Babylonian Journal of Artificial Intelligence
Vol.2026, **pp.** 77-97
DOI: https://doi.org/10.58496/BJAI/2026/013; ISSN: 3006-5437
https://mesopotamian.press/journals/index.php/BJAI/index

Research Article

Coupling-Aware Counterfactual Reasoning for Causal Artificial Intelligence: Identification, Learnability Limits, and Sharp Bounds with Finite-Sample Inference

Jordi Vallverdú[1,*,]

[1]*Departament de Filosofia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Catalonia, Spain*

A R T I C L E I N F O

Article History
Received 05 Sep 2026
Revised 26 Oct 2026
Accepted 05 Nov 2026
Published 08 Nov 2026

Keywords
Causal AI
Counterfactual reasoning
Potential outcomes
Partial identification
Probability of benefit
Probabilities of causation
Optimal transport
LLM evaluation

ABSTRACT

Probabilities of benefit, probabilities of causation and the dispersion of individual effects depend on the joint law of potential outcomes, which interventional data do not fix. Building on Fréchet–Makarov and partial-identification results, we show that every coupling of the interventional laws is induced by an empirically equivalent structural causal model, characterize the feasible joint laws when observational and experimental data are combined, and restate the classical criterion that an expected cross-world score is learnable exactly when it is additively separable. We turn these facts into a coupling audit that reports sharp identification intervals, computed by linear programming, with projection confidence intervals that have finite-sample coverage and asymptotic Imbens–Manski intervals. Over all couplings, the sharp range of the probability that an individual benefits is twice as wide as the Gaussian-copula range; under empirically indistinguishable models, an independence plug-in estimator remains biased by up to 0.16; two public randomized trials yield informative but wide intervals. On a 240-item stress test, three open-source language models answered identified items almost perfectly and non-identified items never correctly. Core results are machine-checked in Lean 4. Conclusions are limited to finite or univariate outcomes, the stated model classes and, for the pilot, small models.

1. INTRODUCTION

Counterfactual reasoning is increasingly embedded in artificial-intelligence systems. It enters individual-level fairness criteria [1, 2], causal formulations of algorithmic recourse [3, 4], causal representation learning [5], unit selection [6], and decision problems that ask whether an intervention would benefit a particular unit. In causal inference the corresponding targets include the probabilities of necessity and sufficiency [7–9], the probability that treatment benefits or harms a unit [10, 11], and the distribution of individual treatment effects [12–14]. These targets compare outcomes under mutually incompatible interventions for the same unit and therefore depend on a *joint* law of potential outcomes [15, 16].

Causal reasoning is also an explicit target of language-model evaluation. CLadder generates natural-language questions at the associational, interventional and counterfactual levels from formal causal models [17]; Corr2Cause tests whether models can infer causal structure from correlational statements [18]; and CounterBench focuses on formal counterfactual inference [19]. In a controlled study of small open-source models on CLadder and Corr2Cause, Vallverdú and Xu [20] reported that 13 of 15 paired comparisons between a same-model role-specialized multi-agent system and the best paired-tested single-agent baseline were statistically indistinguishable after Benjamini–Hochberg correction, while the multi-agent

* *Corresponding author. Email: jordi.vallverdu@uab.cat*

configuration cost about 3.4–5.3 times the per-query latency and 4.2–4.6 times the output tokens of guided chain-of-thought on the primary non-thinking models; they also observed that shared causal misconceptions can survive consensus. These findings motivate a complementary theoretical question: even for a computationally perfect reasoner, which counterfactual answers are determined by the information supplied to it?

Let Y_0 and Y_1 denote the outcomes that would occur under interventions $X = 0$ and $X = 1$. A randomized experiment identifies the marginal laws $\mu_0 = \mathrm{Law}(Y_0)$ and $\mu_1 = \mathrm{Law}(Y_1)$ but not the joint law $\mathrm{Law}(Y_0, Y_1)$, which requires a *coupling* of the two marginals. This is classical: Fréchet bounds the cells of a table with given margins [21]; Makarov and others give sharp bounds on the distribution of a sum or difference with fixed marginals [22–25]; econometrics has used these bounds for the distribution of treatment effects [12–14]; and machine learning has recently represented counterfactual assumptions through transport maps or random processes with fixed marginals [26, 27].

A fully specified structural causal model (SCM) does provide a joint counterfactual law. If $Y = g(X, U)$, SCM semantics hold the exogenous state U fixed across interventions and set $(Y_0, Y_1) = (g(0, U), g(1, U))$ [28, 29]. The learning question is different: if several latent SCM representations agree on everything the learner can observe, which joint counterfactual conclusions are invariant across them, and what should an AI system report when they are not? The aim of this paper is to separate two sources of information that are often blended in applications: what is learned from observational and interventional data, and what is supplied by an assumption about cross-world dependence.

Contributions and their status

Table I separates established results, which we restate with proofs only to keep the argument self-contained, from the contributions of this paper.

1. **Representation theorems (Section 4).** For a randomized binary action, the set of shared-disturbance couplings induced by empirically equivalent SCMs is the full coupling set $\Pi(\mu_0, \mu_1)$ (Theorem 1); this elementary statement is closely aligned with canonical representations [27]. We extend it to confounded designs that combine an observational study with an experiment (Theorem 2). The resulting linear program reproduces the Tian–Pearl bounds on probabilities of causation exactly (verified on 2,000 random instances to 3×10^{-13}) and applies unchanged to multi-valued outcomes and arbitrary linear cross-world targets. Linear-programming bounds over response types are classical [30] and have been automated for general discrete models [31, 32]; our contribution here is the explicit SCM-level characterization of the feasible set that the audit optimizes over, not the LP technique.

2. **A precise notion of learnability (Section 5).** We define learnability as consistent estimability over the model class (Definition 4) and show that, for finite expectation-type targets, learnability, identification and additive separability of the score coincide (Propositions 1 and 3 and Corollary 3). The separability criterion is a classical property of transportation polytopes and the impossibility statement is a direct corollary of non-identification; our contribution is to package them as an operational test. We also make explicit how PN and PS fit the criterion, including when their denominators are identified and nonzero (Proposition 2).

3. **A coupling audit with finite-sample inference (Sections 6 and 7).** The audit reports a sharp identification interval computed by transportation linear programming, optionally sharpened by pre-treatment covariates, together with projection confidence intervals whose finite-sample coverage of the whole identified set holds uniformly over the model class, and asymptotic Imbens–Manski intervals for the target [33, 34]. A stability lemma (Lemma 2) gives consistency of the plug-in bounds.

4. **Numerical and empirical evidence (Sections 8 to 10).** We compare the Gaussian-copula range with the sharp interval over all couplings and prove that, for equal-variance Gaussian marginals, the Gaussian family spans exactly half of the length of the sharp range of every threshold probability $P(Y_1 - Y_0 > d)$ (Proposition 6). A Monte Carlo study with empirically equivalent SCMs shows that data cannot distinguish them and that a plug-in estimator that adds a cross-world independence assumption remains biased as n grows; the projection intervals covered the identified set in every replication, and the Imbens–Manski intervals had coverage between 94.4% and 100%, with the lowest values, at the least favourable endpoint, compatible with the nominal 95% within Monte Carlo error. Two public randomized trials [35, 36] illustrate how the audit changes decision-relevant statements.

5. **A coupling stress test for AI reasoners (Section 11).** We release C-STRESS, a 240-item benchmark with gold answers, a three-part scoring rubric (identification recognition, assumption provenance, conditional computation) and an automatic scorer tested on scripted reference responders. In a pilot with three open-source 7–9B language models under one prompt (Section 11.3), all three answered identified items almost perfectly and gave no correct coupling-aware answer on any non-identified item; we do not generalize beyond these models and conditions.

The finite-outcome versions of the main results (Theorems 1 and 2, Propositions 1 and 2, Lemma 2, the deterministic core of Proposition 5, the binary formulas of Appendix A, the validity of the Tian–Pearl bounds for the program of Theorem 2, and the key inequality of Proposition 6 for the standard Gaussian distribution function) have been machine-checked in the Lean 4 proof assistant; Appendix C lists exactly what is and is not formalized, and the formal development is part of the supplementary material.

TABLE I. Status of the formal results in this paper. †: the finite-outcome statement is machine-checked in Lean 4 with Mathlib. ‡: only a core step is machine-checked (stated in Appendix C); the remainder is proved in the text. Unmarked: proved in the text and, where stated, verified numerically.

Result	Status	Relation to prior work
Theorem 1 (randomized representation)†	elementary; restated	aligned with canonical representations [27]; Fréchet classes [21]
Theorem 2 (combined-data representation, LP)†	SCM-level statement; LP technique classical	recovers Tian–Pearl bounds [8]; response-type LPs [30]; automated bounds [31, 32]
Proposition 1 (separability ⇔ invariance)†	classical	transportation polytopes and LP duality [37]
Corollary 3 (no data-only learner)	direct corollary	standard consequence of non-identification [38, 39]
Proposition 3 (learnable ⇔ separable)	new formulation	combines the two previous results with Lemma 2
Proposition 2 (PN, PS in the criterion)‡	restated + explicit conditions	bounds of [8]; denominator conditions made explicit
Lemma 2 (stability of LP bounds)‡	standard type; proof given	stability of optimal transport [40]
Proposition 5 (projection CI)‡	standard method	projection and partial-identification inference [33, 38]
Proposition 6 (Gaussian half-width)‡	new (elementary)	uses Makarov bounds [22, 24]
Audit, simulations, data analyses, C-STRESS	new	–

2. RELATED WORK

2.1 Joint potential-outcome distributions and Fréchet–Makarov bounds

The joint law of (Y_0, Y_1) is not identified by the marginal laws, so functionals of it are partially identified [38, 39, 41]. For binary outcomes the relevant object is a 2×2 table with known margins, whose cells obey the Fréchet bounds [21]. For real-valued outcomes, Makarov [22], Rüschendorf [23] and Frank, Nelsen and Schweizer [24] derived best-possible bounds (the pointwise infimum and supremum over couplings) on the distribution function of a sum (equivalently a difference) of two variables with fixed marginals; Williamson and Downs [25] gave numerical methods for these dependency bounds. Heckman, Smith and Clements [12] brought these ideas to programme evaluation; Fan and Park [13] developed estimation and inference for sharp bounds on the treatment-effect distribution, and Firpo and Ridder [14] characterized sharp bounds for its functionals. Recent work bounds the fraction negatively affected by treatment [10], studies prediction intervals and sharp bounds for individual treatment effects [42], and uses optimal-transport duality to obtain covariate-assisted bounds and inference [43]. Our audit is a computational specialization of this literature to finite-outcome linear targets, with two additions: the SCM-level representation results that justify the feasible set, and a reporting protocol for AI systems.

2.2 Probabilities of causation and unit-level targets

Pearl [7] defined the probabilities of necessity (PN), sufficiency (PS) and necessity-and-sufficiency (PNS); Tian and Pearl [8] derived sharp bounds from experimental data, observational data, or both, and identification under monotonicity. Li and Pearl [6] used these bounds for unit selection, and Mueller, Li and Pearl [9] tightened them with covariates and causal structure. Ciocănea-Teodorescu, Gabriel and Sjölander [11] proposed a sensitivity analysis for the probability of benefit in binary randomized trials. Linear programming over response types goes back to Balke and Pearl [30], and general algorithms now derive sharp symbolic or numerical bounds for discrete causal models from combinations of observational and experimental data [31, 32]. Our Theorem 2 is a special case stated at the SCM level: it identifies the feasible set as the set of SCM-induced joint laws of (X, Y_0, Y_1), so that the Tian–Pearl combined-data bounds are the values of one transportation-type program that also covers multi-valued outcomes and arbitrary linear cross-world scores.

2.3 Treatment-effect heterogeneity

Conditional average treatment effects $\mathbb{E}(Y_1 - Y_0 \mid Z)$ are identified in a randomized trial and are routinely estimated with machine learning [44, 45]. They are additively separable targets in the sense of Section 5. Heterogeneity *within* covariate strata, such as the probability of benefit given Z or the variance of individual effects given Z, is not identified; covariates can narrow but not close the identification region [9, 43]. Sections 9 and 10 quantify this narrowing.

