The Principle of Proxy Instability: Why Causal Inference Must Move Beyond Static Feature Maps

Jake Peace
mail@jakepeace.me
https://github.com/hotprotato/autoite

November 2025

Contents

Abstract

A fundamental, yet often unstated, limitation in causal inference is the **instability of proxies**: the mapping between observed features (X) and the latent causal state (U) is rarely invariant across time or domains. Because standard methods (e.g., Causal Forests) lock in static feature-splits, they fail when the observable proxy relationship shifts or is incomplete. Moreover, I argue that the prevailing penalty applied to divergence from population averages in medicine, education, and psychology actively causes harm—by treating individuals according to group means rather than latent state, we systematically mistreat those who deviate from the average.

This paper is not a silver bullet. It is a blueprint for a silver bullet. I argue that causal inference must shift from static feature engineering to dynamic primitive discovery, and provide a proof-of-concept demonstrating this is achievable. I present **AutoITE** (Automated Individual Treatment Effect Estimation), which replaces static feature maps with **dynamic residual-based discovery** of latent environments. The specific implementation choices are illustrative—the community is invited to investigate adaptive parameters, alternative architectures, and domain-specific calibrations.

Key findings: (1) Residual-based matching achieves $r = -0.94$ correlation with hidden confounders where Causal Forests achieve $r = 0.00$; (2) With 15% triage, AutoITE achieves MAE of 0.042 compared to Causal Forest's 0.230—AutoITE's error is only 18% of theirs; (3) The method is robust to irrelevant features (200 noise features cause only 12% MAE increase), and unlike traditional causal inference, **more information is always better**—practitioners can safely include all available features; (4) Feature sparsity is acceptable as long as at least one feature strongly relates to latent environments; (5) Diagnostic metrics for unknown unknowns are limited but valuable for **monitoring emergence** in economics and psychology. I also establish fundamental limits: interaction-only confounders are irreducibly undetectable, and optimal k requires meta-analysis across sample sizes (likely dependent primarily on n, since Ridge produces well-behaved residuals).

This work opens directions for future research: adaptive parameter selection, methods for determining feature importance with respect to U, T, and Y jointly, and domain-specific calibration of diagnostic thresholds. Triaged cases currently require expert review, but the community may develop additional modeling approaches to further process uncertain samples.

Keywords: Causal Inference, Heterogeneous Treatment Effects, Latent Confounders, Proxy Instability, Residual-Based Learning, Identifiability Limits

1 Introduction: The Static Proxy Fallacy

1.1 The Problem

Treatment effect heterogeneity—the reality that identical treatments produce different outcomes for different individuals—is central to personalized medicine, education, and policy. Current state-of-the-art methods (Causal Forests [Wager and Athey, 2018], X-Learners [Künzel et al., 2019], Double Machine Learning [Chernozhukov et al., 2018]) estimate **Conditional Average Treatment Effects (CATE)**:

$$\tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X] \tag{1}$$

where $Y(1)$ and $Y(0)$ are potential outcomes under treatment and control, and X represents observed covariates.

These methods share a common assumption: **all causally relevant heterogeneity is explainable by observed features** X. They partition the covariate space (via splits, weights, or kernels) and assume that individuals within the same partition share similar treatment effects.

1.2 The Philosophical Failure

I argue that the failure of current SOTA models is not merely computational, but **philosophical**. They assume that once a subgroup G is identified from X, its causal properties are fixed. I posit that G is merely a **temporary container** for the true driver: the **latent environment** U.

Consider the following:

- In 1990, "high income" proxied for "status-seeking behavior" in consumer studies. By 2025, this relationship has inverted in many demographics.
- Pre-treatment glucose levels may proxy for "metabolic dysfunction" in one population, but "athletic carbohydrate loading" in another.
- "High test scores" may indicate "intrinsic motivation" for some students, but "parental pressure" for others.

The correlation between X (observable features) and U (latent causal state) is **unstable**. Therefore, models that lock in static feature-to-subgroup mappings fail when:

1. The proxy relationship shifts over time (distribution shift)
2. The proxy relationship is incomplete (unobserved confounding)
3. The proxy relationship is orthogonal ($U \perp X$)

1.3 The Case for Just-in-Time Discovery

If the mapping $X \to U$ is non-stationary, one cannot learn a global function that reliably recovers U from X. Instead, one must perform **Just-in-Time (JIT)** discovery: at prediction time, using instance-specific information—not just the static feature vector X_i, but the **dynamic residual pattern** R_i—to identify the latent state U_i.

This is the core insight of AutoITE: **residuals are more stable proxies than features**. While the relationship "Income → Status" may shift, the relationship "Baseline residual → Latent metabolic state" is grounded in a fundamental coupling: latent environments that affect treatment response *also* affect baseline outcomes, leaving a consistent fingerprint.

1.4 Contributions

This paper is not a silver bullet—it is a blueprint for a silver bullet. I argue that causal inference must shift from static feature engineering to dynamic primitive discovery, and provide a proof-of-concept demonstrating this is achievable. The specific parameter choices (e.g., $k = 1000$, Ridge regularization) are illustrative rather than definitive—the community is invited to investigate adaptive parameters, alternative architectures, and methods for determining feature importance with respect to U, T, and Y.

1. **Philosophical (Primary)**: I formalize the **Principle of Proxy Instability** and demonstrate that residual-based discovery of latent environments is both possible and practical. This represents a paradigm shift from "who looks like you" (feature similarity) to "who behaves like you" (residual similarity). Crucially, I argue that the penalty applied to divergence from population averages in medicine, education, and psychology actively causes harm: by treating individuals according to group means rather than their latent state, we systematically mistreat those whose true response differs from the average. AutoITE provides a framework for accounting for both latent environments and individual treatment effects.

2. **Proof of Concept**: I present AutoITE, a four-stage architecture demonstrating that baseline residuals can serve as dynamic pointers to latent state. Key empirical results:
    - $r = -0.94$ correlation with hidden U where Causal Forest achieves $r = 0.00$
    - With 15% triage, AutoITE achieves MAE of 0.042 vs Causal Forest's 0.230—AutoITE's error is only 18% of theirs (an 82% reduction)
    - 97% recovery of latent subgroups via bimodality detection when confounders affect baseline
    - 356 fewer predicted deaths compared to when confounders are undetectable (23 vs 379)—a 16× reduction when baseline coupling enables detection
3. **Quantifying Unknown Unknowns**: I introduce diagnostic metrics (bimodality, UHI) that quantify hidden structure in baseline residuals. These metrics are:
    - **Limited**: They detect only confounders that affect baseline outcomes, not interaction-only confounders
    - **Domain-specific**: Interpretation requires calibration to the specific application domain
    - **Valuable for monitoring**: Particularly in economics and psychology, where tracking metric shifts over time can detect emergence of new latent factors or structural changes in the data-generating process
4. **Practical Safeguards**: In absence of perfect unknown-unknown detection:
    - **Feature robustness**: Adding 200 irrelevant features causes only 12% MAE increase. Unlike traditional causal inference where more features can hurt, **more information is always better**—practitioners can safely include all available features ("better safe than sorry")
    - **Sparsity tolerance**: Feature sparsity is acceptable as long as at least one feature strongly relates to latent environments
    - **Triage mechanism**: Deferring 15% of uncertain cases achieves MAE of 0.042 at 85% coverage; triaged cases currently require expert review, but the community may develop additional modeling approaches to further process uncertain samples
    - **Ongoing monitoring**: Diagnostic metrics enable drift detection for triggered investigation
5. **Fundamental Limits and Future Work**: I establish irreducible boundaries and open questions:
    - Interaction-only confounders are **fundamentally undetectable**—no function of observables can detect them
    - All methods share max error $\sim$2.5—an identifiability barrier, not algorithm limitation
    - Optimal k requires meta-analysis: likely dependent primarily on sample size n (since Ridge produces well-behaved residuals), not feature count or complexity
    - The community is invited to investigate adaptive parameters and methods for determining feature importance with respect to U, T, and Y jointly

Invitation to the Community

This work opens several directions for future research:

- **Adaptive parameters**: How should k, regularization strength, and model complexity adapt to data characteristics?
- **Feature importance**: How can one determine which features are important with respect to U (latent state), T (treatment), and Y (outcome) jointly?
- **Architecture variants**: Can neural approaches, Bayesian methods, or ensemble techniques improve upon the Ridge-based JIT framework?
- **Domain-specific calibration**: How should diagnostic thresholds be set for different application domains?

