COUNTERFACTUAL GENERATION FROM AUDIT TRAILS IN MULTI-AGENT NEGOTIATION SYSTEMS

MARTIN LOTZ, PIETRO ALUFFI, MARYA BAZZI, MATT ARDERNE, AND VLADIMIRS MUREVICS

ABSTRACT. When autonomous agents negotiate on behalf of human principals, adverse outcomes demand an answer to the question: *which facts, disclosures, or contract clauses were decision-critical, and what changes would have altered the result?* We formalise this as a constrained optimisation over a structural causal model (SCM) of the multi-agent coordination pipeline, organise the resulting counterfactual interventions into three interpretable classes — evidence, clause, and protocol counterfactuals — and prove that the cheapest single-class intervention upper-bounds the counterfactual margin, with exactness under an explicit layer-local pivotality condition. Building on the established correspondence between counterfactual explanations and adversarial perturbations, we derive closed-form margin formulae for linear decision boundaries under weighted ℓ_∞ cost functions via Lagrange duality and the Fenchel conjugate. The layered structure of the causal graph enables per-layer verification of counterfactual claims, connecting to recent work on verifiable causality analysis. We extend the framework to adversarial settings by modelling the counterparty's disclosure strategy as an endogenous variable, distinguishing between factual, disclosure, and policy counterfactuals: separating "the counterparty was truly risky" from "the counterparty strategically withheld information."

1. INTRODUCTION

With the advent of LLM-based agents as autonomous economic actors, negotiating contracts, executing transactions, and managing portfolios on behalf of human principals, the nature of multi-party coordination is changing profoundly. By dramatically reducing search, contracting, and enforcement costs, these systems approach what has been termed a "Coasean singularity" [Wangsomcharoen and Dworczak, 2025, Krier, 2025]: flexible, context-dependent workflows that span organisational boundaries [Kim et al., 2026, Li et al., 2026] and enable coordination patterns that were previously infeasible. But this capability creates an accountability gap. When an agent-mediated decision produces an adverse outcome, neither the principal nor the regulator can readily determine which inputs, disclosures, or policy clauses were decision-critical, nor what changes would have altered the result. The accountability infrastructure must scale with the coordination capability it enables.

Counterfactual audit analysis is a core component of that infrastructure: given an observed decision, identify the minimal set of facts, disclosures, or contract clauses whose alteration would have changed the outcome. Formalising this capability is nontrivial for three reasons that distinguish the agentic setting from the standard counterfactual explanation literature:

- **Unknown counterparty mechanisms.** In classical algorithmic recourse [Wachter et al., 2018, Karimi et al., 2022], the decision-maker's model is known or queryable. In agent-to-agent coordination, the counterparty's internal process is a black box.
- **Adversarial and strategic behaviour.** The counterparty is a strategic agent whose disclosures are the output of an optimisation [Akerlof, 1970, Stiglitz and Weiss, 1981]. A counterfactual analysis that treats evidence as exogenous conflates "the borrower was truly risky" with "the borrower strategically withheld information."

Date: March 2026 — Working Draft.
2020 *Mathematics Subject Classification.* 68T42, 91A28, 94A60.
Key words and phrases. agent coordination, counterfactual audit.

- **Hybrid variable types.** The coordination pipeline mixes continuous variables (financial metrics), discrete variables (clause verdicts, protocol selections), and structured objects (negotiation transcripts). Standard gradient-based methods assume a continuous, differentiable feature space.

This paper develops a formal framework that addresses all three challenges. We model the coordination pipeline as a structural causal model (SCM), define minimal counterfactual interventions via constrained optimisation, and decompose counterfactuals into interpretable classes (evidence, clause, and protocol counterfactuals). We extend the framework to strategic behaviour and develop both replay-based and learned computational strategies.