2.4 Counterfactuals in AI

The term *counterfactual explanation* covers several objects. Some methods construct nearby feature vectors that flip a classifier prediction; these need not be causal counterfactuals in the SCM sense, and recent work studies their robustness [46]. Causal recourse interprets recommended actions as interventions in an SCM [3, 4]. Counterfactual fairness reasons across factual and counterfactual worlds [1], and Russell et al. [2] already confronted competing counterfactual assumptions by optimizing fairness across several causal worlds. Our results apply when an AI target depends on a joint or conditional cross-world law; they do not imply that every feature-perturbation explanation or every interventional policy value is coupling-sensitive, and Section 5 identifies the insensitive targets exactly.

2.5 Language-model benchmarks for causal reasoning

CLadder tests formal causal inference from specified models with oracle answers [17]; Corr2Cause tests causal-discovery reasoning from correlational premises [18]; CounterBench targets formal counterfactual inference [19]. In these benchmarks each item has a unique correct answer. The empirical study of Vallverdú and Xu [20] found that guided chain-of-thought and self-consistency were often competitive with a more expensive multi-agent architecture and that consensus can preserve shared misconceptions. The present paper addresses a different limit: consensus, prompting, model scale and additional inference calls cannot recover a unique value for a coupling-sensitive target when the supplied information does not identify it. This motivates separating *reasoning errors* from *identification errors* in evaluation (Section 11).

2.6 Transport and canonical counterfactual representations

De Lara et al. [26] formulate transport-based counterfactual models as joint distributions between observable laws and study their links with causal counterfactuals. De Lara [27] introduces canonical representations of Markovian SCMs through random-process laws with preassigned marginals, separating counterfactual choices from interventional constraints, and clarifies when potential-outcome and do-intervention formalisms coincide [47].

What this paper adds. We do not claim novelty for the observation that fixed marginals admit many couplings, for the separability property of transportation programs, or for the Fréchet–Makarov bounds. Relative to identification theory and transport-based approaches, the paper contributes (i) an SCM-level characterization of the feasible joint laws in combined observational–experimental designs with confounding (Theorem 2), which places the audit within the response-type linear-programming tradition [30–32]; (ii) a formal link between learnability, identification and separability that yields a one-line test for any finite expectation-type objective; (iii) an audit that reports sharp intervals with projection confidence intervals that have finite-sample coverage and asymptotic Imbens–Manski intervals, and separates data from assumptions; (iv) a quantitative comparison between parametric sensitivity families and sharp bounds, a simulation with empirically equivalent SCMs, and applications to two public trials; and (v) a released benchmark and scorer that turn the distinction into an evaluation protocol for AI reasoners.

3. SETUP, ASSUMPTIONS AND DEFINITIONS

Table II collects the notation. We begin with a randomized binary action; the action module of an AI decision system in which the action is externally assigned has the same structure. Let $X \in \{0, 1\}$ and let Y take values in a measurable space $\mathcal{Y}$.

Definition 1 (Randomized two-action SCM)**.** *A model is a tuple $M = (\mathcal{U}, P_U, P_X, g)$, where $U \sim P_U$ is exogenous, $X \sim P_X$, and $Y = g(X, U)$ for a measurable map $g : \{0, 1\} \times \mathcal{U} \to \mathcal{Y}$.*

The analysis rests on four explicit assumptions.

Assumption A1 (Randomization and positivity)**.** $X \perp\!\!\!\perp U$ *and* $0 < P_X(1) < 1$.

Assumption A2 (SCM counterfactual semantics)**.** *Potential outcomes are $Y_x = g(x, U)$ with the same U in both worlds; no restriction is placed on g, $\mathcal{U}$ or P_U beyond measurability.*

TABLE II. Notation.

Symbol	Meaning
$X \in \{0,1\}, Y, U$	action, outcome, exogenous state
$Y_x = g(x,U)$	potential outcome under $\mathrm{do}(X=x)$
$\mu_x = \mathrm{Law}(Y_x)$	interventional outcome law
$\Pi(\mu_0,\mu_1)$	set of couplings (joint laws with margins μ_0,μ_1)
$\gamma_M = \mathrm{Law}_M(Y_0,Y_1)$	cross-world law induced by SCM M
$\mathrm{Emp}(M)$	empirical content available to the learner
p_i, q_j	finite-support masses of Y_0 and Y_1
$H = (h_{ij})$	score matrix of an expectation-type target ψ_H
p_0, p_1	$P(Y_0=1)$, $P(Y_1=1)$ in the binary case
$\mathrm{PoB} = \mathrm{PNS}$	$P(Y_0=0, Y_1=1)$ (probability of benefit)

Assumption A3 (Data regime)**.** *The learner has access to (samples from) the observational law $P(X,Y)$ and the single-action interventional laws μ_0,μ_1; no outcome is ever observed under both actions for the same unit.*

Assumption A4 (No added cross-world restriction)**.** *Unless stated otherwise, the model class $\mathcal{M}$ consists of all models satisfying Assumptions A1 and A2; any further restriction (monotonicity, independence, a copula, a mechanism) is treated as an explicit added assumption.*

Assumption A1 is relaxed in Theorem 2. The interventional law under action x is $\mu_x := \mathrm{Law}_M\{g(x,U)\} = P_M(Y \mid \mathrm{do}(X=x))$.

Lemma 1 (Observational data are redundant under randomization)**.** *Under Assumption A1, $P_M(X,Y) = P_X(X)\,\mu_X(Y)$. Hence the observational law is a function of (P_X,μ_0,μ_1) and imposes no restriction on the cross-world law beyond those imposed by μ_0 and μ_1.*

Proof. For measurable B, $P(X=x, Y\in B) = P(X=x, g(x,U)\in B) = P_X(x)\,P(g(x,U)\in B) = P_X(x)\mu_x(B)$ because $X \perp\!\!\!\perp U$. □

Definition 2 (Empirical content and equivalence)**.** *The empirical content is $\mathrm{Emp}(M) := (P_X,\mu_0,\mu_1)$. Models M, M' are empirically equivalent, $M \sim M'$, when $\mathrm{Emp}(M) = \mathrm{Emp}(M')$. By Lemma 1 this is agreement on all observational and single-action interventional distributions.*

Definition 3 (Couplings and induced couplings)**.** *$\Pi(\mu_0,\mu_1)$ is the set of probability measures on $\mathcal{Y}\times\mathcal{Y}$ with marginals μ_0,μ_1. The cross-world law induced by M is $\gamma_M := \mathrm{Law}_M(g(0,U),g(1,U))$, and for an equivalence class $[M]$, $\mathcal{R}([M]) := \{\gamma_{M'} : M'\sim M\}$.*

Definition 4 (Identification, sharp interval, learnability)**.** *Let ψ be a real functional of the cross-world law and $\mathcal{M}$ a model class.*

(a) ψ is identified *at M if $\psi(\gamma_{M'}) = \psi(\gamma_M)$ for all $M'\in\mathcal{M}$ with $M'\sim M$ [38, 48].*

(b) The sharp identification interval *at M is $[\inf,\sup]\{\psi(\gamma_{M'}) : M'\in\mathcal{M},\ M'\sim M\}$.*

(c) ψ is learnable *over $\mathcal{M}$ from the data regime of Assumption A3 if there is a sequence of estimators $\widehat{\psi}_n$, each a measurable function of n i.i.d. draws from the data-generating distributions, such that $\widehat{\psi}_n \to \psi(\gamma_M)$ in probability under every $M\in\mathcal{M}$.*

We call an interval *coupling-robust* if it contains $\psi(\gamma_{M'})$ for every $M'\sim M$; the sharp interval is the shortest such interval. Throughout, "robust" is used only in this sense.

4. REPRESENTATION THEOREMS

Theorem 1 (Full coupling representation)**.** *Under Assumptions A1, A2 and A4, for every M with interventional laws μ_0,μ_1, $\mathcal{R}([M]) = \Pi(\mu_0,\mu_1)$.*

Proof. For $M'\sim M$ the marginals of $\gamma_{M'}$ are μ_0 and μ_1, so $\mathcal{R}([M])\subseteq\Pi(\mu_0,\mu_1)$. Conversely, fix $\gamma\in\Pi(\mu_0,\mu_1)$ and define M_γ with $U=(U_0,U_1)\sim\gamma$ independent of X, the same P_X, and $g_\gamma(x,U)=U_x$. Under $\mathrm{do}(X=x)$ the outcome is $U_x\sim\mu_x$, so $M_\gamma\sim M$, and $\gamma_{M_\gamma} = \mathrm{Law}(U_0,U_1) = \gamma$. □

Remark 1. *Theorem 1 does not say that parameterizations of one fixed scientific model are interchangeable for every purpose. It says that observational and single-action interventional data do not privilege one shared-disturbance coupling. Structural knowledge that excludes members of* $\Pi(\mu_0, \mu_1)$ *is precisely what carries cross-world information. The coordinate construction in the proof is closely aligned with canonical representations [27].*

4.1 Confounded designs with combined observational and experimental data

The randomized setting is the most favourable one for learning interventional laws, and Lemma 1 shows that observational data then add nothing. Without randomization the situation changes: observational data constrain the cross-world law *within* each action group. Replace Assumption A1 by allowing arbitrary dependence between X and U (for example $X = f(U)$), and suppose the learner sees the observational law $P(X, Y)$ and, from an experiment on the same population, μ_0 and μ_1. Cross-world targets such as $\mathrm{PN} = P(Y_0 = 0 \mid X = 1, Y = 1)$ are functionals of the law of the triple (X, Y_0, Y_1), so we study $J_M := \mathrm{Law}_M(X, g(0, U), g(1, U))$.

Theorem 2 (Combined-data representation)**.** *Let* $\mathcal{M}_c$ *be the class of models* $Y = g(X, U)$ *with arbitrary joint law of* (X, U) *and Assumption A2. For data* $(P(X, Y), \mu_0, \mu_1)$ *arising from some* $M \in \mathcal{M}_c$*, the set* $\{J_{M'} : M' \in \mathcal{M}_c$ *with the same data*$\}$ *equals the set* $\mathcal{J}$ *of laws* J *of* (X, Y_0, Y_1) *such that*

$$J(X = 0,\, Y_0 \in B) = P(X = 0, Y \in B), \qquad J(X = 1,\, Y_1 \in B) = P(X = 1, Y \in B), \tag{1}$$

$$J(Y_0 \in B) = \mu_0(B), \qquad J(Y_1 \in B) = \mu_1(B), \tag{2}$$

for all measurable B*.*

Proof. ($\subseteq$) Consistency $Y = Y_X$ gives (1), and the experimental margins give (2). ($\supseteq$) Given $J \in \mathcal{J}$, let $U = (X^*, U_0, U_1) \sim J$, set $X := X^*$ and $g(x, U) = U_x$. Then $P(X = x, Y \in B) = J(X = x, U_x \in B)$ matches (1), $\mathrm{Law}(U_x)$ matches (2), and $J_{M'} = J$. □

For finite outcomes, every target that is linear in J (and every ratio of such a target to an identified observational probability) has sharp bounds given by a linear program in the $2mn$ cells of J with constraints (1)–(2). For binary X and Y and $\mathrm{PN} = J(1, 0, 1)/P(X = 1, Y = 1)$, $\mathrm{PS} = J(0, 0, 1)/P(X = 0, Y = 0)$ and $\mathrm{PNS} = \sum_x J(x, 0, 1)$, the program's values coincide with the closed-form combined-data bounds of Tian and Pearl [8]. This is a consequence, not a coincidence: Tian and Pearl's bounds are sharp over all models compatible with the same data, Theorem 2 shows that $\mathcal{J}$ is exactly the set of joint laws induced by such models, and the program attains its optima on $\mathcal{J}$ (Proposition 4). The numerical check (2,000 random joint laws, maximum absolute deviation 2.9×10^{-13}) verifies the implementation; the validity of the Tian–Pearl PN bounds on $\mathcal{J}$ is also machine-checked (Appendix C).
The constraints (1)–(2) do not by themselves imply $X \perp\!\!\!\perp (Y_0, Y_1)$. When the model class is restricted to randomized assignment, this independence is an additional structural restriction that must be imposed on J. Under it, the conditional law of (Y_0, Y_1) given $X = x$ equals one common coupling $\gamma \in \Pi(\mu_0, \mu_1)$ for every arm with $P_X(x) > 0$, i.e. $J(x, \cdot, \cdot) = P_X(x)\,\gamma$, and the representation reduces to Theorem 1. Randomization thus restricts the feasible set; what it removes, relative to the confounded class, is any information that the observational arm could carry about the cross-world law within each action group (Lemma 1).