The remainder of this paper is organized as follows: Section 2 formalizes the Principle of Proxy Instability and presents the AutoITE methodology. Section 3 describes three validation experiments. Section 4 presents results demonstrating orthogonality recovery, safety lift, completeness detection, Simpson's paradox analysis, and the fundamental identifiability limit. Section 5 discusses implications for the field and calls for a shift toward Primitive Discovery. Section 6 concludes.

2 Methodology: From Static Features to Dynamic Primitives

2.1 The Axiom of Latent Context

I begin by formalizing the core worldview that motivates AutoITE.

Definition 1 (Latent Environment)**.** *A **latent environment** U is an unobserved variable that causally influences both baseline outcomes and treatment effects. Formally, for individual i:*

$$Y_i = f(X_i, T_i, U_i, \epsilon_i) \tag{2}$$

where ϵ_i is random noise independent of (X, T, U).

***Critical clarification**: A latent environment is not a physical or administrative label such as "Hospital A" or "Hospital B." Two patients from different hospitals may share the same latent environment (e.g., similar metabolic dysfunction), while two patients from the same hospital may have entirely different latent environments. The latent environment captures the individual's underlying causal state—their genetic predisposition, physiological subtype, or behavioral pattern—not their membership in an observable group. This is precisely why feature-based methods fail: they condition on observable labels when the true driver is the latent state that cuts across those labels.*

Assumption 1 (The Axiom of Latent Context)**.** *The **true** Individual Treatment Effect (ITE) is a function of both observed covariates X and latent environment U:*

$$\tau(i) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i, U_i] = g(T_i, U_i) \tag{3}$$

Assumption 2 (Proxy Instability)**.** *The mapping between observed features X and latent environment U is **non-stationary**:*

$$P(U_i \mid X_i) \neq P(U_j \mid X_j) \quad \text{for } i \sim \mathcal{D}_{train},\ j \sim \mathcal{D}_{test} \tag{4}$$

where $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ may differ by time, domain, or population.

Implication: Because the proxy mapping is unstable, one cannot learn a fixed function $h : \mathcal{X} \to \mathcal{U}$ that generalizes. Standard CATE methods, which condition only on X, will fail under proxy instability.

2.2 Dual Heterogeneity Decomposition

I decompose treatment effect heterogeneity into two components:

$$\tau(X, U) = \tau_{\text{obs}}(X) + \tau_{\text{lat}}(U) \tag{5}$$

where:

- $\tau_{\text{obs}}(X)$: Heterogeneity explainable by observed features (e.g., age, gender)
- $\tau_{\text{lat}}(U)$: Heterogeneity driven by latent environments (e.g., motivation, frailty, genetic risk)

Standard CATE methods estimate $\tau_{\text{obs}}(X)$ and average over $\tau_{\text{lat}}(U)$. AutoITE seeks to recover both components by using residuals as dynamic proxies for U.

2.3 The Baseline Coupling Assumption

The key enabling assumption for AutoITE is **baseline coupling**: latent environments that affect treatment response must also affect baseline outcomes.

Assumption 3 (Baseline Coupling)**.** *Latent environment U affects baseline outcomes Y_0:*

$$Y_0 = f(X) + g(U) + \epsilon \tag{6}$$

where $g(U)$ is a non-constant function. This implies that baseline residuals $R = Y_0 - \hat{f}(X)$ capture information about U.

Intuition: If high genetic risk makes you both (a) more likely to have elevated baseline biomarkers and (b) more susceptible to treatment side effects, then your baseline residual (deviation from expected biomarkers given demographics) is a fingerprint of your genetic risk.

When Coupling Fails: If U affects *only* treatment effect but not baseline ("interaction-only confounders"), AutoITE cannot recover U from baseline residuals. I address this limitation in Section 5.

2.4 The AutoITE Architecture

AutoITE operates in four stages:

2.4.1 Stage 1: Global Anchor (The Naive Prior)

Train a deliberately **simple** Ridge regression on observed features:

$$\hat{Y}_{\text{global}} = \arg\min_{f} \sum_{i=1}^{n} (Y_i - f(X_i, T_i))^2 + \alpha_{\text{global}} \|f\|^2 \tag{7}$$

Why naive? A complex non-linear model (e.g., Gradient Boosting) would explain away all variance, leaving no signal in residuals. A simple linear model with $\alpha_{\text{global}} = 1.0$ captures obvious trends (age, sex effects) and **forces the latent signal into the residual**.

2.4.2 Stage 2: Residual Embedding (The Latent Fingerprint)

Compute baseline residuals as dynamic proxies for U:

$$R_i = Y_{0,i} - \hat{f}_{\text{global}}(X_i, T_i = 0) \tag{8}$$

Under Assumption 3, $R_i \approx g(U_i) + \epsilon_i$. Individuals with similar residual patterns $R_i \approx R_j$ likely share similar latent states $U_i \approx U_j$, **even if their feature vectors** $X_i \neq X_j$ **differ**.

This is the key distinction from feature-based methods: clustering by **behavior** (how you deviate from expectation) rather than **appearance** (what your demographics are).

2.4.3 Stage 3: Just-in-Time Local Learning (The Last Meter)

For each test individual i, find k nearest neighbors in **residual space**:

$$\mathcal{N}_i = \{j : j \in \text{top-}k \text{ by } |R_j - R_i|\} \tag{9}$$

Train a local Ridge regression on this neighborhood:

$$\hat{Y}^{(i)}_{\text{local}} = \arg\min_f \sum_{j \in \mathcal{N}_i} (Y_j - f(X_j, T_j))^2 + \alpha_{\text{local}} \|f\|^2 \tag{10}$$

with $\alpha_{\text{local}} = 0.01$ (much smaller than global) to allow fitting strong local effects (including negative effects that contradict the global prior).

Metaphor: The global model flies you to the right neighborhood (based on demographics). The local model lands on the specific house (based on residual fingerprint).

2.4.4 Stage 4: Density Fusion (Adaptive Weighting)

Blend global and local predictions:

$$\hat{\tau}_i = \lambda_i \cdot \hat{\tau}^{(i)}_{\text{local}} + (1 - \lambda_i) \cdot \hat{\tau}_{\text{global}} \tag{11}$$

where $\lambda_i = \frac{1}{1+\text{median}(d_{ij})}$ and d_{ij} are distances to neighbors in residual space. Individuals with **dense** neighborhoods (many similar residuals) get high λ (trust local model). Individuals in sparse regions get low λ (fall back to global prior).

2.5 Explicit Boosting and The Last Meter

I define AutoITE's architecture as **Explicit Boosting**. Unlike Gradient Boosting, which iteratively fits residuals to minimize loss through implicit corrections, AutoITE performs a **single, semantically grounded correction step**:

- **Global Model (The Prior)**: Establishes the population-level rule: "On average, treatment helps."
- **Local Model (The Posterior Update)**: Provides a context-specific correction: "But for your latent group (negative residuals), treatment harms."

The global model provides **The Rule**, the local model identifies **The Exception**. This approximates an **N-of-1 trial**: for each individual, identifying their latent cohort (via residuals) and estimating treatment effects within that micro-environment, achieving individualization without requiring a separate trial for each person.

The Last Meter: After feature-based methods get you to the right neighborhood, AutoITE uses residuals to deliver predictions to the specific individual—the final meter of precision that separates population averages from true personalization.