Our contributions are:

(1) We formalise the counterfactual audit problem as a constrained optimisation over an SCM of the coordination pipeline (Section 3) and derive closed-form counterfactual margins for linear decision boundaries under weighted ℓ_∞ audit costs via Lagrange duality (Section 3.1).
(2) We organise counterfactual interventions into three interpretable classes — evidence, clause, and protocol counterfactuals — and prove that the cheapest single-class intervention upper-bounds the counterfactual margin, with exactness under an explicit layer-local pivotality condition (Section 3.4).
(3) We develop two computational strategies — replay-based search with causal pruning and a learned counterfactual generator (Section 4) — and robustness safeguards — counterfactual confidence scores and multiplicity reporting — against post-hoc rationalisation (Section 5).
(4) We give a three-tier identifiability regime for pipelines with opaque (LLM-based) layers, based on monotone-transport counterfactual identification (Section 6), and extend the framework to adversarial settings, distinguishing true-state, disclosure, and policy counterfactuals (Section 7).

2. Related Work

Counterfactual audit analysis for multi-agent systems sits at the intersection of several research traditions: causal inference and structural equation modelling, which provide the formal language for counterfactual reasoning; adversarial robustness, which studies the dual problem of minimal decision-changing perturbations; algorithmic recourse and explainability, which develop computational methods for counterfactual generation; information economics and mechanism design, which model the strategic disclosure behaviour of self-interested agents; and verifiable computation, which addresses the trust problem when audit results must be checked by third parties. We review each in turn, focusing on the aspects most relevant to our framework.

Counterfactual explanations and algorithmic recourse. Wachter et al. [2018] introduced the modern formulation of counterfactual explanations: given an unfavourable classifier decision, find the minimal change to the input features that would produce a favourable outcome, framed as a constrained optimisation over the feature space. Subsequent work has refined this formulation along several axes: Mothilal et al. [2020] generate *diverse* counterfactual sets to address multiplicity (multiple equally minimal explanations); Ustun et al. [2019] impose actionability constraints (some features, such as age, cannot be changed) and integer constraints for discrete variables; Joshi et al. [2019] extend the approach to black-box models using learned generative proxies; Karimi et al. [2021] recast recourse as *minimal interventions on a causal model* rather than feature perturbations — the formulation we adopt; and Karimi et al. [2022] provide a comprehensive survey, distinguishing methods by whether they respect causal structure, actionability, and plausibility. Halpern and Pearl [2005] address a complementary problem, *actual causation*, formalising when an event can be said to have caused an outcome in a specific instance. Our work builds directly on these formulations and differs from this literature in three respects: the "model" is a multi-agent coordination pipeline rather than a single classifier; the goal is audit (understanding what happened) rather than recourse (advising change); and some variables are controlled by an adversarial counterparty whose disclosure strategy is endogenous. The individual ingredients —

minimal counterfactuals, causal propagation, the adversarial correspondence discussed below — are established; the contribution here is their combination in a layered audit pipeline with per-layer margins, verifiability, and strategic counterparties.

Information economics and strategic disclosure. The distinction between factual and disclosure counterfactuals (Section 7) draws on the information economics of adverse selection [Akerlof, 1970, Stiglitz and Weiss, 1981]. Classical disclosure theory shows that with verifiable messages and commonly known information endowments, unravelling forces full disclosure in equilibrium [Grossman, 1981, Milgrom, 1981]; when these conditions fail — as they do in LLM-mediated negotiation, where the counterparty's information endowment is itself uncertain — strategic withholding survives, and the disclosure counterfactual of Section 7 becomes substantive. In the agentic setting, a counterparty can optimise its disclosure strategy in real time. The revelation principle [Myerson, 1979] implies that the auditor should model the counterparty's evidence as the output of an optimal disclosure strategy applied to the true state, rather than treating it as exogenous. The robustness of coordination under adversarial conditions also connects to Byzantine-robust federated learning [Das and Sen, 2026], where adaptive aggregation without parametric attack assumptions informs our counterfactual confidence measures.