Corollary 1 (Identification criterion)**.** *Under Assumptions A1 to A4, a functional* ψ *on* $\Pi(\mu_0, \mu_1)$ *is identified if and only if it is constant on* $\Pi(\mu_0, \mu_1)$*. Under Theorem 2 the same holds with* $\mathcal{J}$ *in place of* $\Pi(\mu_0, \mu_1)$*.*

5. WHICH JOINT COUNTERFACTUAL OBJECTIVES ARE LEARNABLE?

Assume finite supports $\mathcal{Y}_0 = \{a_1, \ldots, a_m\}$ and $\mathcal{Y}_1 = \{b_1, \ldots, b_n\}$ with $p_i = \mu_0(a_i) > 0$ and $q_j = \mu_1(b_j) > 0$. A coupling is an $m \times n$ nonnegative matrix $\Gamma = (\gamma_{ij})$ with row sums p_i and column sums q_j (the transportation polytope). Many targets are expectations of a cross-world score,

$$\psi_H(\Gamma) = \sum_{i=1}^{m} \sum_{j=1}^{n} h_{ij}\gamma_{ij}, \qquad h_{ij} = h(a_i, b_j). \tag{3}$$

5.1 The additive-separability criterion

The following equivalence is a classical property of transportation linear programs: the polytope has dimension $(m - 1)(n - 1)$ around any strictly positive point, and a linear objective is constant on it exactly when it lies in the span of the row and column constraints [37]. We restate it with a short proof because the rectangle form (ii) is what the audit computes.

Proposition 1 (Separability criterion; classical)**.** *Assume $p_i, q_j > 0$. The following are equivalent: (i) ψ_H is constant on $\Pi(\mu_0, \mu_1)$; (ii) every rectangle contrast vanishes,*

$$h_{ij} + h_{k\ell} - h_{i\ell} - h_{kj} = 0 \quad \textit{for all } i, k, j, \ell; \tag{4}$$

(iii) $h_{ij} = \alpha_i + \beta_j$ for some numbers α_i, β_j.

Proof. (iii)⇒(i): $\psi_H(\Gamma) = \sum_i \alpha_i p_i + \sum_j \beta_j q_j$ for every feasible Γ. (ii)⇒(iii): fix i_0, j_0, put $\alpha_i = h_{ij_0}$ and $\beta_j = h_{i_0 j} - h_{i_0 j_0}$, and apply (4) to (i, i_0, j, j_0). (i)⇒(ii): the independent coupling $p_i q_j$ is strictly positive, so for distinct i, k and j, ℓ and small $\varepsilon > 0$ the perturbation $+\varepsilon$ at $(i, j), (k, \ell)$ and $-\varepsilon$ at $(i, \ell), (k, j)$ is feasible; constancy forces $\varepsilon(h_{ij} + h_{k\ell} - h_{i\ell} - h_{kj}) = 0$. □

Corollary 2 (Examples)**.** *In a finite-outcome randomized two-action problem: (a) $\mathbb{E}(Y_1 - Y_0)$ is identified, since $h = -y_0 + y_1$ is separable; (b) every target $\mathbb{E}\{a(Y_0) + b(Y_1)\}$, including conditional average effects within covariate strata, is identified; (c) for binary outcomes,* $\mathrm{PoB} = P(Y_0 = 0, Y_1 = 1)$ *is not identified when $0 < p_0 < 1$ and $0 < p_1 < 1$: its score $h = \mathbf{1}\{y_0 = 0, y_1 = 1\}$ has rectangle contrast 1; (d) if both supports contain two distinct values, $\mathbb{E}[(Y_1 - Y_0)^2]$ has rectangle contrast $2(a_i - a_k)(b_\ell - b_j) \neq 0$, so $\mathrm{Var}(Y_1 - Y_0)$ is not identified.*

5.2 Probabilities of necessity and sufficiency

PN and PS are conditional, not unconditional, cross-world probabilities. Their relation to the criterion requires their denominators.

Proposition 2 (PN and PS under randomization)**.** *Let X, Y be binary and assume Assumptions A1 and A2. Then*

$$\mathrm{PN} := P(Y_0 = 0 \mid X = 1, Y = 1) = \frac{\mathrm{PoB}}{p_1}, \qquad \mathrm{PS} := P(Y_1 = 1 \mid X = 0, Y = 0) = \frac{\mathrm{PoB}}{1 - p_0}.$$

(i) The denominators $p_1 = P(Y = 1 \mid X = 1)$ and $1 - p_0 = P(Y = 0 \mid X = 0)$ are observational probabilities, hence identified; PN *is defined iff $p_1 > 0$ and* PS *iff $p_0 < 1$. (ii) With an identified nonzero denominator,* $\mathrm{PN} = \psi_{H/p_1}$ *and* $\mathrm{PS} = \psi_{H/(1-p_0)}$ *with $H = \mathbf{1}\{y_0 = 0, y_1 = 1\}$; scaling by a constant preserves separability, so by Proposition 1* PN *and* PS *are identified iff a marginal is degenerate ($p_0 \in \{0, 1\}$ or $p_1 \in \{0, 1\}$) and are otherwise partially identified. (iii) Their sharp intervals are*

$$\mathrm{PN} \in \left[\max\left\{0, 1 - \tfrac{p_0}{p_1}\right\}, \min\left\{1, \tfrac{1-p_0}{p_1}\right\}\right], \qquad \mathrm{PS} \in \left[\max\left\{0, \tfrac{p_1-p_0}{1-p_0}\right\}, \min\left\{1, \tfrac{p_1}{1-p_0}\right\}\right],$$

which coincide with the experimental-data bounds of Tian and Pearl [8]; under monotonicity ($Y_1 \geq Y_0$), $\mathrm{PN} = 1 - p_0/p_1$ *and* $\mathrm{PS} = (p_1 - p_0)/(1 - p_0)$.

Proof. Under Assumption A1, $X \perp\!\!\!\perp (Y_0, Y_1)$ and $Y = Y_1$ on $\{X = 1\}$, so $P(Y_0 = 0 \mid X = 1, Y = 1) = P(Y_0 = 0, Y_1 = 1)/P(Y_1 = 1)$; PS is analogous. (ii) follows from Proposition 1, since PoB's rectangle contrast is 1 and a degenerate marginal reduces the polytope to a point. (iii) divides the Fréchet range $\mathrm{PoB} \in [\max(0, p_1 - p_0), \min(p_1, 1 - p_0)]$ (Appendix A) by the identified denominators. □

In confounded designs (Theorem 2) the denominators become $P(X = 1, Y = 1)$ and $P(X = 0, Y = 0)$, again observational and identified; they must be positive, and the bounds are the combined-data Tian–Pearl bounds computed by the linear program.

5.3 Learnability

Corollary 3 (No data-only learner for a coupling-sensitive target)**.** *If $\psi(\gamma) \neq \psi(\gamma')$ for some $\gamma, \gamma' \in \Pi(\mu_0, \mu_1)$, then no sequence of estimators based only on data of the regime in Assumption A3 is consistent under both M_γ and $M_{\gamma'}$ of Theorem 1.*

Proof. The data distributions are identical under M_γ and $M_{\gamma'}$, so each $\widehat{\psi}_n$ has the same law under both; it cannot converge in probability to two different constants. □

Proposition 3 (Learnable ⇔ identified ⇔ separable)**.** *For finite outcomes with positive masses, ψ_H is learnable over $\mathcal{M}$ (Definition 4) if and only if H is additively separable.*

Proof. If H is separable, $\psi_H = \sum_i \alpha_i p_i + \sum_j \beta_j q_j$, and replacing p, q by empirical frequencies of the two arms gives a consistent estimator under every model. If H is not separable, Proposition 1 gives two couplings with different values and Corollary 3 applies. □

The obstruction persists with infinite data. Finite-sample uncertainty is layered on top of it: a valid analysis should not let a narrow confidence interval for a single selected coupling hide a wide identification region.

6. THE COUNTERFACTUAL COUPLING AUDIT

6.1 Sharp bounds as transportation programs

For finite supports and score H, define

$$\underline{\psi}_H := \min_{\Gamma \geq 0} \sum_{i,j} h_{ij}\gamma_{ij}, \qquad \overline{\psi}_H := \max_{\Gamma \geq 0} \sum_{i,j} h_{ij}\gamma_{ij} \quad \text{s.t.} \quad \sum_j \gamma_{ij} = p_i,\ \sum_i \gamma_{ij} = q_j. \tag{5}$$

Proposition 4 (Sharpness)**.** *For finite supports, $[\underline{\psi}_H, \overline{\psi}_H]$ is the sharp identification interval of ψ_H over $[M]$, and both endpoints are attained by SCMs empirically equivalent to M. The analogous statement holds for the combined-data program of Theorem 2.*

Proof. The feasible set is nonempty and compact, so both optima are attained; by Theorem 1 (resp. Theorem 2) every feasible point, including the optimizers, is induced by an empirically equivalent SCM. □

Covariates. If a pre-treatment covariate Z with finitely many values is observed, the identified quantities are $A_{zi} = P(Z = z, Y_0 = a_i)$ and $B_{zj} = P(Z = z, Y_1 = b_j)$. The sharp interval is obtained by optimizing over arrays $J_{zij} \geq 0$ with $\sum_j J_{zij} = A_{zi}$ and $\sum_i J_{zij} = B_{zj}$; the stratified program decomposes into one transportation problem per stratum, and its interval is contained in the unconditional one.

6.2 Audit procedure

Counterfactual Coupling Audit (finite outcomes)

1. Estimate the interventional marginals p, q (and, if available, their stratified versions A, B and the observational law for combined designs).
2. Encode the target as a score matrix H; for PN or PS, divide by the identified denominator after checking that it is positive (Proposition 2).
3. Compute the residual $R_{ij} = h_{ij} - h_{i1} - h_{1j} + h_{11}$. If $\max |R_{ij}| \leq 10^{-10}$, the target is identified: report its plug-in value with a standard confidence interval.
4. Otherwise solve (5) (or its stratified or combined-data version) and report the plug-in sharp interval together with a confidence interval from Section 7.
5. If a cross-world assumption is adopted, encode it as a restriction on the feasible set (e.g. forbidden cells for monotonicity, a fixed cell for a known joint frequency), recompute, and report it separately from the empirical information.

Implementation. The released Python package (Appendix C) solves all programs with the HiGHS dual simplex solver [49] through `scipy.optimize.linprog` [50] with default tolerances (primal and dual feasibility 10^{-7}) and sparse constraint matrices. A transportation problem has mn variables and $m + n$ constraints; the largest instance in this paper (169×141 support points, Section 10) solves in about 0.5 s on a laptop CPU.

6.3 Continuous outcomes

For continuous outcomes, $\inf_\gamma$ and $\sup_\gamma$ of $\int h\, d\gamma$ over $\Pi(\mu_0, \mu_1)$ are optimal-transport problems [40]. For threshold events of the individual effect $D = Y_1 - Y_0$, the infimum and supremum of $P(D \leq d)$ over all couplings are given by the Makarov bounds [22, 24, 25]

$$\sup_y \max\{F_1(y) - F_0(y-d), 0\} \leq P(D \leq d) \leq 1 + \inf_y \min\{F_1(y) - F_0(y-d), 0\}, \tag{6}$$

(best possible as infimum and supremum; attainment is discussed case by case in Appendix B), and the extremes of $\mathrm{Var}(D)$ are attained by the comonotone and countermonotone couplings. With empirical marginals, which have finite support, the finite audit applies directly (Section 10).

7. FINITE-SAMPLE INFERENCE

The audit's endpoints are estimated from finite samples. We use two complementary procedures.

