).

Human evaluators are limited by their perception *capabilities* regardless of their perception *preferences*. An evaluator may attempt at executing a perception preference faithfully and consistently, yet they may not choose the action maximizing $Q_a(\lambda_Y, \lambda_U, \lambda_V)$ as they have access to only a noisy version of $\{Q_a\}_{a \sim \pi}$. We set $\delta = 0.2$ in (c) so that the perceived quality of each action is multiplied by an independent factor uniformly drawn from $[0.8, 1.2]$; the order statistics, especially the maximum, are likely perturbed by this random factor. We observe that `GenNOP` is reasonably robust under this perturbation, as the perturbed case largely resembles the unperturbed case with only a slight increase in minimum regret and in the size of the undesirable region near the Y-vertex.

K.5 Real Experiment Details

Human experts solve utility-maximization problems involving measurable and human-centered objectives implicitly in their decision-making. Without access to ground-truth ϵ-optimal policies, we can nonetheless treat the decisions made by expert-level human decision-makers as a proxy for the ground truth. To this end, we study medications prescribed to patients admitted to intensive care units (ICUs) to showcase the alignment between the recommendations from `GenNOP` and the prescriptions given by critical care practitioners. The goal of the decision problem facing the latter can be characterized as significant in both measurable and human-centered objectives: critical care practitioners are tasked with stabilizing patients as their immediate goal, while they must take a holistic approach towards caregiving with their expertise and intuition. Additionally, there are vast pools of past patient records available electronically for us to apply `GenNOP`. Among them is Medical Information Mart for Intensive Care (MIMIC)-IV (Johnson et al., 2023), a large de-identified dataset available for credentialed access. From its `icu` module, we extract 2 datasets explained as follows:

`mimic-icu-sepsis` Sepsis is a life-threatening condition often treated in ICUs and is a major cause of mortality worldwide. It requires timely and appropriate antibiotic therapy to improve outcomes. The challenge lies in selecting the correct antibiotic regimen due to the diverse range of potential pathogens, including bacteria, viruses, and fungi (Gauer et al., 2020). Recent advancements in machine learning have shown promise in predicting sepsis outcomes and optimizing treatment strategies (Raghu et al., 2017; Moor et al., 2021). Prompt and effective intervention, supported by machine learning models and clinical tools, can significantly enhance patient recovery and reduce the burden of sepsis on healthcare systems.

In our study, we selected $2,783$ patients whose diagnoses fell under the sepsis category using a set of International Classification of Diseases Version 9/10 (ICD-9/10) codes for sepsis-related diseases, who were admitted to the ICU, and who were eventually discharged from the hospital having survived, by joining tables and filtering columns. We regard individual age, sex, and SOFA score as covariates (X). SOFA (Sequential Organ Failure Assessment) scores are a valuable tool in critical care for assessing organ dysfunction and predicting patient outcomes in sepsis patients (Raith et al., 2017). We use the recorded dosages of five antibiotics (Vancomycin, Meropenem, Piperacillin/Tazobactam, Levofloxacin, and Azithromycin) as the action vector. The total length of stay (LOS) in the hospital is a significant treatment indicator for sepsis patients, as it reflects the severity of the illness, the effectiveness of the treatment, and the patient's response to therapy. Prolonged LOS is often associated with increased hospital costs, higher mortality rates, and a greater likelihood of long-term complications (PL et al., 2024). We thus consider negative hospital LOS (to conform to the maximization problem of `GenNOP`) as the measurable objective (Y).

`mimic-icu-cardio` Cardiovascular diseases are a leading cause of morbidity and mortality worldwide and often necessitate intensive care for appropriate management. In our study, we selected $4,219$ patients whose diagnoses fell under the cardiovascular disease category, who were admitted to the ICU, and who were eventually discharged from the hospital having survived, by joining tables and filtering columns. We condition generation on age and sex; lactate is used to define the held-out slice. The treatments of interest are the dosages of four commonly administered cardiovascular medications: norepinephrine (a vasopressor), heparin (an anticoagulant), furosemide (a diuretic), and metoprolol (a beta-blocker). As with the sepsis dataset, the total hospital length of stay (LOS) is used as the outcome variable, reflecting both the severity of cardiac conditions and response to treatment.

Evaluation Since we have no access to the counterfactual outcomes under actions other than those recorded in the datasets, we cannot calculate those metrics reported for synthetic and semi-synthetic datasets which compare generated actions with ground-truth ϵ-optimal actions. For the displayed comparison, we select recorded actions whose hospital LOS is below the mean LOS within the held-out slice. We reserve a small range of covariates solely for this purpose and exclude all actions within that range from the training set to prevent leakage that inflates evaluation metrics.

The archived clinical notebooks use a conditional Gaussian-process generator with GEV-derived observation weights and $\epsilon = 15$ in units of negative hospital LOS (days). This is a fixed experimental setting. They fit the generator on $\log(1 + a)$ of the full medication vectors, generate ten samples per held-out covariate vector, invert the transformation, and set negative dosages to zero. A separate two-dimensional VAE embeds actions for visualization after generation. Gaussian KDE displays the saved generated-action embeddings.

For local perturbation, the cardiovascular notebook uses recorded actions with LOS of one day and independent Gaussian noise with standard deviation 0.1 times each medication's training standard deviation. The sepsis notebook uses the first recorded action with LOS of two days and noise with standard deviation equal to each medication's training standard deviation. Negative perturbed dosages are set to zero.

The displayed observations use the held-out LOS rule above; they are not selected by a separately fitted model of $y^\star(x)$. The comparison assesses the distributions of generated and recorded actions in this slice.