Strategic classification and performative prediction. When decision subjects respond strategically to a decision rule, the distribution of observed features becomes endogenous to the rule itself. Hardt et al. [2016] formalise this as *strategic classification*, in which agents manipulate their features at a cost to obtain favourable decisions, and Perdomo et al. [2020] generalise it to *performative prediction*, in which deployed models shift the data distribution they are evaluated on. Our strategic extension (Section 7) transposes this concern to the audit problem: the counterparty's evidence is modelled as the output of a disclosure strategy, and the audit must distinguish counterfactuals over the true state from counterfactuals over the disclosure map — a distinction that does not arise when evidence is treated as exogenous.

Structured transparency. The disclosure protocols whose conformance our framework audits are an instance of *structured transparency* [Trask et al., 2020]: information-flow designs that let mutually distrusting parties share, verify and act on sensitive information under explicit bounds (field whitelists, query budgets, anti-enumeration limits), rather than forcing a choice between full disclosure and refusal. Counterfactual audit is the natural retrospective complement of such designs: where structured transparency constrains which information *may* flow, the audit quantifies which flows were, in fact, decision-critical — and the disclosure counterfactual of Section 7 makes the dependence on the counterparty's use of those channels explicit.

Adversarial robustness. The minimal counterfactual problem is structurally similar to the adversarial perturbation problem: both seek the smallest input change that flips a decision. Szegedy et al. [2014] first demonstrated that neural networks are vulnerable to imperceptibly small perturbations, and subsequent work has developed both attacks (projected gradient descent, Carlini–Wagner [Carlini and Wagner, 2017], DeepFool [Moosavi-Dezfooli et al., 2016]) and fundamental limits on robustness [Fawzi et al., 2018]. That the adversarial distance and the counterfactual margin $\mathrm{cost}(\delta^*)$ are instances of the same optimisation with different cost functions and feasible sets has been made precise in the explanation literature [Freiesleben, 2022, Pawelczyk et al., 2022]; we do not claim this correspondence as a contribution. Our use of it (Sections 3.1 and 3.4) is to deploy it in a layered, causally structured perturbation space, where it yields per-layer margins and imports both algorithms and fundamental limits from adversarial robustness into counterfactual audit.

Verifiable causality analysis. Song et al. [2026] introduce vCAUSE, which uses authenticated data structures (a graph accumulator and a verifiable provenance graph) to enable third-party validation of causality analysis over system logs. Their approach processes 25M logs in 2 minutes and achieves provable unforgeability. The architecture maps directly to our setting: agent interaction traces can be stored as verifiable provenance graphs, and the graph accumulator's proof mechanism can be adapted for verifiable counterfactual audit statements. We develop this connection in Section 3.5.

3. Structural Causal Model for Agent Decisions

We model the agent coordination pipeline as a *structural causal model* (SCM) [Pearl, 2009] over the trace variables. Informally, a *causal graph* is a directed acyclic graph whose nodes are the variables of the system and whose edges record direct mechanistic dependence: an edge $X \to Y$ asserts that X is an input to the mechanism that generates Y. While the graph encodes the qualitative causal assumptions (which variables respond to which), the quantitative content is supplied by attaching to each node a *structural equation*, a function producing the node's value from the values of its graph parents together with an exogenous noise term that captures everything outside the model. Acyclicity ensures the system can be evaluated by a single forward pass in topological order (parents evaluated before their children). What this structure buys, over an ordinary probabilistic model, is a well-defined notion of *intervention*: to model "setting Y to y" one deletes Y's structural equation (equivalently, severs its incoming edges) and substitutes the fixed value, leaving every other mechanism untouched — a surgery on the model, which is quite different from conditioning on the observation $Y = y$. In our setting the nodes are the trace variables of a negotiation (evidence items, clause verdicts, protocol state, decision), the edges follow the pipeline's data flow, and interventions formalise audit questions of the form "what would the decision have been, had this input been different?"