2.6 Algorithm

Algorithm 1 AutoITE: Automated Individual Treatment Effect Estimation

Require: Training data $(X_{\text{train}}, T_{\text{train}}, Y_{\text{train}}, Y_{\text{pre,train}})$, test data $(X_{\text{test}}, Y_{\text{pre,test}})$, hyperparameters $(\alpha_{\text{global}}, \alpha_{\text{local}}, k)$

Ensure: Predicted treatment effects $\hat{\tau}_{\text{test}}$

Stage 1: Global Anchor
Train Ridge(α_{global}) on $(X_{\text{train}}, T_{\text{train}}, Y_{\text{train}}) \rightarrow \hat{f}_{\text{global}}$
Stage 2: Residual Embedding
Compute baseline residuals: $R_{\text{train}} = Y_{\text{pre,train}} - \hat{f}_{\text{global}}(X_{\text{train}}, T = 0)$
Compute test residuals: $R_{\text{test}} = Y_{\text{pre,test}} - \hat{f}_{\text{global}}(X_{\text{test}}, T = 0)$
Scale residuals using RobustScaler (preserves outliers)
Stage 3: Just-in-Time Local Learning
for each test individual i **do**
    Find k nearest neighbors $\mathcal{N}_i$ in residual space: $\{j : |R_j - R_i| \text{ smallest}\}$
    Train local Ridge(α_{local}) on $\{(X_j, T_j, Y_j) : j \in \mathcal{N}_i\} \rightarrow \hat{f}^{(i)}_{\text{local}}$
    Compute density weight: $\lambda_i = 1/(1 + \text{median}(|R_j - R_i|))$
end for
Stage 4: Density Fusion
for each test individual i **do**
    $\hat{\tau}_{\text{global}} = \hat{f}_{\text{global}}(X_i, T = 1) - \hat{f}_{\text{global}}(X_i, T = 0)$
    $\hat{\tau}^{(i)}_{\text{local}} = \hat{f}^{(i)}_{\text{local}}(X_i, T = 1) - \hat{f}^{(i)}_{\text{local}}(X_i, T = 0)$
    $\hat{\tau}_i = \lambda_i \cdot \hat{\tau}^{(i)}_{\text{local}} + (1 - \lambda_i) \cdot \hat{\tau}_{\text{global}}$
end for
return $\hat{\tau}_{\text{test}}$

2.7 Hyperparameters and Implementation

AutoITE uses the following hyperparameters (validated across all experiments):

- $\alpha_{\text{global}} = 1.0$: Standard Ridge regularization for global model (prevents overfitting while preserving latent signal)
- $\alpha_{\text{local}} = 1.0$: Regularization for local model scaled with neighborhood size
- $k = 1000$: Neighborhood size (optimal for $n = 10{,}000$; see Section 2.7.1)
- Scaler: StandardScaler (z-score normalization; equivalent to RobustScaler for well-behaved residuals)

2.7.1 The k Parameter: Sample Size Dependency

A critical finding is that optimal k scales with sample size. For $n = 10{,}000$:

k	MAE	Q99	Max Error	MAE Reduction
20 (original)	0.274	2.26	2.5	–
200	0.120	1.88	2.5	56%
500	0.102	1.90	2.5	63%
1000	**0.095**	**1.73**	2.5	**65%**
1500	0.107	1.71	2.5	61%
2000	0.114	1.71	2.5	58%

Why large k works: The residual-based neighbor matching relies on Ridge regression producing well-behaved, approximately Gaussian residuals. With sufficient neighbors, the local model achieves stable estimation of treatment effects within latent groups. Small k creates hard boundaries between environments; large k implements **soft environment membership**—patients draw from mixed neighborhoods, and the local model learns a probability-weighted blend of treatment effects.

Meta-analysis needed: The optimal k likely depends on sample size as the primary driver. With $n = 10{,}000$, $k = 1000$ (10% of data) is optimal. For smaller datasets, proportionally smaller k may be appropriate. Future work should establish the relationship $k^* = f(n, \sigma_\epsilon^2)$ where σ_ϵ^2 is residual variance.

Implementation available at: `https://github.com/hotprotato/autoite`

3 Experiments: Validating the Worldview

I design three experiments to validate the Principle of Proxy Instability and demonstrate AutoITE's ability to recover latent heterogeneity when static feature maps fail.

3.1 Experiment 1: Orthogonality Recovery (The Mechanism)

Objective: Prove that when $U \perp X$ (perfect proxy instability), feature-based methods achieve $r = 0$ correlation with latent heterogeneity, while AutoITE recovers $r \approx -1$ via residuals.

Data Generation: Synthetic trial with $N = 10{,}000$ patients:

- Features: $X \sim \mathcal{N}(0, I_5)$ (5 independent covariates)
- Latent environment: $U \sim$ Bernoulli(0.5) (binary hidden genetic marker)
- **Orthogonality**: $U \perp X$ (features provide zero information about latent state)
- Treatment: $T \sim$ Bernoulli(0.5) (randomized)
- Baseline outcome: $Y_{\text{pre}} = X\beta - 2U + \epsilon$ (coupling: U affects baseline)
- Treatment effect: $\tau(U) = \begin{cases} +1.0 & \text{if } U = 0 \\ -2.0 & \text{if } U = 1 \end{cases}$ (strong negative effect for $U = 1$)
- Observed outcome: $Y = Y_{\text{pre}} + T \cdot \tau(U) + \epsilon$

Metrics:

- Correlation between $\hat{\tau}$ and true U

- Detection rate: Percentage of high-risk ($U = 1$) patients correctly predicted to have negative treatment effect
- MAE between $\hat{\tau}$ and true τ

Baselines:

- Global Ridge (no baseline information)
- Causal Forest (EconML CausalForestDML)
- AutoITE (with baseline residuals)

3.2 Experiment 2: Safety Lift (The Stakes)

Objective: Demonstrate that AutoITE's ability to recover latent heterogeneity translates to **real-world safety benefits** in a high-stakes medical trial simulation.

Scenario: Hidden toxicity trial where treatment is beneficial on average but lethal for a latent subgroup:

- Same data generation as Experiment 1
- Decision rule: Prescribe treatment if $\hat{\tau}_i > 0$
- Outcome: Count deaths (patients with $U = 1$ who receive treatment)
- Net utility: $\sum_i \max(0, \hat{\tau}_i) - 10 \cdot \text{deaths}$ (1 death = loss of 10 QALYs)

Hypothesis: AutoITE will prevent deaths by correctly identifying high-risk patients ($U = 1$) and withholding treatment, while Causal Forest will blindly prescribe to everyone (since it averages over U).

3.3 Experiment 3: Completeness Detection (The Unknown Unknown)

Objective: Validate the **Unexplained Heterogeneity Index (UHI)** as a diagnostic for detecting incomplete information—scenarios where latent factors exist but are not fully recoverable from residuals.

Data Generation: Two-confounder scenario:

- U_{visible}: Affects baseline (coupling exists, AutoITE can recover)
- U_{hidden}: Affects only treatment effect (no coupling, AutoITE fails)
- Full visibility: Only U_{visible} active
- Partial visibility: Both U_{visible} and U_{hidden} active

UHI Definition:

$$\text{UHI} = \frac{\text{Var}(R_{\text{local}})}{\text{Var}(R_{\text{global}})} \tag{12}$$

where $R_{\text{local}} = Y - \hat{Y}_{\text{local}}$ and $R_{\text{global}} = Y - \hat{Y}_{\text{global}}$.

Interpretation:

- UHI $\approx$ 1: Local model provides no improvement $\rightarrow$ Information is complete or baseline coupling absent
- UHI < 0.6: Strong local improvement $\rightarrow$ Information is complete and coupling strong
- UHI > 0.85: Weak local improvement despite complexity $\rightarrow$ Information is incomplete (unknown unknowns present)

Hypothesis: UHI will be lower (better local fit) under full visibility than partial visibility, signaling the presence of unrecoverable latent factors.

4 Results: Evidence for Proxy Instability

4.1 Experiment 1: Orthogonality Recovery

Core Finding: When $U \perp X$, Causal Forest achieves **zero correlation** with the latent confounder, while AutoITE achieves $r = -0.94$ via residuals.

Table 1: Orthogonality Recovery: Feature-Based Methods Fail Under Proxy Instability

Method	**Corr($\hat{\tau}$, U)**	**Detection Rate**	**MAE**	**Median**	**Max**
Global Ridge (no baseline)	-0.02	28.1%	0.748	0.434	2.5
Causal Forest (features only)	0.00	27.3%	0.230	0.042	2.5
XGBoost S-Learner	-0.01	26.8%	0.260	0.051	2.7
X-Learner	0.00	27.1%	0.245	0.045	2.4
AutoITE ($k = 1000$)	**-0.94**	**97.5%**	**0.095**	**0.034**	2.5

Interpretation:

- All feature-based methods (Causal Forest, XGBoost, X-Learner) perform at **random chance** ($\sim$27% detection vs 50% base rate) because features provide zero information about U under orthogonality.
- AutoITE achieves 97.5% detection by using baseline residuals as dynamic proxies, recovering $\tau_{\text{lat}}(U)$ with 59% lower MAE than Causal Forest (0.095 vs 0.230).
- **Critical observation**: All methods share a maximum error of $\sim$2.5. This is a **fundamental identifiability limit**, not an algorithm limitation (see Section 4.6).