Lemma 2 (Stability of the sharp bounds)**.** *Let $V(p,q) = \max_{\Gamma\in\Pi(p,q)}\langle H,\Gamma\rangle$ and $\|H\|_\infty = \max_{ij}|h_{ij}|$. For probability vectors p, p' on m points and q, q' on n points, $|V(p,q) - V(p',q')| \le \|H\|_\infty\big(2\|p-p'\|_1 + \|q-q'\|_1\big)$, and the same holds for the minimum.*

Proof. Let Γ be optimal for (p,q). Rescale row i by p'_i/p_i (or, if $p_i = 0$, put mass p'_i in row i distributed as q); the result Γ'' has row sums p', satisfies $\|\Gamma'' - \Gamma\|_1 \le \|p-p'\|_1$, and has column sums q'' with $\|q''-q'\|_1 \le \delta := \|p-p'\|_1 + \|q-q'\|_1$. Remove mass $(q''_j - q'_j)_+$ from each column in proportion to its entries, collect the removed row totals r (total $\le \delta/2$), and add back $r_i(q'_j - q''_j)_+ / \sum_k (q'_k - q''_k)_+$ in each cell. Row sums remain p', column sums become q', and the change has ℓ_1 norm at most δ. The resulting $\Gamma' \in \Pi(p',q')$ satisfies $\|\Gamma' - \Gamma\|_1 \le 2\|p-p'\|_1 + \|q-q'\|_1$, so $V(p',q') \ge V(p,q) - \|H\|_\infty(2\|p-p'\|_1 + \|q-q'\|_1)$; exchange the roles to conclude. □

Corollary 4 (Consistency of plug-in bounds)**.** *For fixed finite supports, if $\hat p, \hat q$ are the empirical frequencies of the two arms, the plug-in endpoints are consistent for the sharp endpoints at rate $O_P(n^{-1/2})$ under every model in $\mathcal{M}$ (pointwise in the model; by Lemma 2 and $\|\hat p - p\|_1 + \|\hat q - q\|_1 = O_P(n^{-1/2})$).*

Proposition 5 (Projection confidence interval)**.** *Let $\mathcal{B}_A, \mathcal{B}_B$ be boxes such that $P(A \in \mathcal{B}_A,\ B \in \mathcal{B}_B) \ge 1-\alpha$ for the true stratified margins. Let $[\ell, u]$ be the minimum and maximum of $\langle H, J\rangle$ over $J \ge 0$ with $\sum J = 1$, $(\sum_j J_{zij})_{zi} \in \mathcal{B}_A$ and $(\sum_i J_{zij})_{zj} \in \mathcal{B}_B$. Then $P\big([\underline{\psi}_H, \overline{\psi}_H] \subseteq [\ell, u]\big) \ge 1-\alpha$ uniformly over $\mathcal{M}$; in particular $[\ell, u]$ covers the true value of the target with probability at least $1-\alpha$ whatever the true coupling.*

Proof. On the event that the true margins lie in the boxes, every coupling compatible with them is feasible for the outer program. □

We build the boxes from Clopper–Pearson intervals [51] for each cell $P(Z = z, Y_x = y)$, estimated from arm x, with a Bonferroni correction over all cells. Conditionally on the arm sizes, each cell count is binomial under randomization, so the joint box has coverage at least $1-\alpha$ in finite samples for every data-generating distribution, and the projection interval inherits this guarantee; it is conservative. As a less conservative alternative targeting the true parameter rather than the whole set, we report the Imbens–Manski interval $[\hat L - c\,\hat s_L, \hat U + c\,\hat s_U]$, where c solves $\Phi(c + (\hat U - \hat L)/\max(\hat s_L, \hat s_U)) - \Phi(-c) = 1-\alpha$ [33]. It is asymptotically valid when the endpoint estimators are asymptotically normal, i.e. away from kinks of the $\max$ / $\min$ in the bounds; Stoye [34] discusses uniform refinements. For continuous outcomes we use bootstrap standard errors of the LP endpoints inside the Imbens–Manski formula; this is a heuristic without a proven coverage guarantee, in particular where the optimal LP basis changes, and Fan and Park [13] give formal theory for Makarov-type bounds. Supplementary file `inference_claims.md` lists every coverage statement with its regime and justification.

8. NUMERICAL DEMONSTRATIONS

8.1 Binary probability of benefit

Let $Y \in \{0,1\}$ with $p_0 = P(Y_0 = 1) = 0.30$ and $p_1 = P(Y_1 = 1) = 0.60$. Every coupling is fixed by $q = P(Y_0 = 1, Y_1 = 1) \in [\max(0, p_0 + p_1 - 1), \min(p_0, p_1)] = [0, 0.30]$. The average effect is $p_1 - p_0 = 0.30$ for every q, whereas $P(Y_0 = 0, Y_1 = 1) = p_1 - q$, $P(Y_0 = 1, Y_1 = 0) = p_0 - q$ and $P(Y_0 \ne Y_1) = p_0 + p_1 - 2q$.

Derivation of the variance of the individual effect. For $D = Y_1 - Y_0 \in \{-1, 0, 1\}$, $D^2 = \mathbf{1}\{Y_0 \ne Y_1\}$, so $\mathbb{E}D^2 = p_0 + p_1 - 2q$ and $\mathbb{E}D = p_1 - p_0$. Hence

$$\mathrm{Var}(D) = p_0 + p_1 - 2q - (p_1 - p_0)^2, \tag{7}$$

affine and decreasing in q. With $p_0 = 0.3, p_1 = 0.6$: $q = 0.30$ gives $0.90 - 0.60 - 0.09 = 0.21$; $q = 0$ gives $0.90 - 0.09 = 0.81$; independence ($q = p_0 p_1 = 0.18$) gives $0.90 - 0.36 - 0.09 = 0.45$. Table III reports all values; the LP of (5) reproduces them to machine precision.

TABLE III. Coupling audit for $p_0 = 0.30$ and $p_1 = 0.60$. "Independence" is the value under the added selector $Y_0 \perp\!\!\!\perp Y_1$ ($q = 0.18$), not an estimate.

Target	Identified?	Sharp lower	Independence	Sharp upper
$\mathbb{E}(Y_1 - Y_0)$	yes	0.30	0.30	0.30
$P(Y_0 = 0, Y_1 = 1)$ (benefit)	no	0.30	0.42	0.60
$P(Y_0 = 1, Y_1 = 0)$ (harm)	no	0.00	0.12	0.30
$P(Y_0 \neq Y_1)$	no	0.30	0.54	0.90
$\mathrm{Var}(Y_1 - Y_0)$, eq. (7)	no	0.21	0.45	0.81
PN (Proposition 2)	no	0.50	0.70	1.00
PS (Proposition 2)	no	0.43	0.60	0.86

8.2 Gaussian sensitivity slice versus the full coupling set

Let $Y_0 \sim \mathcal{N}(0, 1)$ and $Y_1 \sim \mathcal{N}(0.5, 1)$. The bivariate-normal family with correlation ρ keeps both marginals fixed, with $\mathrm{Var}(Y_1 - Y_0) = 2(1 - \rho)$ and $P(Y_1 > Y_0) = \Phi(0.5/\sqrt{2(1 - \rho)})$ (Appendix B). Table IV compares the whole Gaussian range with the sharp interval over the complete coupling set, computed from the Makarov bounds (6) and the Fréchet–Hoeffding couplings, and cross-checked by the discretized transportation program with 150 quantile atoms per marginal (deviations of at most 0.007 for the probabilities; 0.034 for the upper variance bound, because equally weighted quantile atoms understate tail variance).

TABLE IV. Gaussian-copula range versus sharp bounds over all couplings for $Y_0 \sim \mathcal{N}(0, 1)$, $Y_1 \sim \mathcal{N}(0.5, 1)$. Gaussian column: the open range of the family $\rho \in (-1, 1)$ (endpoints not attained). Sharp column: infimum and supremum over all couplings, written as the closed interval they span; an endpoint marked * is attained by an explicit coupling (comonotone or countermonotone, Appendix B); for the unmarked endpoint we claim only that it is the infimum or supremum. The last column is (Gaussian width)/(sharp width).

Target	Gaussian family, $\rho \in (-1, 1)$	Sharp, all couplings	Width ratio
$\mathrm{Var}(Y_1 - Y_0)$	$(0,\ 4)$	$[0^*,\ 4^*]$	1.00
$P(Y_1 > Y_0)$ (benefit)	$(0.599,\ 1)$	$[0.197,\ 1^*]$	0.50
$P(Y_1 < Y_0)$ (harm)	$(0,\ 0.401)$	$[0^*,\ 0.803]$	0.50
$P(Y_1 - Y_0 > 1)$	$(0,\ 0.401)$	$[0^*,\ 0.803]$	0.50
$P(Y_1 - Y_0 > -1)$	$(0.773,\ 1)$	$[0.547,\ 1^*]$	0.50

Proposition 6 (Gaussian half-width)**.** *Let $Y_0 \sim \mathcal{N}(m_0, \sigma^2)$, $Y_1 \sim \mathcal{N}(m_1, \sigma^2)$, $\delta = m_1 - m_0$ and $d \neq \delta$, and write $c = (\delta - d)/(2\sigma)$. Over all couplings of the two marginals: (i) if $\delta > d$, the infimum of $P(Y_1 - Y_0 > d)$ is $2\Phi(c) - 1$ and its maximum is 1, attained by the comonotone coupling; (ii) if $\delta < d$, its minimum is 0, attained by the comonotone coupling, and its supremum is $2\Phi(c)$. Over the Gaussian family the probability ranges over the open intervals $(\Phi(c), 1)$ and $(0, \Phi(c))$ respectively. Hence the Gaussian family spans exactly half of the length of the sharp range for every $d \neq \delta$ and every σ. For $\mathrm{Var}(Y_1 - Y_0)$ the minimum 0 and maximum $4\sigma^2$ over all couplings are attained by the comonotone and countermonotone couplings, and the Gaussian range $(0, 4\sigma^2)$ has the same closure.*

The proof is in Appendix B. Throughout, "sharp interval" for a continuous target means the closed interval spanned by the infimum and supremum over all couplings; we state explicitly which endpoints are attained. For finite supports (all binary examples and the empirical NSW margins) both endpoints are attained (Proposition 4). The practical message is direct: the Gaussian slice misses the most pessimistic half of the benefit range. A system that reports $P(Y_1 > Y_0) = 0.638$ because it silently used $\rho = 0$ has injected an assumption, and even a sensitivity analysis restricted to ρ would claim that more than 59.9% of units benefit, whereas couplings compatible with the data bring the share arbitrarily close to 19.7% (Figure 1).

9. SIMULATION WITH EMPIRICALLY EQUIVALENT STRUCTURAL CAUSAL MODELS

Design. A pre-treatment covariate $Z \in \{0, 1, 2, 3\}$ has $P(Z) = (0.3, 0.3, 0.2, 0.2)$; the stratum margins are $p_0(Z) = (0.15, 0.30, 0.45, 0.40)$ and $p_1(Z) = (0.55, 0.60, 0.50, 0.70)$; the action is randomized with $P(X = 1) = 1/2$; and $V, V' \sim U(0, 1)$ are independent of everything else. Three SCMs share every observational and interventional distribution

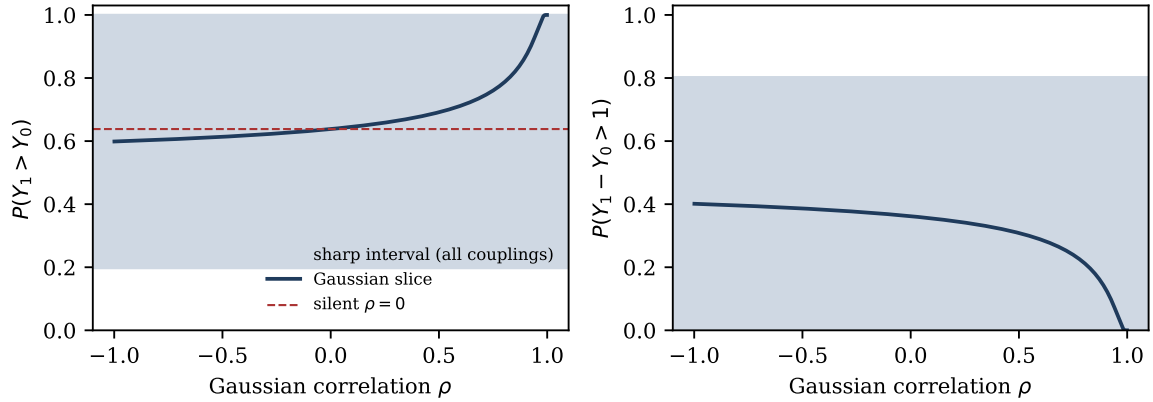


Fig. 1. Gaussian sensitivity path (line) versus sharp interval over all couplings (shaded) for $P(Y_1 > Y_0)$ (left) and $P(Y_1 - Y_0 > 1)$ (right). The dashed line is the value under silent independence.