Formally, an SCM is a tuple $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{F})$, where $\mathbf{V}$ is the set of endogenous variables, $\mathbf{U}$ is the set of exogenous (latent) variables, and $\mathcal{F} = \{f_V\}_{V \in \mathbf{V}}$ is the set of structural equations, with each f_V determining V as a function of its parents $\mathrm{pa}(V)$ in the causal graph $\mathcal{G}$ and the relevant exogenous variables. The central operation on an SCM is the *intervention*, Pearl's do-operator.

Definition 3.1 (Counterfactual intervention)**.** Let $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathcal{F})$ be an SCM with causal graph $\mathcal{G}$, and let $D \in \mathbf{V}$ be a designated outcome variable with observed value d. A *counterfactual intervention* is a set $\delta = \{\mathrm{do}(V_i := v_i')\}_{i \in S}$ for some $S \subseteq \mathbf{V} \setminus \{D\}$, which replaces the structural equations of the variables in S with the prescribed values v_i' and recomputes all downstream variables by forward-propagating through $\mathcal{F}$ in topological order, holding the exogenous variables $\mathbf{U}$ fixed at their abducted values. The resulting counterfactual outcome is denoted $d_{\mathcal{M}}(\delta)$.

We say δ is *decision-changing* if $d_{\mathcal{M}}(\delta) \neq d$. The requirement that $\mathbf{U}$ is held fixed is the abduction step of Pearl's three-step procedure (see Pearl, 2009, Chapter 7, §§7.1.2–7.1.3): the exogenous variables are first inferred from the observed data, then kept constant while the intervention is applied and the outcome recomputed. This ensures that the counterfactual pertains to the *same individual* (or, in our setting, the same interaction), not to a generic draw from the population.

Definition 3.2 (Minimal counterfactual identification)**.** Given an SCM $\mathcal{M}$, an outcome variable D with observed value d, a feasible set Δ of interventions, and a cost function $\mathrm{cost}\colon \Delta \to \mathbb{R}_+$ measuring the "size" of an intervention, the *minimal counterfactual* is

$$\delta^* = \operatorname*{arg\,min}_{\delta \in \Delta} \ \mathrm{cost}(\delta) \quad \text{subject to} \quad d_{\mathcal{M}}(\delta) \neq d. \tag{3.1}$$

The value $\mathrm{cost}(\delta^*)$ is the *counterfactual margin* of the decision.

We require cost to be *monotone* across disjoint variable sets: if $\delta = \delta_A \cup \delta_B$ with δ_A and δ_B acting on disjoint subsets of $\mathbf{V}$, then $\mathrm{cost}(\delta) \geq \max\big(\mathrm{cost}(\delta_A), \mathrm{cost}(\delta_B)\big)$. This is the property used in the layer-wise decomposition of Proposition 3.7. All cost functions considered in this paper are monotone: additively separable costs (weighted ℓ_1 norms, combinatorial per-variable costs) satisfy $\mathrm{cost}(\delta_A \cup \delta_B) = \mathrm{cost}(\delta_A) + \mathrm{cost}(\delta_B) \geq \max\big(\mathrm{cost}(\delta_A), \mathrm{cost}(\delta_B)\big)$ by non-negativity of the summands, while the weighted ℓ_∞ norm satisfies monotonicity with equality, $\|(\delta_A, \delta_B)\|_{w,\infty} = \max\big(\|\delta_A\|_{w,\infty}, \|\delta_B\|_{w,\infty}\big)$, and is *not* additively separable. Where a result requires additivity rather than monotonicity, we say so explicitly.

The cost function $\mathrm{cost}(\delta)$ must be chosen to produce meaningful counterfactuals; its design is domain-specific and discussed in detail in Section 3.3. We first illustrate both definitions with a toy example.