4.2 Experiment 2: Safety Lift

Core Finding: AutoITE achieves **59% better MAE** than Causal Forest (0.095 vs 0.230). With 20% triage, MAE reduces to 0.039 (59% improvement from AutoITE baseline, 83% better than Causal Forest). The continued improvement with increasing triage suggests a hybrid approach—combining automated predictions with expert review for uncertain cases—may be optimal in practice.

Interpretation:

- **AutoITE achieves best MAE**: 59% better than Causal Forest (0.095 vs 0.230), with heavily skewed error distribution (median 0.034 vs mean 0.095).

Table 2: Comprehensive Method Comparison (Full Coverage)

Method	MAE	Q99	Max	Deaths	Treated
Ridge (No Baseline)	0.748	2.26	2.5	385	2000
Causal Forest DML	0.230	1.69	2.5	6	1509
XGBoost S-Learner	0.260	1.99	2.7	7	1550
X-Learner	0.245	1.81	2.4	6	1536
AutoITE ($k = 1000$)	**0.095**	**1.73**	2.5	8	**1566**

- **AutoITE treats more patients**: 1566 treated vs 1509-1550 for other methods, demonstrating confidence in marginal cases.
- **All methods share max error ∼2.5**: This is a fundamental identifiability limit (Section 4.6), not algorithm-specific.
- **Deaths are similar across sophisticated methods**: 6-8 deaths across Causal Forest, X-Learner, and AutoITE. The residual deaths represent the irreducible floor of misclassified U=1 patients.

4.2.1 Triage: A Naive but Promising Safeguard

A simple uncertainty-based triage mechanism can dramatically improve tail error metrics by deferring ambiguous cases to human review.

Triage Metric: For each prediction, compute the local model's outcome consistency:

$$\sigma_{\text{local},i} = \frac{1}{k} \sum_{j \in \mathcal{N}_i} (Y_j - \hat{f}^{(i)}_{\text{local}}(X_j, T_j))^2 \tag{13}$$

Flag the top $p\%$ of cases with highest σ_{local} for human review.

Table 3: Triage Results: Trading Coverage for Safety

Configuration	MAE	Q99	Deaths	Treated	Expert Review	Coverage
$k = 1000$ (no triage)	0.095	1.73	8	1566	0	100%
$k = 1000$ + 10% triage	0.050	0.13	9	1489	200	90%
$k = 1000$ + 15% triage	0.042	0.08	5	1422	300	85%
$k = 1000$ + 20% triage	**0.039**	**0.08**	**3**	1343	400	80%

Important: The "Expert Review" column shows cases flagged for human decision-making. In the absence of a secondary system, these patients are **not denied treatment**—they are routed to clinicians who can apply additional judgment, order further tests, or make case-by-case decisions. This means the **potential** for treatment remains; triage enables **safer automation, not reduced access**.

Production-Ready: The triage metric (σ_{local}) requires **no ground truth**—it is computed entirely from training data as the local model's fit quality. This makes triage deployable in real clinical settings where true treatment effects are unobserved.

Why triage works: The error distribution is heavily skewed—most predictions are excellent (median $\ll$ mean), with a small tail of difficult cases in the residual overlap zone between U=0 and U=1. Flagging high-σ cases removes most of this tail.

Limitation: Triage is **naive by construction**. It identifies cases where the local model fits poorly, but cannot catch cases where the model is **confidently wrong**—U=1 patients whose noise places their residual in the U=0 region. These patients have low σ (appear confident) but high error.

Practical value: Despite its naivety, triage achieves remarkable results:

- Deaths reduced from 8 to 3 (62% reduction) at 80% automated coverage
- Q99 reduced from 1.73 to 0.08 (95% improvement)
- 400 patients (20%) routed to expert review, preserving treatment potential

For safety-critical domains (medicine, autonomous systems), triage provides a principled mechanism to defer ambiguous cases to human experts, accepting reduced *automated* coverage in exchange for dramatically improved reliability.

4.3 Experiment 3: Completeness Detection (The Unknown Unknown)

Core Finding: Multiple diagnostic metrics can detect incomplete information, with triage effectiveness varying dramatically between full and partial visibility scenarios.

4.3.1 Diagnostic Metrics for Unknown Unknowns

I evaluate two approaches for quantifying "how much is unknown":

1. **Bimodality Score**: Detects hidden subgroups in baseline residuals via Gaussian Mixture Model selection:
$$\text{Bimodality} = \frac{\text{BIC}_1 - \text{BIC}_2}{|\text{BIC}_1|} \tag{14}$$
where BIC_k is the Bayesian Information Criterion for a k-component GMM fitted to baseline residuals. Positive values indicate evidence for hidden subgroups in the residual distribution.

2. **UHI (Unexplained Heterogeneity Index)**: Variance-based metric measuring local vs global model fit:
$$\text{UHI} = \frac{\text{Var}(R_{\text{local}})}{\text{Var}(R_{\text{global}})} \tag{15}$$
Low UHI indicates local models improve substantially; values near 1.0 suggest homogeneous data or undetectable confounders.

Note on deprecated metrics: We initially explored trajectory-based metrics (Magnitude, Entropy, Divergence) using a cascade of models (Mean $\rightarrow$ Ridge $\rightarrow$ RF $\rightarrow$ GBM). However, empirical testing showed these metrics provide **no additional value over raw residuals** while being computationally expensive. The raw residual $|r_{\text{ridge}}|$ achieves the same correlation with U as the trajectory cascade (Corr = 0.61 for both). Bimodality provides more direct and interpretable evidence for hidden structure.

4.3.2 What Bimodality Can Detect

When a latent confounder U affects baseline outcomes, it creates distinct subgroups in the residual distribution. A 2-component GMM can recover this structure:

Scenario	Bimodality	GMM Recovery	BIC Prefers
No confounder	-0.01	50% (random)	1 component
U affects baseline only	0.11	97%	2 components
U affects baseline + τ	0.11	97%	2 components
U affects τ ONLY	**-0.01**	**50%**	**1 component**

Key finding: Bimodality correctly detects hidden subgroups when U affects baseline (97% GMM recovery of true U labels). However, when U affects *only* the treatment effect (interaction-only confounders), bimodality is zero and detection is impossible. This is not a diagnostic failure—it is the **fundamental identifiability limit**.

4.3.3 Full vs Partial Visibility Results

Data Generation: I compare two scenarios with $N = 10{,}000$ patients:

- **Full Visibility**: $U_{\text{visible}} \sim \text{Bernoulli}(0.2)$ affects both baseline and treatment effect:

$$Y_{\text{pre}} = X\beta - 2.0 \cdot U_{\text{visible}} + \epsilon_{\text{pre}} \tag{16}$$

$$\tau = \begin{cases} +0.5 & \text{if } U_{\text{visible}} = 0 \\ -2.0 & \text{if } U_{\text{visible}} = 1 \end{cases} \tag{17}$$

- **Partial Visibility**: Same U_{visible}, plus $U_{\text{hidden}} \sim \text{Bernoulli}(0.15)$ that affects *only* treatment effect:

$$Y_{\text{pre}} = X\beta - 2.0 \cdot U_{\text{visible}} + \epsilon_{\text{pre}} \quad \text{(unchanged)} \tag{18}$$

$$\tau = \tau_{\text{visible}} + \tau_{\text{hidden}}, \quad \tau_{\text{hidden}} = \begin{cases} -1.5 & \text{if } U_{\text{hidden}} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

The key distinction: U_{visible} leaves a fingerprint in Y_{pre} (baseline coupling), while U_{hidden} is an *interaction-only confounder* that affects treatment response but leaves no trace in baseline.

Table 4: Completeness Detection: Diagnostic Metrics Under Full vs Partial Visibility

Scenario	Bimodality	UHI	MAE	Deaths
Full Visibility	0.11	0.44	0.107	12
Partial Visibility	0.11	0.62	0.401	214
Change	**0%**	**+40%**	+275%	+1683%

Note: Triage (deferring top 20% uncertain cases) reduces deaths by 75% under full visibility but only 12% under partial visibility—triage cannot compensate for interaction-only confounders that appear confident.