of (Z, X, Y) but differ in their structural equations:

$$\begin{aligned} M_{\text{co}}: &\quad Y_0 = \mathbf{1}\{V < p_0(Z)\}, \quad Y_1 = \mathbf{1}\{V < p_1(Z)\} \\ M_{\text{ind}}: &\quad Y_0 = \mathbf{1}\{V < p_0(Z)\}, \quad Y_1 = \mathbf{1}\{V' < p_1(Z)\} \\ M_{\text{anti}}: &\quad Y_0 = \mathbf{1}\{V < p_0(Z)\}, \quad Y_1 = \mathbf{1}\{1 - V < p_1(Z)\} \end{aligned}$$

with $Y = Y_X$. Their probabilities of benefit are 0.280, 0.405 and 0.565, which are, respectively, the lower endpoint, an interior point and the upper endpoint of the covariate-conditional sharp interval $[0.280, 0.565]$ (the unconditional interval is $[0.280, 0.585]$). For each SCM and $n \in \{200, 1000, 5000, 25000\}$ we draw 1,000 data sets (seed 20261018) and compute: the *independence plug-in* $\sum_z \hat{P}(z)\hat{p}_1(z)\{1 - \hat{p}_0(z)\}$; the plug-in conditional sharp interval; the 95% projection interval with Bonferroni Clopper–Pearson boxes over the 16 cells; and the unconditional Imbens–Manski interval. The independence plug-in combines two distinct ingredients. The estimation step, stratum-specific arm frequencies $\hat{p}_x(z)$, is a T-learner in the sense of Künzel et al. [45]: it estimates each treatment-specific outcome function separately and identifies only the margins $p_0(z)$ and $p_1(z)$. A T-learner by itself makes no assumption about the joint law of (Y_0, Y_1). The product $p_1(z)\{1 - p_0(z)\}$ instead requires the additional cross-world assumption $Y_0 \perp\!\!\!\perp Y_1 \mid Z$, which is not implied by the estimator and is not learned from the data. We use this estimator because combining fitted treatment-specific models under conditional independence is a common way to turn CATE-type models into unit-level benefit statements.

Results. (i) *Indistinguishability*: a χ^2 test comparing the (Z, X, Y) counts of independent samples from M_{co} and M_{anti} rejected at level 5% in 4.3%, 5.3%, 4.6% and 5.0% of 1,000 replications for the four sample sizes, i.e. at the nominal 5% rate within Monte Carlo error (standard error about 0.7 percentage points), as Corollary 3 predicts. (ii) *Non-vanishing bias*: the independence plug-in concentrates at 0.405 for all three models; its bias is $+0.125$ under M_{co} and -0.160 under M_{anti} at every n, and its RMSE falls to 0.004 only under M_{ind} (Table V and figure 2). (iii) *Bounds*: the plug-in sharp interval converges to $[0.280, 0.565]$ under all three models. (iv) *Coverage*: the projection interval covered the entire identified set, and hence the true value, in 100% of replications under every model and n (it is conservative: mean width 0.371 at $n = 25{,}000$ versus a set width of 0.285); the Imbens–Manski interval covered the true value in 94.4%–100% of replications, with the lowest coverage at the lower endpoint (M_{co}), as expected for an interval calibrated for the least favourable endpoint. With 1,000 replications the Monte Carlo standard error of a coverage estimate near 0.95 is about 0.007. For M_{co} the 95% Wilson intervals of the four coverage estimates are $[0.932, 0.960]$, $[0.935, 0.962]$, $[0.944, 0.969]$ and $[0.928, 0.957]$; all contain 0.95, so the observed 94.4% at $n = 25{,}000$ is compatible with nominal coverage at the least favourable endpoint but is not evidence of over- or exact coverage there. At interior and upper-endpoint models the interval is conservative (coverage ≥ 0.978). For the projection interval, 1,000 of 1,000 replications cover the set (Wilson interval $[0.996, 1]$), consistent with its finite-sample guarantee; we make no claim that either procedure has exact nominal coverage.

A causal-AI decision example. Consider a targeting module that offers the action to a stratum when its probability of benefit exceeds 0.35. The stratum-level sharp intervals are $[0.40, 0.55]$, $[0.30, 0.60]$, $[0.05, 0.50]$ and $[0.30, 0.60]$, while the independence values are 0.468, 0.420, 0.275 and 0.420. An independence-based module treats strata 0, 1 and 3 (80% of the population). Under M_{co} only stratum 0 meets the threshold and under M_{anti} all four do, so the independence rule is correct under neither extreme model, and no amount of data can tell which model holds. A coupling-aware module instead reports

TABLE V. Monte Carlo results (1,000 replications per cell). PoB: probability of benefit. Indep.: independence plug-in (T-learner margins combined under the added assumption $Y_0 \perp\!\!\!\perp Y_1 \mid Z$). Proj.: 95% projection interval. IM: Imbens–Manski interval. Monte Carlo standard error of a coverage proportion near 0.95: about 0.007; Wilson intervals in the text.

n	Model	True PoB	Indep. mean	Indep. RMSE	Mean $[\hat{L}_Z, \hat{U}_Z]$	Proj. cover / width	IM cover
200	M_{co}	0.280	0.404	0.131	$[0.288, 0.546]$	1.000 / 0.828	0.948
200	M_{ind}	0.405	0.407	0.046	$[0.291, 0.548]$	1.000 / 0.830	1.000
200	M_{anti}	0.565	0.405	0.166	$[0.289, 0.546]$	1.000 / 0.828	0.978
1000	M_{co}	0.280	0.405	0.127	$[0.282, 0.563]$	1.000 / 0.652	0.950
1000	M_{ind}	0.405	0.405	0.019	$[0.281, 0.561]$	1.000 / 0.650	1.000
1000	M_{anti}	0.565	0.405	0.161	$[0.282, 0.563]$	1.000 / 0.652	0.999
5000	M_{co}	0.280	0.405	0.126	$[0.280, 0.565]$	1.000 / 0.465	0.958
5000	M_{ind}	0.405	0.405	0.009	$[0.280, 0.565]$	1.000 / 0.465	1.000
5000	M_{anti}	0.565	0.406	0.160	$[0.281, 0.565]$	1.000 / 0.465	1.000
25000	M_{co}	0.280	0.405	0.125	$[0.280, 0.565]$	1.000 / 0.371	0.944
25000	M_{ind}	0.405	0.405	0.004	$[0.280, 0.565]$	1.000 / 0.371	1.000
25000	M_{anti}	0.565	0.405	0.160	$[0.280, 0.565]$	1.000 / 0.371	1.000

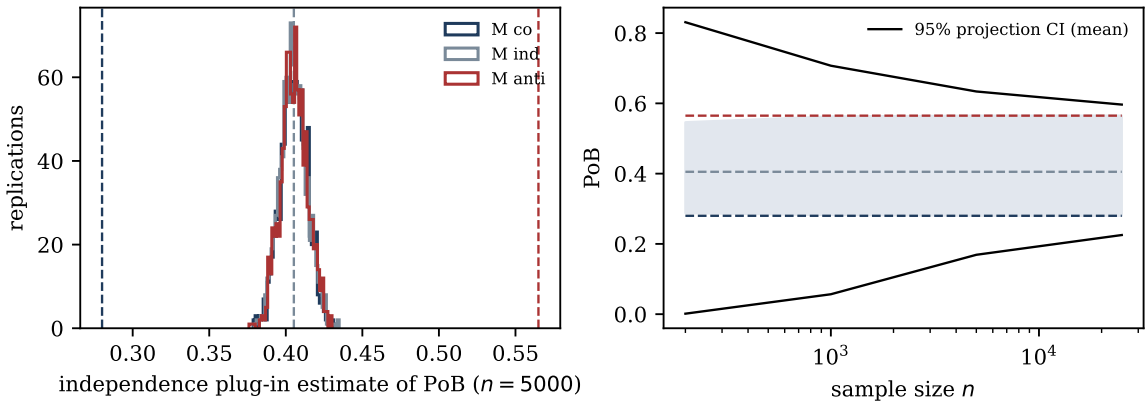


Fig. 2. Left: sampling distribution of the independence plug-in at $n = 5{,}000$ under the three SCMs (dashed lines: true values). Right: mean plug-in conditional sharp interval (shaded), true values of the three SCMs (dashed) and mean 95% projection interval (black) as n grows.

that stratum 0 qualifies robustly, that strata 1–3 qualify only under added assumptions, and which assumption would be needed.

10. APPLICATION TO PUBLIC RANDOMIZED TRIALS

Data. (a) Thornton [35] randomized cash incentives for learning HIV test results in rural Malawi. We use the public replication extract distributed with [52] (Python package `causaldata` 0.1.5), with $X = 1$ if any incentive was offered and $Y = 1$ if the participant obtained the result; after dropping records with missing X or Y, $n = 2{,}834$ ($n_1 = 2{,}211$, $n_0 = 623$). Tertiles of distance to the testing centre serve as a pre-treatment covariate. (b) The National Supported Work experiment [36, 53], Dehejia–Wahba subsample from the same source, $n = 445$ ($n_1 = 185$, $n_0 = 260$): X is assignment to training and Y is 1978 earnings, analysed both as employment ($Y > 0$) and as earnings. Prior employment (1975 earnings > 0) and absence of a high-school degree form four covariate strata.

Findings and decision implications. In the Thornton trial the average effect of the incentive, 45.1 percentage points, is identified and precisely estimated; this is all that a population-level cost-effectiveness calculation needs. Statements about *who* was moved by the incentive are not identified. Between 45.1% and 66.1% of participants would learn their status with an incentive and not without one, and among incentivized participants who learned their status, between 57% and 84% would not have done so without the incentive (PN). The upper part of these ranges corresponds to couplings in which up to 21% of participants would learn their status without an incentive but not with one. Monotonicity ("an incentive never deters") removes these couplings and point-identifies $\mathrm{PNS} = 0.451$ and $\mathrm{PN} = 0.571$; it is plausible here, but it is an assumption, and the audit makes it visible. Conditioning on distance did not narrow the interval because the stratum margins

TABLE VI. Coupling audit of two public randomized trials. Intervals are plug-in sharp intervals; "Indep." and "Monot." are point values under the corresponding added assumption. CIs are 95%: Wald for identified targets, projection (Bonferroni Clopper–Pearson) or Imbens–Manski (IM) for partially identified ones. Projection intervals have finite-sample coverage; IM intervals are asymptotic; the earnings intervals plug 200-replication bootstrap standard errors into the IM formula and are heuristic (no proven coverage, Section 7).