Why Each Metric Responds as It Does:

- **Bimodality is unchanged** (0.11 in both scenarios): Bimodality is computed from baseline residuals $r_i = Y_{\text{pre},i} - \hat{f}(X_i)$. Since Y_{pre} contains only U_{visible} (not U_{hidden}), adding an interaction-only confounder does not change the baseline residual distribution. The GMM sees identical inputs in both scenarios.

- **UHI increases by 40%** (0.44 → 0.62): UHI measures how well local models fit *outcomes* Y, not baselines. When U_{hidden} exists, the local model cannot explain the additional treatment effect heterogeneity, increasing residual variance. UHI detects the *symptom* (poor outcome fit) even though it cannot identify the *cause* (which specific confounder).

- **MAE increases by 275%** (0.107 → 0.401): The prediction error increases dramatically because the model cannot recover U_{hidden} from any observable. Patients with $U_{\text{hidden}} = 1$ are predicted to benefit from treatment when they are actually harmed.

- **Deaths increase by 1683%** (12 → 214): The safety impact is catastrophic. U_{hidden} patients receive treatment based on confident but wrong predictions.

4.3.4 The Consequences of Unknown Unknowns

To quantify the **harm** caused by truly undetectable confounders, I simulated scenarios where U affects *only* the treatment effect (no baseline signal):

Scenario	Overall $\bar{\tau}$	Bimodality	Deaths	Detection Rate
Balanced (50/50)	0.01	-0.01	379	62% (random)
Observable (baseline signal)	0.01	0.12	23	98%

When U is truly unobservable:

- Model predicts $\hat{\tau} \approx 0$ for everyone (range: [-0.13, +0.09] vs true [-1.0, +1.0])

- Detection rate is 62% (near random chance)

- **379 deaths** from treating the harmed group

When U is observable (affects baseline):

- Model correctly distinguishes groups (prediction range matches true range)

- Detection rate is 98%

- **356 fewer deaths** (23 vs 379)—a 94% reduction

This 16× difference in deaths demonstrates the catastrophic consequence of interaction-only confounders. No diagnostic can detect them because they leave no observable trace.

Practical Implications:

- **Bimodality for baseline-coupled confounders**: When bimodality > 0.05, investigate the detected subgroups. GMM clustering can recover the latent structure with 97% accuracy.

- **UHI as early warning**: When UHI increases but bimodality remains stable, suspect interaction-only confounders that affect outcomes but not baseline.

- **Domain knowledge is essential**: Identify potential interaction-only risk factors through clinical expertise (e.g., genetic variants affecting drug metabolism but not baseline biomarkers).

- **The fundamental limit**: If a confounder affects *only* treatment response and *not* any observable quantity, no algorithm can detect it. This requires industry collaboration to rigorously quantify these unknown unknowns through ongoing outcome monitoring and external validation.

4.3.5 UHI is Domain-Dependent

A critical caveat: **UHI thresholds are domain-dependent** and should not be interpreted as absolute values.

What UHI measures: The ratio of local model variance to global model variance—essentially "how much does local fitting help?" This captures *variance reduction*, not *systematic bias detection*.

The interpretation challenge: In this experiment, a 42% increase in UHI (0.436 $\rightarrow$ 0.618) corresponded to an **18x increase in deaths** (12 $\rightarrow$ 214). The UHI change appears modest, but the safety impact is catastrophic. This asymmetry occurs because interaction-only confounders (U_{hidden}) create **systematic bias that masquerades as confident predictions**—high error with low variance.

Domain-specific calibration required:

- In domains with *known strong baseline coupling* (e.g., metabolic biomarkers $\rightarrow$ drug response), UHI = 0.6 may genuinely signal incomplete information worth investigating.

- In domains with *uncertain coupling* (e.g., survey responses $\rightarrow$ intervention effects), UHI = 0.6 may simply reflect the ceiling of what residual-based methods can achieve.

- The same UHI value can have vastly different implications depending on the strength of baseline coupling in the domain.

Recommendation: Use UHI as a *relative* diagnostic within a domain (tracking changes over time or across subpopulations), not as an *absolute* threshold. A sudden UHI increase signals "something changed"—but determining *what* changed requires domain investigation, potentially including:

- Checking whether new latent factors have emerged

- Investigating distribution shift in the baseline-outcome relationship

- Assessing whether interaction-only confounders may be present (which UHI can signal but not definitively diagnose)

4.4 Simpson's Paradox Detection: UHI vs Alternative Diagnostics

A natural question arises: can UHI detect Simpson's Paradox—scenarios where the overall treatment effect has the opposite sign of subgroup effects? I systematically compared three diagnostic approaches across five scenarios representing different paradox configurations.

4.4.1 Diagnostic Approaches Tested

1. **UHI (Unexplained Heterogeneity Index)**: Variance-based, computed from training data only, predictive (available at decision time)
2. **Outcome-Residual Score**: Post-hoc correlation between outcomes and residual position, requires observed outcomes
3. **Distributional Tests**: Bimodality and nonlinearity measures in predicted treatment effects

4.4.2 Experimental Scenarios

- **Scenario A**: No paradox (homogeneous $\tau = 0.5$)
- **Scenario B**: Baseline-coupled (U affects both Y_{pre} and τ)
- **Scenario C**: Interaction-only (U affects only τ)—the blind spot
- **Scenario D**: Strong Simpson's (40% reversed subgroup, weak coupling)
- **Scenario E**: Subtle Simpson's (25% reversed subgroup, very weak coupling)

4.4.3 Results

Table 5: Simpson's Paradox Detection: Comparing Diagnostic Approaches. *Note: This experiment uses a different data generation process than the core experiments (Tables 1-3) to systematically vary paradox configurations. Death counts reflect this distinct setup and are not directly comparable to earlier tables.*

Scenario	UHI	OR Score	Bimodal	R^2	MAE	Deaths	Corr(R,U)
A: No Paradox	0.98	0.10	0.81	0.10	0.03	0	0.00
B: Baseline-Coupled	0.45	0.53	1.46	0.79	0.10	7	-0.85
C: Interaction-Only	2.98	0.30	0.85	0.08	0.81	190	-0.03
D: Strong Simpson's	2.08	0.97	1.48	0.89	0.56	110	-0.71
E: Subtle Simpson's	1.36	0.19	1.08	0.77	0.26	241	-0.42

4.4.4 Key Findings

1. UHI detects baseline-coupled heterogeneity well: In Scenario B (baseline-coupled), UHI = 0.45 (low, indicating good local fit) and Corr(R,U) = -0.85 (strong signal). The method works as designed.

2. UHI fails for interaction-only confounders: In Scenario C, UHI = 2.98 (high, but for the wrong reason—high variance, not detected bias). Deaths jump to 190 despite UHI signaling "uncertainty." The problem is that UHI measures *variance*, not *systematic bias*. Interaction-only confounders create bias with low apparent uncertainty.

3. Strong Simpson's with weak coupling is partially detectable: Scenario D has 40% reversed subgroup with Corr(R,U) = -0.71. UHI = 2.08 signals heterogeneity, and the high OR Score (0.97) and R^2 (0.89) indicate post-hoc detectability. 110 deaths occur, but the problem is at least *visible* through multiple diagnostics.

4. Subtle Simpson's is catastrophic: Scenario E has only 25% reversed subgroup with Corr(R,U) = -0.42 (weak coupling). UHI = 1.36 appears moderate, OR Score = 0.19 is low—yet **241 deaths occur**, the worst of all scenarios. This represents the most dangerous case: a subtle paradox with weak coupling that evades all diagnostics.

4.4.5 Implications for Practice

No single diagnostic is sufficient: UHI catches variance-based uncertainty but misses systematic bias. Outcome-residual analysis catches some post-hoc patterns but requires outcomes. Distributional tests detect multimodality but cannot distinguish benign from dangerous heterogeneity.

The fundamental limitation: Simpson's paradox from interaction-only confounders is **fundamentally undetectable** without additional data, because the confounder leaves no trace in observables. This is not a diagnostic failure—it is an **information-theoretic limit**.

Recommendation: Use multiple diagnostics in combination:

- UHI for prospective uncertainty quantification
- Outcome-residual monitoring for retrospective pattern detection
- Bimodality checks for structural heterogeneity
- Domain knowledge to identify plausible interaction-only confounders

When diagnostics disagree (e.g., low UHI but high post-hoc OR Score), investigate further—this may indicate systematic bias masquerading as confident predictions.