Data / target	Identified?	Sharp interval	Indep.	Monot.	95% CI (method)
(a) Thornton HIV-results trial, $n = 2{,}834$*;* $\hat{p}_1 = 0.789$*,* $\hat{p}_0 = 0.339$					
Average effect	yes	0.451	0.451	0.451	$[0.410,\ 0.491]$ (Wald)
Benefit (PNS)	no	$[0.451,\ 0.661]$	0.522	0.451	$[0.416,\ 0.693]$ (IM); $[0.386,\ 0.704]$ (proj.)
Harm	no	$[0.000,\ 0.211]$	0.071	0.000	$[0.000,\ 0.231]$ (proj.)
PN	no	$[0.571,\ 0.838]$	0.661	0.571	$[0.502,\ 0.915]$ (proj.)
PS	no	$[0.681,\ 1.000]$	0.789	0.681	$[0.626,\ 1.000]$ (proj.)
Benefit given distance tertile	no	$[0.451,\ 0.661]$	–	–	$[0.299,\ 0.750]$ (proj.)
(b) NSW, employment in 1978, $n = 445$*;* $\hat{p}_1 = 0.757$*,* $\hat{p}_0 = 0.646$					
Average effect	yes	0.111	0.111	0.111	$[0.026,\ 0.195]$ (Wald)
Benefit (PNS)	no	$[0.111,\ 0.354]$	0.268	0.111	$[0.039,\ 0.403]$ (IM); $[0.000,\ 0.424]$ (proj.)
Harm	no	$[0.000,\ 0.243]$	0.157	0.000	$[0.000,\ 0.321]$ (proj.)
PN	no	$[0.146,\ 0.468]$	0.354	0.146	$[0.000,\ 0.625]$ (proj.)
PS	no	$[0.313,\ 1.000]$	0.757	0.313	$[0.000,\ 1.000]$ (proj.)
Benefit given 4 strata	no	$[0.097,\ 0.353]$	–	–	$[0.000,\ 0.566]$ (proj.)
(c) NSW, 1978 earnings (in \$1,000), LP on empirical margins (169×141 *support points)*					
Average effect	yes	1.794	1.794	–	$[0.464,\ 3.125]$ (Wald, bootstrap SE)
$P(Y_1 > Y_0)$	no	$[0.132,\ 0.757]$	0.527	–	$[0.067,\ 0.809]$ (IM)
$P(Y_1 < Y_0)$	no	$[0.000,\ 0.646]$	0.387	–	$[0.000,\ 0.690]$ (IM)
$P(Y_1 - Y_0 > 1)$	no	$[0.069,\ 0.692]$	0.474	–	$[0.012,\ 0.751]$ (IM)
$\mathrm{Var}(Y_1 - Y_0)$	no	$[7.70,\ 142.32]$	91.52	–	$[0.00,\ 179.23]$ (IM)

are similar, and it widened the projection interval because more cells are estimated. In the NSW experiment the 11.1-point employment effect is identified, but the data cannot exclude that training *reduced* employment for up to 24.3% of trainees (32.1% at the upper projection limit); a monotonicity assumption would set this to zero, and in a job-training context where lock-in effects are possible that assumption needs a substantive argument. The stratified plug-in lower bound (0.097) is slightly below the unconditional one (0.111) only because the stratified estimate re-weights arms by the pooled covariate distribution; in the population the stratified interval is contained in the unconditional one. For earnings, the average gain of \$1,794 is identified, but the share of participants whose earnings rose ranges from 13.2% to 75.7%, the share whose earnings fell from 0% to 64.6%, and the standard deviation of individual gains from \$2,800 to \$11,900. An AI decision or recourse module that tells an applicant "training will raise your earnings" can support the average statement from these data, but not the individual one; the independence values (52.7% gain, 38.7% lose), which such a module would report if it silently assumed independence, are an assumption rather than a finding.

11. IMPLICATIONS FOR CAUSAL AI SYSTEMS AND A COUPLING STRESS TEST

11.1 Separate learned marginals from assumed cross-world dependence

A causal pipeline may estimate structural equations, latent variables or conditional interventional laws with flexible machine learning. None of that changes the logical point: two fitted systems with the same validated interventional content can imply different cross-world couplings, and hence different joint counterfactuals, with identical predictive and interventional fit. A reporting interface should expose three layers: (1) the empirically learned single-world quantities; (2) the cross-world restriction or selector (rank invariance, monotonicity, independence, a copula, a mechanism); and (3) the resulting value or identification interval with its confidence interval.

11.2 C-STRESS: a coupling stress test for AI reasoners

Systems can fail because they do not know a causal rule, because they execute it incorrectly, or because they confidently answer a query that the information does not determine. These failure modes should not be collapsed into one accuracy score. Consensus does not create missing cross-world information: if every agent shares the same unstated coupling assumption, agreement can increase without the answer becoming warranted, which is consistent with the persistence of shared errors under consensus reported in [20]. We therefore release C-STRESS, an item bank, answer schema, scorer and runner implementing a three-part protocol, and report a pilot evaluation in Section 11.3.

Item families. The 240 items are generated deterministically (seed 20261018) over five application domains (medicine, education, finance, marketing, public health):

F1 (40 items). Identified targets (binary average effect; Gaussian $\mathbb{E}(Y_1 - Y_0)$). Gold: point.

F2 (80). Coupling-sensitive targets given marginals only: benefit, harm, disagreement, PN, PS, variance of the individual effect (binary), $P(Y_1 > Y_0)$ and $\mathrm{Var}(Y_1 - Y_0)$ (Gaussian). Gold: sharp interval.

F3 (80, in 40 pairs). The same targets with a supplied assumption (independence, monotonicity, a known joint frequency, or a Gaussian correlation). Pair members share marginals and differ only in the assumption. Gold: point.

F4 (20). Identified targets with an irrelevant supplied assumption; the answer must not change. Gold: point.

F5 (20). PN from combined observational and experimental data with confounding. Gold: Tian–Pearl interval (Theorem 2).

Example items (verbatim stems). *F2:* "A large randomized controlled trial assigned patients at random to receive a new drug or not. Among patients assigned to receive a new drug, 62% ended up recovering within 30 days. Among patients assigned not to receive it, 29% ended up recovering within 30 days. Treat these percentages as exact population values. Question: What is the variance of the individual treatment effect $Y(1) - Y(0)$, where $Y(x)$ is 1 if the outcome occurs?" Gold: not identified, $[0.221, 0.801]$ by (7). *F3 pair:* the same stem with "Assume that each unit's outcome under treatment is statistically independent of its outcome under no treatment" (gold 0.441) versus "Assume monotonicity: the treatment never prevents the outcome for any unit" (gold 0.221). *F4:* an average-effect question with the independence sentence added; the gold answer is the same as without it.

Answer schema and implementation. The system prompt (released verbatim) asks for a JSON object with fields `identified` (Boolean), `answer_type` (point or interval), `value`, `lower`, `upper`, `cross_world_assumption` (one of none, independence, monotonicity, rank preservation, gaussian, joint specified, other) and a short explanation. It does not instruct the model that items may be unidentified. The runner calls any OpenAI-compatible endpoint (e.g. a local Ollama or vLLM server) with temperature 0 and at most 600 output tokens, strips reasoning traces, extracts the first JSON object and logs raw outputs.

Scoring rubric. For item k with gold g_k and parsed response r_k:

- **S1, identification recognition** (all items): $S1_k = \mathbf{1}\{r_k.\texttt{identified} = g_k.\texttt{identified}\}$, where items with a supplied assumption (F3, F4) count as identified relative to the stated class.
- **S2, assumption provenance** (F2, F3, F5): for F2/F5, $S2_k = 1$ if the response is an interval or declares a cross-world assumption other than "none"; an unsupported point answer scores 0 even if it coincides with a compatible value. For F3, $S2_k = 1$ iff a point answer declares the supplied assumption.
- **S3, conditional computation**: the response is scored against the gold answer *conditional on the provenance it declares*: an interval against the sharp interval, a point with a declared assumption against the value under that assumption (tolerance ± 0.01 for probabilities, $\pm 2\%$ for variances above 1). S3 is not applicable to unsupported point answers or unparseable assumptions.
- **Strict score**: all applicable components equal 1. **Pair consistency**: both members of an F3 pair have $S2 = S3 = 1$, i.e. the answer moves exactly when the assumption moves.

Proportions are reported with Wilson 95% intervals.

How to read S2 and S3 together. S2 measures only whether a response *declares* provenance: on F2 and F5 it credits any interval and any named cross-world assumption, whether or not that assumption was used, is plausible, or yields the reported number. It therefore cannot distinguish an honest declaration from a label attached to an unrelated computation, and a high S2 alone is not evidence of identification awareness. S3 supplies the missing check by recomputing the gold answer under the declared provenance, but it scores numbers and not reasoning, so it can credit a correct number reached by an invalid argument and is not applicable to undeclared point answers. Only the conjunction of S1, S2 and S3 (the strict score) certifies a coupling-aware answer, and S2 should never be reported or interpreted in isolation. Section 11.3 shows a model for which S2 is high (0.84) while S3 reveals that the declared assumption was not the one used.

Validation of the rubric. To check that the scores separate the intended failure modes, we scored five deterministic scripted responders (programs, not language models): an oracle; an oracle with 20% arithmetic slips; a "silent independence" responder that always gives the independence point value without declaring it; a "declared independence" responder that declares independence but ignores supplied assumptions; and an over-cautious responder that always returns the no-assumption interval. Table VII shows the intended pattern: arithmetic slips lower only S3; silent selection drives S2 to 0 while S3 stays perfect on the 60 items where it is applicable; ignoring supplied assumptions is caught by S2 and pair consistency; over-caution is caught by S1.

TABLE VII. C-STRESS rubric validation with scripted responders (240 items; S2 applies to 180 items).

Scripted responder	S1	S2	S3	Strict	Pair consistency
Oracle	1.00	1.00	1.00	1.00	1.00
Oracle with 20% arithmetic slips	1.00	1.00	0.81	0.81	0.62
Silent independence	0.58	0.00	1.00	0.25	0.00
Declared independence	0.58	0.72	1.00	0.38	0.00
Always interval (over-cautious)	0.42	0.56	1.00	0.42	0.00

11.3 Pilot evaluation with open-source language models

Setup. We ran all 240 C-STRESS items once on three open-weight instruction-tuned models without explicit reasoning traces: Qwen2.5-7B-Instruct, Llama-3.1-8B and Gemma-2-9B, served by Ollama with the default model tags on a single GPU in Google Colab (October 2026). Decoding used temperature 0 and at most 600 output tokens, with the system prompt and parser of Section 11.2 unchanged; the prompt does not mention identification or couplings beyond the definition of the JSON field. Mean latency was 3.5–4.5 s per item. Raw outputs, parsed answers and per-item scores are released. Rescoring the raw outputs with the released scorer reproduces every reported number, and rescaling answers given as percentages changed no score.

TABLE VIII. C-STRESS pilot (240 items per model, one run, temperature 0). Proportions with Wilson 95% intervals. S2 applies to 180 items; the number of items on which S3 is applicable is given in parentheses.

Model	Parse	S1	S2	S3 (applicable)	Strict
Qwen2.5-7B-Instruct	0.98	0.68 [0.62, 0.74]	0.57 [0.49, 0.64]	0.48 [0.41, 0.55] (182)	0.32 [0.26, 0.38]
Llama-3.1-8B	1.00	0.62 [0.55, 0.68]	0.84 [0.78, 0.89]	0.35 [0.29, 0.42] (220)	0.30 [0.25, 0.36]
Gemma-2-9B	1.00	0.59 [0.52, 0.65]	0.28 [0.22, 0.35]	0.63 [0.53, 0.71] (110)	0.29 [0.23, 0.35]

Results by family. Strict scores separate the item families sharply. On identified targets all models are close to perfect: F1 1.00, 0.95 and 1.00 (Qwen, Llama, Gemma, in that order) and F4 1.00 for all three, so an irrelevant cross-world assumption never changed a correct average-effect answer. On coupling-sensitive targets given only the marginals (F2) and on PN from combined data (F5), the strict score is 0.00 for every model: no model returned the correct sharp interval, or a correctly computed answer under a declared assumption together with a correct identification judgement, on any of these 100 items. With a supplied assumption (F3) strict scores were 16/80, 14/80 and 9/80, and pair consistency was 0.00 for all three models: in no F3 pair did both answers track the stated assumption correctly.

Anatomy of the F2 answers. Table IX classifies the point answers to F2 items. Recognition that the target is not identified was rare (S1 on F2: 0.20, 0.11 and 0.01). Gemma gave a point value without any declared assumption on 79 of 80 items. Llama declared "independence" on every F2 item, which earns S2 credit, but only 5 of its 71 point values equal the independence value; its declared provenance does not describe its computation, and S3 on F2 (0.06) exposes this. Qwen labelled 8 items (all variance targets) as intervals but gave no numerical bounds. The most common identifiable pattern is the monotonicity value (e.g. benefit $= p_1 - p_0$), which the models never declared as an assumption. Between 9 and 21 point answers per model lie *outside* the sharp identification set, i.e. they are incompatible with every SCM that fits the stated data. F5 showed the same split: Qwen recognized non-identification on 19 of 20 items but gave correct bounds on none; Llama and Gemma gave point values throughout.

TABLE IX. Point answers to the 80 F2 items, matched (tolerance ± 0.01) to the value implied by a cross-world selector. "Outside" counts values incompatible with the stated marginals under every coupling.