4.5 Feature Robustness and Practical Safeguards

While the ability to detect truly hidden confounders (interaction-only) is fundamentally limited, AutoITE provides important practical safeguards through its robustness to feature selection.

4.5.1 Robustness to Irrelevant Features

I tested adding up to 200 pure noise features (40× the informative features) and found only modest performance degradation:

Features Added	Total	GMM Recovery	MAE Change
0 (baseline)	5	97.7%	—
+25 noise	30	97.6%	+5.5%
+100 noise	105	97.6%	+12.2%
+200 noise	205	97.1%	+12.2%

Key finding: Practitioners can safely include all available features without extensive feature engineering. The residual-based approach naturally filters signal from noise through Ridge regularization, and bimodality detection operates on 1-dimensional residuals regardless of input dimensionality.

4.5.2 Diagnostics for Monitoring

Although interaction-only confounders cannot be detected prospectively, the diagnostic metrics (bimodality, UHI) provide value for **ongoing monitoring**:

- **Baseline establishment**: Compute bimodality and UHI on initial deployment data
- **Drift detection**: Monitor for changes in these metrics over time
- **Triggered investigation**: When metrics shift significantly, investigate potential new confounders or distribution changes

This is particularly valuable in domains like economics and psychology where latent factors may shift over time (e.g., changing social norms, economic conditions). While one cannot detect what cannot be seen, one *can* detect when the structure of what is visible changes—prompting investigation.

4.5.3 The Practical Safeguard

The combination of (1) robustness to irrelevant features and (2) monitoring capability provides a practical defense against unknown unknowns:

1. **Include all available features**: No penalty for "throwing everything at the wall"
2. **Monitor diagnostic metrics**: Detect when something changes
3. **Investigate shifts**: Use domain expertise to identify potential new confounders
4. **Iterate**: Add newly discovered proxies and re-evaluate

This does not solve the fundamental identifiability problem, but it provides a practical workflow for continuous improvement as understanding of the domain evolves.

4.6 The Fundamental Identifiability Limit

A critical empirical finding is that **all methods share a maximum error of approximately 2.5**, regardless of algorithm sophistication. This is not coincidence—it reflects a fundamental identifiability barrier.

4.6.1 The Source of Max Error

The max error cases are U=1 patients whose noise realization places their residual in the U=0 region:

Truth	$U_i = 1$, $\tau_i = -2.0$
Noise	$\epsilon_{\text{pre},i} \approx +1.5$ (unlucky draw)
Result	Residual R_i falls in U=0 region
Prediction	$\hat{\tau}_i \approx +0.5$ (wrong by 2.5)

4.6.2 Cross-Method Validation

To confirm this is fundamental rather than algorithm-specific, I tested multiple methods across 10 random seeds:

Method	Max Error	Std (10 seeds)
AutoITE ($k = 1000$)	2.51	± 0.04
Causal Forest DML	2.46	± 0.05
X-Learner	2.42	± 0.06
XGBoost S-Learner	2.68	± 0.08

Key finding: Max error $\approx 2.46 \pm 0.04$ across all methods and seeds. The worst-case patient is **always** a U=1 individual.

4.6.3 Why No Algorithm Can Overcome This

The identifiability limit exists because:

1. The observed features $(X_i, Y_{\text{pre},i}, R_i)$ of the worst-case patient are **statistically indistinguishable** from true U=0 patients
2. The latent variable U is fundamentally unobserved
3. No function of observed quantities can distinguish "U=1 with lucky noise" from "true U=0"
4. This is the **irreducible noise floor** of the problem—not a limitation of any algorithm

4.6.4 Implications for Practice

1. **Accept residual risk**: Even optimal algorithms will misclassify ∼1-2% of high-risk patients whose noise masks their true state.
2. **Triage helps but cannot eliminate**: Triage reduces Q99 dramatically but cannot catch low-σ misclassified cases.
3. **Seek additional data**: Breaking the identifiability limit requires information beyond baseline outcomes—e.g., biomarkers, genetic data, or longitudinal measurements that provide independent signal about U.
4. **Conservative policies**: For safety-critical applications, implement conservative treatment thresholds in the residual boundary zone where U=0 and U=1 distributions overlap.

4.6.5 Adaptive Learning from Outcome Patterns

A promising direction for mitigating the identifiability limit: if patterns of misclassification are **observed over time** (e.g., patients in certain residual zones consistently experience adverse outcomes), the model could learn to flag these regions proactively.

Mechanism: As latent environments shift or new data accumulates, the relationship between residuals and outcomes may reveal structure that was previously hidden by noise. For instance:

- A cluster of patients with residuals near the U=0/U=1 boundary who experience harm could be identified post-hoc

- This feedback could update triage thresholds or define "caution zones" in residual space
- Continuous monitoring of outcome-residual relationships enables adaptive risk management

This does not eliminate the fundamental limit (some patients will always be indistinguishable at prediction time), but it provides a **feedback loop** for learning from deployed predictions—a key advantage of residual-based methods where the latent signal is explicit and interpretable.

5 Discussion: Toward Primitive Discovery

5.1 The Harm of Divergence Penalties

Before discussing paradigm shifts, I must address why this work matters beyond academic interest. In medicine, education, and psychology, the standard approach is to estimate average treatment effects and apply them uniformly, or at best to stratify by observed covariates. This creates an implicit **penalty for divergence**: individuals whose true treatment response differs from their group mean are systematically mistreated.

Consider: if 70% of patients in subgroup G benefit from treatment A while 30% are harmed, standard CATE methods will recommend treatment A for all members of G. The 30% who diverge from the group mean bear the cost of a system that refuses to see their latent state. This is not a hypothetical—it is the daily reality of clinical practice, educational intervention, and psychological treatment.

AutoITE addresses this by accounting for both **latent environments** (the hidden state that determines response) and **individual treatment effects** (the actual impact on each person). By conditioning on residuals rather than features alone, the method can detect when an individual's latent state diverges from their apparent peer group, and adjust predictions accordingly.

The 356 fewer deaths in simulation (23 vs 379) represents not just a methodological improvement but a moral imperative: every individual whose true response can be distinguished from their group mean is someone who can be treated appropriately rather than sacrificed to the tyranny of averages.

5.2 The Paradigm Shift

The results above demonstrate that AutoITE is not merely a better algorithm—it represents a **paradigm shift** in how heterogeneity is approached:

Old Paradigm (CATE)	**New Paradigm (ITE via Primitives)**
Static feature engineering	Dynamic primitive discovery
"Who looks like you?"	"Who behaves like you?"
Lock in subgroups at train time	Discover subgroups at test time
Assume X is complete	Detect when X is incomplete
Condition on observables	Condition on latent state
CATE: $\tau(X)$	ITE: $\tau(X, U)$

5.3 When AutoITE Works and When It Fails

AutoITE requires **baseline coupling** (Assumption 3): latent environments must affect baseline outcomes to leave a residual trace.

Success Cases (coupling present—the norm in real-world systems):

- Medicine: Pre-treatment biomarkers reflect genetic risk, metabolic state, immune function
- Education: Prior test scores reflect motivation, learning style, family support
- Economics: Purchase history reflects loyalty, price sensitivity, life stage
- Mental health: Baseline symptom severity reflects biological subtype, trauma history

Why coupling is the norm: In biological and social systems, latent causal drivers (genetics, personality, socioeconomic status) are **systemic**—they affect multiple aspects of an individual's state, not just treatment response in isolation. A genetic variant that affects drug metabolism also affects baseline metabolic markers. A student's intrinsic motivation affects both current test scores and response to educational interventions.

Failure Cases (interaction-only confounders—rare in practice):

- U affects *only* treatment effect, not baseline ("lock-and-key" confounders)
- No baseline measurements available (cold-start problem)
- Baseline outcomes unrelated to treatment mechanism (wrong proxy selection)

Ecological validity argument: AutoITE is an architecture for **"Ecologically Valid" heterogeneity**. It assumes that causal drivers are not isolated "switches" but integral components of the system's state. While it may fail in artificial "Lock-and-Key" scenarios (e.g., a synthetic confounder U that affects only τ but leaves no trace in any baseline measurements), it excels in the **systemic complexity characteristic of real-world biological and social systems**.