Model	Point answers	= independence	= monotone/rank	= average effect	Other	Outside sharp set
Qwen2.5-7B-Instruct	61	6	21	3	31	9
Llama-3.1-8B	71	5	28	5	33	15
Gemma-2-9B	79	9	25	15	30	21

Scorer audit and parser sensitivity. We manually audited a stratified sample of 60 responses (20 per model, all five families, all parse failures eligible; released as `pilot_scorer_audit.csv`). In all 60 the scores applied the rubric as specified. Five cases were debatable rather than wrong: in four, S3 credited a correct number given with an incorrect explanation (e.g. PS under independence equals $P(Y_1 = 1)$, so a naive answer is numerically right), and in one a correct value under the supplied assumption was penalised because the model declared a different assumption. The audit also

showed that 7 of the 20 F5 items have the trivial gold interval $[0, 1]$. Of the six unparseable outputs, one was truncated by a repetition loop and five (all Qwen) wrote arithmetic expressions inside the JSON; evaluating those expressions changes no S3 or strict score and raises Qwen's S1 by 0.01 and S2 by 0.03.

Interpretation and limits. All statements in this subsection concern the three evaluated models under the stated conditions: quantized 7–9B instruction-tuned models without reasoning traces, one system prompt that does not mention identification, temperature 0, one run, and the 240 generated items. For these models, aggregate accuracy would hide the result: strict scores of about 0.3 arise from near-perfect performance on identified items combined with no correct answer on non-identified ones. Within this scope the pilot illustrates the distinction between reasoning errors and identification errors, and it shows why S2 cannot be read alone: Llama's S2 of 0.84 reflects declared, not used, assumptions, as its S3 of 0.06 on F2 reveals. The pilot does not estimate how often the failure occurs in other model families, larger models, reasoning-trace models, other prompts (including prompts that warn about identification) or other item distributions, and we draw no conclusion about language models in general.
This protocol complements CLadder, Corr2Cause and CounterBench [17–19]: those benchmarks test computation from fully specified premises, while C-STRESS asks whether a system recognizes when a joint counterfactual is not determined by weaker information.

11.4 Fairness and recourse

Counterfactual fairness asks whether a prediction for the same individual would remain unchanged under a counterfactual change of a protected attribute [1]. Such an assessment is tied to a causal model of how the individual's other variables change across worlds; Russell et al. [2] integrate several counterfactual assumptions for this reason. Theorem 1 explains why the robustness question is fundamental: unless additional structure restricts the coupling, empirically equivalent representations support different same-unit comparisons. Causal recourse distinguishes feature perturbations from interventions that propagate through a causal system [3, 4]. If a recourse objective only asks for an expected outcome under a proposed intervention, the relevant marginal suffices; if it compares a given individual's outcomes across alternative actions, cross-world structure becomes relevant and should be audited.

12. WHAT ADDITIONAL STRUCTURE CAN IDENTIFY A COUPLING?

Real applications may possess information that narrows $\Pi(\mu_0, \mu_1)$:

Mechanistic structural knowledge. A justified latent mechanism privileges a mapping from exogenous states to outcomes; the coupling is then supplied by the mechanism.

Monotonicity or no-harm restrictions. For binary outcomes, $Y_1 \geq Y_0$ removes a cell of the coupling table and identifies PNS, PN and PS [8]; in the audit it is a forbidden cell.

Rank or transport restrictions. Rank-preserving or optimal-transport constructions select cross-world matchings [26].

Parametric dependence models. A copula family restricts the feasible set [54]; Proposition 6 shows how much a Gaussian restriction can remove.

Covariates and combined data. Pre-treatment covariates narrow the interval (Sections 9 and 10); observational data under confounding add constraints (Theorem 2).

Additional designs. Repeated outcomes, crossover designs, proxies or longitudinal structure may carry information about latent dependence beyond single-action marginals.

The audit accommodates all cases by replacing the transport polytope with the subset consistent with the added assumption, and the reporting principle is that this restriction is labelled separately from the empirical marginals it augments.

13. LIMITATIONS

First, the main theorems use a binary action. More actions lead to the multi-marginal version of the same construction ($U = (Y_x)_{x \in \mathcal{A}}$, $g(x, U) = U_x$); the corresponding audit is a multi-marginal transportation program whose size grows exponentially in $|\mathcal{A}|$. Second, randomization (Assumption A1) makes observational data uninformative about the coupling (Lemma 1); Theorem 2 removes this restriction for designs that also contain experimental margins, but purely observational designs with unmeasured confounding leave even the marginals partially identified [41], and we do not treat them. Third, the separability theorem covers expectation-type targets; nonlinear functionals can be audited with Corollary 1 but may

need non-linear optimization. Fourth, our model class is deliberately weak. Graphical identification theory identifies richer counterfactual queries in stronger classes [48]; our conclusions are relative to the stated classes. Fifth, the projection interval is conservative, especially with many strata (Section 10), and the Imbens–Manski interval relies on asymptotic normality away from kinks; continuous-outcome inference uses bootstrap standard errors without a formal uniformity guarantee. Sixth, the language-model pilot is small (three quantized 7–9B models, one prompt, one run); larger models, reasoning-trace models and alternative prompts, including prompts that warn about identification, may behave differently, and the S2 score credits a declared assumption even when the computation does not use it, which is why S2 must be read together with S3; S3 scores numbers, not reasoning, and can credit a numerically correct answer reached by an invalid argument. Finally, we do not claim that independence, comonotonicity, optimal transport or any other selector is wrong; a selector can be useful when justified or used transparently for sensitivity analysis. The results are provenance statements: for a coupling-sensitive target, a point value contains information not supplied by the two interventional marginals.

14. CONCLUSION

A structural causal model determines joint counterfactuals once its latent structure is fixed, but observational and interventional data do not fix that structure. In a randomized two-action system every coupling of the two marginals is realized by an empirically equivalent SCM, and in combined observational–experimental designs the realizable joint laws form an explicit polytope. For finite outcomes the boundary of what can be learned is simple: an expectation-type target is learnable exactly when its score is additively separable across worlds. Average and conditional average effects pass this test; probabilities of benefit, probabilities of causation and the dispersion of individual effects do not. The appropriate computational response is a sharp interval, reported with a projection confidence interval (finite-sample coverage of the identified set) or an Imbens–Manski interval (asymptotic coverage of the target; heuristic when bootstrap standard errors are used for continuous outcomes), and an explicit statement of which cross-world restriction would turn it into a point. The numerical results show that parametric sensitivity families can miss half of that interval, that a plug-in estimator that adds cross-world independence remains biased under indistinguishable models, and that real trials support informative but wide individual-level conclusions. For AI evaluation, the same principle yields a stress test that separates identification recognition from assumption selection and numerical reasoning. In a pilot limited to three open-source 7–9B models, one prompt and one run, these models computed identified items almost perfectly and gave no correct coupling-aware answer on non-identified ones; whether this holds for other models, scales or prompting strategies remains open.

A. BINARY COUPLING FORMULAS

For $Y_0, Y_1 \in \{0,1\}$ with $p_0 = P(Y_0 = 1)$, $p_1 = P(Y_1 = 1)$ and $q = P(Y_0 = 1, Y_1 = 1)$, the Fréchet bounds are $\max(0, p_0 + p_1 - 1) \le q \le \min(p_0, p_1)$ and the cells are $P(1,1) = q$, $P(1,0) = p_0 - q$, $P(0,1) = p_1 - q$, $P(0,0) = 1 - p_0 - p_1 + q$. Benefit, harm, disagreement and, by (7), $\mathrm{Var}(Y_1 - Y_0)$ are affine in q, so their sharp ranges are attained at the two endpoints; $p_1 - p_0$ does not depend on q. In particular $\mathrm{PoB} \in [\max(0, p_1 - p_0), \min(p_1, 1 - p_0)]$.

B. GAUSSIAN CALCULATIONS AND PROOF OF PROPOSITION 6

If (Y_0, Y_1) is bivariate normal with means (m_0, m_1), common variance σ^2 and correlation ρ, then $D = Y_1 - Y_0 \sim \mathcal{N}(\delta, 2\sigma^2(1-\rho))$, so $\mathrm{Var}(D) = 2\sigma^2(1-\rho)$ and $P(D > d) = \Phi\big((\delta - d)/(\sigma\sqrt{2(1-\rho)})\big)$. As ρ ranges over $(-1, 1)$ the standard deviation of D ranges over $(0, 2\sigma)$, giving the Gaussian ranges in the proposition.

For the sharp range, write $F_0(y) = \Phi((y - m_0)/\sigma)$ and $F_1(y) = \Phi((y - m_1)/\sigma)$. The Makarov bounds (6) are best possible in the sense that, for each d, they are the infimum and supremum of $P(D \le d)$ over all couplings [22, 24]; we do not use, and do not claim, that every such extremum is attained. By (6), the infimum of $P(D > d) = 1 - P(D \le d)$ is $\max\{0, \sup_y [F_0(y-d) - F_1(y)]\}$. The difference $\Phi((y - d - m_0)/\sigma) - \Phi((y - m_1)/\sigma)$ is that of two normal distribution functions with the same scale and locations $m_0 + d$ and m_1; it is maximized at the midpoint $y = (m_0 + d + m_1)/2$, where it equals $\Phi(c) - \Phi(-c) = 2\Phi(c) - 1$ with $c = (\delta - d)/(2\sigma)$. This is positive iff $\delta > d$. The supremum is $1 - \max\{0, \sup_y [F_1(y) - F_0(y-d)]\}$, and $\sup_y [F_1(y) - F_0(y-d)] = \max\{0, 2\Phi(-c) - 1\}$ by the same argument (the midpoint maximization is machine-checked for the standard Gaussian, Appendix C). Hence for $\delta > d$ the infimum is $2\Phi(c) - 1$ and the supremum is 1; for $\delta < d$ they are 0 and $2\Phi(c)$. *Attainment.* The comonotone coupling $Y_1 = Y_0 + \delta$ gives $D \equiv \delta$, hence $P(D > d) = 1$ if $\delta > d$ and $P(D > d) = 0$ if $\delta < d$: the trivial endpoint is always attained. For the non-trivial endpoint ($2\Phi(c) - 1$ when $\delta > d$, $2\Phi(c)$ when $\delta < d$) we claim only that it is the infimum or supremum; attainment would require an additional argument for strict threshold events, which none of our conclusions needs, because the reported interval is the closure of the set of values attained by admissible couplings. The harm probability $P(Y_1 < Y_0) = P(Y_0 - Y_1 > 0)$ is the same problem with the roles of the marginals exchanged ($\delta \mapsto -\delta$, $d = 0$). The

Gaussian widths are $1-\Phi(c)$ and $\Phi(c)$, exactly half the sharp widths $2-2\Phi(c)$ and $2\Phi(c)$. For the variance, the comonotone coupling $Y_1 = Y_0 + \delta$ attains 0 and the countermonotone coupling $Y_1 = m_1 - (Y_0 - m_0)$ attains $4\sigma^2$; these are the limits of the Gaussian family as $\rho \to \pm 1$. With $\delta = 0.5$, $\sigma = 1$, $d = 0$: $2\Phi(0.25) - 1 = 0.197$ and $\Phi(0.25) = 0.599$, as in Table IV.

C. REPRODUCIBILITY

All numbers, tables and figures are produced by the supplementary package `coupling_audit` (Python 3.13, NumPy, SciPy 1.18 with HiGHS, pandas, Matplotlib; exact versions in `environment.lock.txt`). The command `make all` runs, in order: `test_audit` (12 numerical checks of the analytic claims, including the Tian–Pearl agreement, the Makarov bounds and the discretized-LP cross-checks); `exp1_gaussian_sharp` (Table IV, Figure 1); `exp2_simulation` (Table V, Figure 2; 12,000 data sets); `exp3_real_rct` (Table VI; data files and preprocessing rules included); `benchmark.validate_rubric` (Table VII; writes the 240 C-STRESS items to `items.jsonl`); and `make pilot-analysis`, which re-parses and re-scores the archived raw outputs of the pilot (no model calls), checks every record against the item bank, and produces Tables VIII and IX. Seeds are fixed (20261018 unless stated). LP settings: HiGHS via `linprog(method="highs")`, default tolerances, sparse constraint matrices. Projection intervals use Bonferroni-corrected Clopper–Pearson boxes; Imbens–Manski intervals solve for the critical value with Brent's method. New language-model runs use `benchmark.run_pilot` (OpenAI-compatible endpoint, temperature 0, 600 tokens) or the released Colab notebook, which also records the server and model versions. The non-LLM pipeline takes under one hour on a two-core CPU; a full re-run in a fresh environment is documented in `REPRODUCIBILITY_REPORT.md`.