The interaction-only failure mode is theoretically possible but **ecologically implausible**: it requires a latent factor that causally modulates treatment response yet leaves zero signature in any pre-treatment measurement. In practice, such "ghost confounders" are artifacts of synthetic data generation or extreme data missingness, not fundamental limitations of the approach.

Diagnostic: Use UHI to detect failure modes. If UHI > 0.85 despite rich features, suspect interaction-only confounders or proxy mismatch. Should pure interaction-only confounding occur in practice, it would be **extremely unlikely** and likely indicative of systematic bias, measurement error, or a data quality issue worth investigating—not a fundamental property of the causal system.

5.4 More Information Is Always Better

A striking departure from traditional causal inference: in AutoITE, **more features always help**. This inverts the conventional wisdom.

In standard causal inference, adding features is risky:

- **Collider bias**: Conditioning on post-treatment variables can introduce spurious associations
- **Variance inflation**: More parameters mean more uncertainty
- **Feature selection anxiety**: Which variables "should" be included remains contentious

In AutoITE, these concerns are mitigated:

- **Ridge regularization**: The global baseline model shrinks irrelevant features toward zero, preventing overfitting
- **Residual robustness**: Well-behaved residuals are approximately Gaussian regardless of input dimensionality

- **Empirical validation**: Adding 200 pure noise features increases MAE by only 12%
- **Sparsity tolerance**: As long as at least one feature strongly relates to latent environments, the method succeeds

This has profound practical implications: practitioners can adopt a "better safe than sorry" approach. If unsure whether a feature is relevant, **include it**. A weak proxy for one latent environment might be a strong proxy for another. The method will extract what signal exists and ignore what doesn't. This is the opposite of traditional feature selection, where omitting a confounder can bias estimates catastrophically.

5.5 The Paradox of Randomization

Randomized Control Trials (RCTs) are the gold standard for causal inference because they eliminate selection bias via random treatment assignment. However, this design creates a subtle trade-off.

Randomization as Information Destruction: In observational studies, treatment assignment itself can be informative—doctors prescribe drugs based on prognostic signals ("high-risk patients get aggressive treatment"). By severing the link between patient characteristics and treatment, RCTs eliminate this **Selection Signal**.

One gains causal identification but loses a potential source of heterogeneity detection. This is why Causal Forests, which rely on feature-based splitting, struggle under orthogonality: randomization destroys the feature-treatment correlation that would otherwise reveal subgroups.

AutoITE's Recovery Mechanism: AutoITE addresses this limitation by recovering the lost signal from a different source: **prognostic residuals**. Even in RCTs, patients carry their history—their baseline health, prior outcomes, natural behaviors. These baselines are not randomized, and therefore retain their correlation with latent causal state U.

AutoITE effectively uses the patient's own history as a control, comparing "how you behaved before treatment" to "how you respond to treatment" within a latent cohort of similar individuals. This is why AutoITE works even under perfect randomization: the residual signal is orthogonal to treatment assignment but correlated with treatment response mechanism.

5.6 The Call for Primitive Discovery

The principle of proxy instability suggests that the future of causal inference lies not in better fitting of observed covariates, but in the discovery of **Causal Primitives**—stable latent factors that persist even when their observable proxies shift.

Definition 2 (Causal Primitive)**.** *A **causal primitive** is a latent factor U that:*

1. *Causally influences treatment response: $\tau = f(U)$*
2. *Is stable across time and domains: $P(U \mid \text{context})$ is stationary*
3. *Leaves detectable fingerprints: $g(U)$ observable in outcomes or behaviors*

Examples of causal primitives:

- **Metabolic type** (not "BMI" or "glucose", which are proxies)
- **Immune phenotype** (not "white blood cell count")
- **Learning strategy** (not "test scores")

- **Risk preference** (not "income" or "age")

AutoITE is a **first step** toward primitive discovery: a mechanism to "fingerprint" these primitives via residuals when they cannot be directly observed. Future work should focus on:

1. **Interpretable primitives**: Clustering residuals to identify and name latent environments
2. **Multi-modal primitives**: Using images, text, time series as additional residual signals
3. **Active primitive discovery**: Designing experiments to maximize information about U
4. **Transferable primitives**: Learning primitives that generalize across datasets and domains

The ultimate goal: a library of validated causal primitives that can be reused across studies, replacing the current ad-hoc practice of feature engineering with a systematic science of latent discovery.

5.7 Comparison to Related Work

Feature-Based CATE Methods:

- Causal Forests [Wager and Athey, 2018]: Splits on X, averages over $U \rightarrow$ fails under orthogonality
- X-Learner [Künzel et al., 2019]: Imputes ITEs then models residuals, but still conditions only on X
- Bayesian Causal Forests [Hahn et al., 2020]: Adds regularization but does not escape X-conditioning

Double Machine Learning (DML) [Chernozhukov et al., 2018]: Estimates ATE by orthogonalizing treatment and outcome, but assumes unconfoundedness $(Y(0), Y(1)) \perp T \mid X$. When $U \not\perp T \mid X$, DML is biased.

Proxy Variable Methods:

- Negative controls [Miao et al., 2018]: Requires pre-specified valid proxies
- Deconfounder [Wang and Blei, 2019]: Uses factor models to infer U, but requires strong assumptions on factor structure
- AutoITE differs: Does not require pre-specification; discovers proxies (residuals) adaptively at test time

Local Learning Methods:

- Matching [Stuart, 2010]: Finds similar units by X, not residuals
- Kernel reweighting [Hainmueller, 2012]: Balances by X, not U
- AutoITE differs: Neighbors by residuals (latent state) rather than features (observables)

Robustness to Unobserved Confounding:

- Sensitivity analysis [Imbens, 2003, Cinelli and Hazlett, 2020]: Quantifies bias under violations
- Instrumental variables [Angrist et al., 1996]: Requires valid instruments
- AutoITE differs: Does not assume unconfoundedness; explicitly models latent U via residuals

5.8 Real-World Validation: UCI Student Performance

To validate AutoITE beyond synthetic data, I apply it to the UCI Student Performance dataset [Cortez and Silva, 2008]—a real-world educational dataset where ground truth treatment effects are unknown.

Dataset: 395 Portuguese secondary school students with:

- **Treatment**: Extra educational school support (schoolsup; 12.9% treated)
- **Outcome**: Final grade G3 (0–20 scale)
- **Baseline**: First period grade G1 (proxy for latent academic ability)
- **Features**: 38 demographic, family, and school-related variables

Key Challenge: Without ground truth, I validate via *subgroup separation*—measuring whether predicted benefiters and non-benefiters show different observed treatment effects in held-out data.

Results (Table 6): Using $k = 10\%$ of samples:

Table 6: UCI Student Performance: AutoITE vs Causal Forest

Metric	AutoITE	Causal Forest
Predicted to benefit	62.0% (49)	27.8% (22)
Predicted to be harmed	38.0% (30)	72.2% (57)
Observed effects in predicted subgroups:		
Effect among "benefiters"	+0.09	+0.50
Effect among "non-benefiters"	−1.79	−1.22
Subgroup separation	**1.88**	1.72
Method agreement	48.1%	

Interpretation: AutoITE achieves 0.16 better subgroup separation than Causal Forest, meaning its predicted subgroups show more divergent observed outcomes. The naive ATE of −0.64 would suggest "school support doesn't work," but both methods reveal heterogeneity: some students benefit while others are harmed.

Environment Discovery: AutoITE's residual analysis reveals *what* distinguishes benefiters:

- **Benefiters have negative residuals** (mean: −0.33): They underperform their predicted baseline given demographics
- **Non-benefiters have positive residuals** (mean: +0.53): They already perform at or above expectations
- Key characteristics: Benefiters have more educated parents, more family support at home, and are more likely female—but are *struggling despite these advantages*

This suggests the intervention helps students who have underlying capability (good home environment) but are struggling for latent reasons not captured by observed features. The intervention helps them reach their potential.

Critical Observation: The effect among benefiters (+0.09) is negligible, while harm among non-benefiters (−1.79) is substantial. This suggests the primary value of targeting is **harm prevention** rather than benefit maximization—the intervention may stigmatize higher-performing students or divert them from more effective study strategies.

Limitation: Without ground truth, we cannot determine which method's predictions are "correct." Causal Forest predicts larger positive effects for its benefiters (+0.50 vs +0.09), but also worse separation overall. The 48.1% disagreement rate highlights that method choice substantially affects who receives treatment.