Formal verification. The file `lean/CouplingAudit.lean` states 22 theorems (21 target statements and one intermediate result, the Gaussian window inequality `stdGauss_window`) plus auxiliary lemmas in Lean 4 (v4.28.0) with Mathlib (v4.28.0) and proves all of them. The target statements are: the finite-outcome versions of Theorems 1 and 2; Proposition 1 (all three equivalences); the Fréchet range, the sharp ranges of PoB, PN and PS, the identification condition of Proposition 2 and eq. (7); Lemma 2 (the coupling-transport bound and its consequence for scores); the deterministic core of Proposition 5; the Tian–Pearl combined-data bounds on PN as valid bounds over the feasible set of Theorem 2; and, for Φ the standard Gaussian distribution function and $c > 0$, $\sup_y[\Phi(y+c) - \Phi(y-c)] = 2\Phi(c) - 1$, the key step of Proposition 6. No statement contains unproved assumptions; the only axioms used are Lean's standard `propext`, `Classical.choice` and `Quot.sound`. Not formalized are: the measurable (non-finite) versions of Theorems 1 and 2; the Makarov bounds and the sharpness part of Proposition 6; the sharpness of the Tian–Pearl bounds (only their validity); and the statistical results (Proposition 3, the consistency corollary and the coverage statements), which involve estimators. The proofs were produced with the Aristotle system (Harmonic) and independently rebuilt from the pinned project in the supplement (Lean v4.28.0, Mathlib commit `8f9d9cf`): `lake build` succeeds and `#print axioms` reports only `propext`, `Classical.choice` and `Quot.sound` for every theorem. Build logs and a statement-by-statement correspondence table (`lean/REPORT.md`) are included.

Contents of the supplementary material. Table X lists every supplementary item referred to in the text. All items are uploaded with the manuscript as *Supplementary File S1*, a PDF that (i) reproduces as text the documentation, the reproducibility report, the register of inference claims, all source code, the result tables, the run logs and examples of benchmark items and raw responses, and (ii) contains every file of the archive, and the complete archive `BJAI-2026-0078_supplementary.zip`, as embedded attachments (Attachments panel of the PDF viewer). `MANIFEST_SHA256.txt` gives a checksum for each file.

DATA AND CODE AVAILABILITY

The two trial data sets are public: Thornton [35] and the NSW experiment [36, 53], both as distributed with [52]. Code, data extracts, the C-STRESS item bank, the raw language-model outputs of the pilot, the Lean 4 formalization and all generated outputs are provided as Supplementary File S1 and are permanently archived in a public repository under an open licence [55], https://doi.org/10.5281/zenodo.23664694.

RESEARCH ETHICS

This study uses only publicly available, de-identified secondary data and synthetic data; it did not involve new data collection from human participants or animals, so institutional ethics approval and informed consent were not required.

TABLE X. Supplementary material (Supplementary File S1 and its embedded archive; paths relative to the archive root).

Path	Content
`README.md`, `Makefile`	installation and the exact commands that regenerate every result
`requirements.txt`, `environment.lock.txt`	minimum and exact package versions (Python 3.13)
`coupling_audit/`	implementation of the audit (LP bounds, combined-data LP, inference)
`experiments/`	scripts for Tables III to VI, Figures 1 and 2 and the 12-check test suite
`data/`	public extracts of the Thornton and NSW trials, with preprocessing rules in `README.md`
`benchmark/`	C-STRESS generator, the 240 items (`items.jsonl`), scorer, runner, rubric validation, Colab notebook
`results/pilot_2026-10-18/`	raw language-model responses (720 records), validation file, manual scorer audit, Tables VIII and IX
`results/`, `figures/`	all generated tables and figures
`lean/`	Lean 4 formalization, pinned project files, correspondence table (`REPORT.md`), rebuild logs
`REPRODUCIBILITY_REPORT.md`, `logs/`	clean-environment re-run and comparison with the reported values
`inference_claims.md`	every coverage claim with its regime (finite-sample, asymptotic, heuristic) and justification

CONFLICTS OF INTEREST

The author declares no conflicts of interest.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Jordi Vallverdú: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

FUNDING

This work was supported by the project PALGAIT (PID2023-148336NB-I00), funded by MICIU/AEI/10.13039/501100011033.

ACKNOWLEDGMENT

The author thanks the editor and the reviewers for comments that substantially improved the article.

USE OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN WRITING

During manuscript preparation and revision, the author, a non-native English speaker, used OpenAI ChatGPT and Anthropic Claude to assist with language editing, LaTeX formatting, and drafting and checking analysis code, and the Aristotle system (Harmonic) to produce Lean 4 proofs of the formalized statements, which were checked by the Lean kernel. The author independently reviewed and verified the mathematical claims, proofs, interpretations, references, code and numerical results, and takes full responsibility for the submitted content.

REFERENCES

[1] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4066–4076.

[2] C. Russell, M. J. Kusner, J. R. Loftus, and R. Silva, "When worlds collide: Integrating different counterfactual assumptions in fairness," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 6414–6423.

[3] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: From counterfactual explanations to interventions," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY: ACM, 2021, pp. 353–362.

[4] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: Contrastive explanations and consequential recommendations," *ACM Computing Surveys*, vol. 55, no. 5, pp. 95:1–95:29, 2022.

[5] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[6] A. Li and J. Pearl, "Unit selection based on counterfactual logic," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019, pp. 1793–1799.
[7] J. Pearl, "Probabilities of causation: Three counterfactual interpretations and their identification," *Synthese*, vol. 121, no. 1–2, pp. 93–149, 1999.
[8] J. Tian and J. Pearl, "Probabilities of causation: Bounds and identification," *Annals of Mathematics and Artificial Intelligence*, vol. 28, no. 1–4, pp. 287–313, 2000.
[9] S. Mueller, A. Li, and J. Pearl, "Causes of effects: Learning individual responses from population data," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, 2022, pp. 2712–2718.
[10] N. Kallus, "What's the harm? sharp bounds on the fraction negatively affected by treatment," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.
[11] I. Ciocănea-Teodorescu, E. E. Gabriel, and A. Sjölander, "Sensitivity analysis for the probability of benefit in randomized controlled trials with a binary treatment and a binary outcome," *Biostatistics*, vol. 26, no. 1, p. kxaf011, 2025.
[12] J. J. Heckman, J. Smith, and N. Clements, "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts," *The Review of Economic Studies*, vol. 64, no. 4, pp. 487–535, 1997.
[13] Y. Fan and S. S. Park, "Sharp bounds on the distribution of treatment effects and their statistical inference," *Econometric Theory*, vol. 26, no. 3, pp. 931–951, 2010.
[14] S. Firpo and G. Ridder, "Partial identification of the treatment effect distribution and its functionals," *Journal of Econometrics*, vol. 213, no. 1, pp. 210–234, 2019.
[15] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.
[16] P. W. Holland, "Statistics and causal inference," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.
[17] Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. Lyu, K. Blin, F. Gonzalez Adauto, M. Kleiman-Weiner, M. Sachan, and B. Schölkopf, "CLadder: Assessing causal reasoning in language models," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.
[18] Z. Jin, J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. Diab, and B. Schölkopf, "Can large language models infer causation from correlation?" in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
[19] Y. Chen, V. K. Singh, J. Ma, and R. Tang, "CounterBench: Evaluating and improving counterfactual reasoning in large language models," 2025, arXiv:2502.11008.
[20] J. Vallverdú and A. Xu, "Do multi-agent LLMs improve causal identification? a controlled, cost-aware study on CLadder," *Array*, vol. 31, p. 101010, 2026.
[21] M. Fréchet, "Sur les tableaux de corrélation dont les marges sont données," *Annales de l'Université de Lyon, Section A*, vol. 14, pp. 53–77, 1951.
[22] G. D. Makarov, "Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed," *Theory of Probability and Its Applications*, vol. 26, no. 4, pp. 803–806, 1982.
[23] L. Rüschendorf, "Random variables with maximum sums," *Advances in Applied Probability*, vol. 14, no. 3, pp. 623–632, 1982.
[24] M. J. Frank, R. B. Nelsen, and B. Schweizer, "Best-possible bounds for the distribution of a sum—a problem of Kolmogorov," *Probability Theory and Related Fields*, vol. 74, no. 2, pp. 199–211, 1987.
[25] R. C. Williamson and T. Downs, "Probabilistic arithmetic I. numerical methods for calculating convolutions and dependency bounds," *International Journal of Approximate Reasoning*, vol. 4, no. 2, pp. 89–158, 1990.
[26] L. De Lara, A. González-Sanz, N. Asher, L. Risser, and J.-M. Loubes, "Transport-based counterfactual models," *Journal of Machine Learning Research*, vol. 25, no. 136, pp. 1–59, 2024.
[27] L. De Lara, "Canonical representations of Markovian structural causal models: A framework for counterfactual reasoning," 2025, arXiv:2507.16370.
[28] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge: Cambridge University Press, 2009.
[29] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: MIT Press, 2017.
[30] A. Balke and J. Pearl, "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, vol. 92, no. 439, pp. 1171–1176, 1997.
[31] M. C. Sachs, G. Jonzon, A. Sjölander, and E. E. Gabriel, "A general method for deriving tight symbolic bounds on causal effects," *Journal of Computational and Graphical Statistics*, vol. 32, no. 2, pp. 567–576, 2023.
[32] G. Duarte, N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser, "An automated approach to causal inference in discrete settings," *Journal of the American Statistical Association*, vol. 119, no. 547, pp. 1778–1793, 2024.
[33] G. W. Imbens and C. F. Manski, "Confidence intervals for partially identified parameters," *Econometrica*, vol. 72, no. 6, pp. 1845–1857, 2004.
[34] J. Stoye, "More on confidence intervals for partially identified parameters," *Econometrica*, vol. 77, no. 4, pp. 1299–1315, 2009.
[35] R. L. Thornton, "The demand for, and impact of, learning HIV status," *American Economic Review*, vol. 98, no. 5, pp. 1829–1863, 2008.
[36] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review*, vol. 76, no. 4, pp. 604–620, 1986.
[37] G. B. Dantzig, *Linear Programming and Extensions*. Princeton, NJ: Princeton University Press, 1963.
[38] C. F. Manski, *Partial Identification of Probability Distributions*. New York: Springer, 2003.
[39] E. Tamer, "Partial identification in econometrics," *Annual Review of Economics*, vol. 2, pp. 167–195, 2010.
[40] C. Villani, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Berlin: Springer, 2009, vol. 338.
[41] C. F. Manski, "Nonparametric bounds on treatment effects," *The American Economic Review*, vol. 80, no. 2, pp. 319–323, 1990.
[42] Z. Zhang and T. S. Richardson, "Individual treatment effect: Prediction intervals and sharp bounds," 2025, arXiv:2506.07469.
[43] W. Ji, L. Lei, and A. Spector, "Model-agnostic covariate-assisted inference on partially identified causal effects," 2024, arXiv:2310.08115.
[44] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.
[45] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.
[46] J. Jiang, F. Leofante, A. Rago, and F. Toni, "Interval abstractions for robust counterfactual explanations," *Artificial Intelligence*, vol. 336, p. 104218, 2024.
[47] L. De Lara, "A clarification on the links between potential outcomes and do-interventions," *Journal of Causal Inference*, vol. 13, no. 1, p. 20240033, 2025.
[48] I. Shpitser and J. Pearl, "Complete identification methods for the causal hierarchy," *Journal of Machine Learning Research*, vol. 9, no. 64, pp. 1941–1979, 2008.
[49] Q. Huangfu and J. A. J. Hall, "Parallelizing the dual revised simplex method," *Mathematical Programming Computation*, vol. 10, no. 1, pp. 119–142,

2018.
[50] P. Virtanen, R. Gommers, T. E. Oliphant *et al.*, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.
[51] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.
[52] S. Cunningham, *Causal Inference: The Mixtape*. New Haven, CT: Yale University Press, 2021.
[53] R. H. Dehejia and S. Wahba, "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1053–1062, 1999.
[54] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 8, pp. 229–231, 1959.
[55] J. Vallverdú, "Coupling-aware counterfactual reasoning for causal artificial intelligence: Code, data extracts, C-STRESS benchmark, pilot outputs and Lean 4 formalization," Zenodo, 2026, version 1.0, https://doi.org/10.5281/zenodo.23664694.