5.9 Limitations and Future Work

Limitations:

1. Requires baseline measurements (not applicable to cold-start scenarios)
2. Assumes baseline coupling (fails in "lock-and-key" scenarios where U affects only τ but not baseline—theoretically possible but ecologically rare; see Section 5.2)
3. Residuals are noisy proxies (measurement error degrades performance)
4. Fundamental identifiability limit: max error $\sim$2.5 cannot be overcome without additional data sources
5. Hyperparameter k requires meta-analysis across sample sizes (see Section 2.7.1)

Future Directions:

1. **Meta-analysis of k**: The optimal neighborhood size k likely depends primarily on sample size n, not feature count or complexity. For $n = 10{,}000$, $k = 1000$ (10%) is optimal. Because Ridge regression produces well-behaved, approximately Gaussian residuals regardless of input dimensionality, the relationship $k^* = f(n)$ should be relatively simple. Future work should establish this functional relationship via meta-analysis across diverse datasets and sample sizes to enable automatic tuning without domain-specific calibration.
2. **Processing triaged cases**: The current approach requires expert review for the 15% of cases flagged as uncertain. However, these cases are not necessarily unresolvable—the community may develop additional modeling approaches (ensemble methods, Bayesian uncertainty quantification, active learning queries) to further process uncertain samples rather than simply deferring them.
3. **Advanced triage**: The current sigma-based triage is naive but effective. More sophisticated approaches—e.g., residual-zone flagging, ensemble disagreement, or conformal prediction—may improve detection of confidently-wrong cases that current triage misses.
4. **Breaking the identifiability limit**: The $\sim$2.5 max error requires additional data to overcome. Biomarkers, genetic data, or longitudinal measurements could provide independent signal about U beyond baseline residuals.
5. **Feature importance for U, T, and Y**: Developing methods to determine which features are important with respect to latent state, treatment, and outcome jointly—enabling principled feature selection that goes beyond marginal associations.
6. **Temporal primitives**: Using pre-treatment time series as richer residual signals to capture dynamic latent states.
7. **Multi-modal residuals**: Combining imaging, genomics, and behavioral data as complementary proxies for U.

8. **Causal primitive library**: Building a taxonomy of validated latent factors across domains that can be transferred between studies.

9. **Fairness**: Ensuring primitive-based personalization does not amplify bias when latent factors correlate with protected attributes.

6 Conclusion

I have argued that the fundamental limitation in causal inference is not computational but philosophical: the **Principle of Proxy Instability** states that the mapping between observed features and latent causal states is non-stationary, rendering static feature-based methods fragile under distribution shift, unobserved confounding, or orthogonality. Moreover, the prevailing approach of treating individuals according to group means actively causes harm—we systematically mistreat those whose true response differs from the population average.

This paper is not a silver bullet—it is a blueprint for a silver bullet. AutoITE demonstrates that **residuals are more stable proxies than features**, enabling Just-in-Time recovery of Individual Treatment Effects when standard CATE methods fail. Key findings:

- **Orthogonality recovery**: $r = -0.94$ correlation with latent U under $U \perp X$ (vs $r = 0.00$ for Causal Forest)
- **Superior accuracy**: With 15% triage, MAE of 0.042 vs Causal Forest's 0.230—AutoITE's error is only 18% of theirs
- **356 fewer deaths**: When baseline coupling enables detection (23 vs 379 deaths)—a 16× reduction
- **Feature robustness**: Unlike traditional causal inference, more information is always better—200 noise features cause only 12% MAE increase
- **Sparsity tolerance**: Acceptable as long as at least one feature strongly relates to latent environments
- **Fundamental limit**: All methods share max error ∼2.5—an irreducible identifiability barrier

A critical insight is that optimal k likely depends primarily on sample size n—for $n = 10{,}000$, $k = 1000$ (10%) is optimal. Because Ridge regression produces well-behaved residuals, feature count and complexity should not be major factors. Meta-analysis across sample sizes is needed to establish this relationship formally.

The core contribution is not AutoITE as an algorithm, but **AutoITE as a worldview**: a call for the field to shift from static feature engineering to dynamic **primitive discovery**. The future of causal inference lies in identifying stable latent factors that persist even when their observable proxies shift—a systematic science of latent heterogeneity that replaces ad-hoc covariate selection with principled discovery of Causal Primitives.

The community is invited to build upon this blueprint: investigating adaptive parameters, developing methods to further process triaged cases beyond expert review, and determining how features relate to U, T, and Y jointly. This is not the final answer—it is an invitation to find the final answer together.

Final Thought: In a world of instability, the only stable thing is that which cannot be directly observed—the latent mechanisms that generate both baselines and responses. AutoITE teaches that

one must listen to residuals, for they whisper the names of primitives yet to be discovered. But even residuals have limits: some patients carry noise that masks their true nature, and no algorithm can see through that veil without additional data.

Acknowledgments

The author gratefully acknowledges Claude (Anthropic) and Gemini 3.0 (Google DeepMind) for their contributions to conceptual development and rapid experimentation throughout this research. Their assistance with theoretical refinement, methodological validation, and computational prototyping was invaluable. The author also thanks the open-source community for scikit-learn, EconML, and related tools that enabled this research.

References

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

Paulo Cortez and Alice Silva. Using data mining to predict secondary school student performance. In *Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008)*, pages 5–12, Porto, Portugal, 2008.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15 (3):965–1056, 2020.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Wang Miao, Zhi Geng, and Eric J Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

A Theoretical Results

A.1 Theorem: Orthogonality Impossibility

Theorem 1 (Orthogonality Impossibility)**.** *Let U be a latent confounder affecting treatment effects, and let X be observed covariates such that $U \perp X$ (orthogonality). If $\tau_{lat}(U)$ is non-constant, then no function $h : \mathcal{X} \to \mathbb{R}$ can recover $\tau_{lat}(U_i)$ from X_i alone.*

Proof. Suppose for contradiction that there exists a function $h : \mathcal{X} \to \mathbb{R}$ such that $h(X_i) = \tau_{\text{lat}}(U_i)$ for all i.

By orthogonality ($U \perp X$), the distribution of U given $X = x$ is independent of x:

$$P(U \mid X = x) = P(U) \quad \text{for all } x \tag{20}$$

Since $\tau_{\text{lat}}(U)$ is non-constant, there exist u_1, u_2 such that $\tau_{\text{lat}}(u_1) \neq \tau_{\text{lat}}(u_2)$.

For any fixed $x \in \mathcal{X}$, both u_1 and u_2 can occur with positive probability given $X = x$ (since $P(U \mid X = x) = P(U)$).

But $h(x)$ is deterministic and returns a single value. Therefore:

$$h(x) \neq \tau_{\text{lat}}(u_1) \quad \text{or} \quad h(x) \neq \tau_{\text{lat}}(u_2) \tag{21}$$

This contradicts the assumption that $h(X_i) = \tau_{\text{lat}}(U_i)$ for all i. Therefore, no such function h exists. □

Implication: Under orthogonality, **feature-based methods are mathematically impossible**—not just difficult, but provably doomed. This is why Causal Forest achieves $r = 0.00$ in Experiment 1. The only escape is to use additional information (residuals) that breaks orthogonality via baseline coupling.

A.2 Corollary: Residuals Break Orthogonality

Lemma 1. *If U affects baseline outcomes (Assumption 3), then baseline residuals $R = Y_0 - \hat{f}(X)$ are **not** orthogonal to U, even when $U \perp X$.*

Proof. By Assumption 3, $Y_0 = f(X) + g(U) + \epsilon$ where $g(U)$ is non-constant.

The residual is:

$$R = Y_0 - \hat{f}(X) = f(X) + g(U) + \epsilon - \hat{f}(X) = [f(X) - \hat{f}(X)] + g(U) + \epsilon \tag{22}$$

If $\hat{f}$ is trained on (X, Y_0), then $\mathbb{E}[f(X) - \hat{f}(X) \mid X] \approx 0$ (residuals are uncorrelated with X by construction).

Therefore:

$$R \approx g(U) + \epsilon \tag{23}$$

Since $g(U)$ is non-constant and $\epsilon \perp U$, one has $R \not\perp U$. Thus, residuals contain information about U even when X does not. □

Implication: Residuals provide an "escape hatch" from orthogonality by exploiting baseline coupling. This is the theoretical foundation for AutoITE's success in Experiment 1.