Example 3.3 (A simple SCM for a loan decision)**.** Consider a minimal lending scenario with four endogenous variables: *income* $X_1 \in \mathbb{R}_+$, *debt* $X_2 \in \mathbb{R}_+$, *credit score* $S \in [0,1]$, and *decision* $D \in \{\text{approve}, \text{reject}\}$. The structural equations are:

$$X_1 = U_1,\ X_2 = U_2,\ S = \sigma\big((\alpha\, X_1 - \beta\, X_2 + U_S)/T\big),\ D = \begin{cases} \text{approve} & \text{if } S \geq \theta, \\ \text{reject} & \text{otherwise,} \end{cases}$$

where σ is the sigmoid function, $\alpha, \beta > 0$ are fixed weights, $T > 0$ is a scale (temperature) parameter setting the sensitivity of the score to the financial inputs, θ is the approval threshold, and U_1, U_2, U_S are independent exogenous variables. Here U_S captures all determinants of the credit score beyond income and debt (payment history, credit utilisation, length of credit history, and measurement noise). It is specific to a given applicant: in Pearl's three-step counterfactual procedure, U_S is recovered during the abduction step and then *held fixed* when evaluating the intervention, encoding the assumption that the applicant's full credit profile stays the same in the counterfactual world. The causal graph and the resulting decision boundary (for $\alpha = \beta = 1$, $U_S = 0$, $T = 32.3$) are shown in Figure 1.

Suppose we observe a rejected applicant with $X_1 = 40$, $X_2 = 60$, $S = 0.35$, $D = \text{reject}$ (with $\theta = 0.5$; the observed score is consistent with the abducted noise, since $\sigma(-20/32.3) \approx 0.35$). The counterfactual question is: *what is the smallest change to the inputs that would have led to approval?* Since D depends on X_1 and X_2 only through S, we can either intervene on the inputs (an *evidence counterfactual*: e.g., $\mathrm{do}(X_2 := 30)$ might raise S above θ) or directly on the score (a *clause counterfactual*: $\mathrm{do}(S := 0.5)$). The minimal counterfactual δ^* is whichever is cheapest under the cost function $\mathrm{cost}(\cdot)$. This toy example already exhibits the layer-by-layer decomposition we formalise in Proposition 3.7: interventions on (X_1, X_2) and on S can be optimised independently because the graph is layered (and, because either layer alone can flip the decision, the pure-layer bound of the proposition is attained here).

3.1. **The counterfactual margin for linear decision boundaries.** Example 3.3 admits a clean closed-form analysis since the decision boundary is linear in the intervened variables. Since $D = \mathbf{1}[S \geq \theta]$ and $S = \sigma\big((\mathbf{a}^T\mathbf{x} + U_S)/T\big)$ with $\mathbf{a} = (\alpha, -\beta)^T$ and $\mathbf{x} = (X_1, X_2)^T$, the boundary $S = \theta$ corresponds to the hyperplane $H = \{\mathbf{x} : \mathbf{a}^T\mathbf{x} + b = 0\}$ where $b = U_S - T\sigma^{-1}(\theta)$. The minimal counterfactual problem (3.1), restricted to evidence interventions on $\mathbf{x}$, becomes

$$\delta^* = \arg\min_{\delta \in \mathbb{R}^2} \|\delta\| \quad \text{s.t.} \quad \mathbf{a}^T(\mathbf{x} + \delta) + b = 0, \tag{3.2}$$

where the norm $\|\cdot\|$ encodes the cost function. The choice of norm determines both the geometry of the counterfactual and the form of the solution.

If we use the Euclidean distance ($\|\delta\| = \|\delta\|_2$), the solution can be derived using elementary geometry (see, for example, the lecture notes [Lotz, 2024, Chapter 22]):

$$\delta_2^* = -\frac{\mathbf{a}^T\mathbf{x} + b}{\|\mathbf{a}\|_2^2}\,\mathbf{a}, \qquad \|\delta_2^*\|_2 = \frac{|\mathbf{a}^T\mathbf{x} + b|}{\|\mathbf{a}\|_2}. \tag{3.3}$$

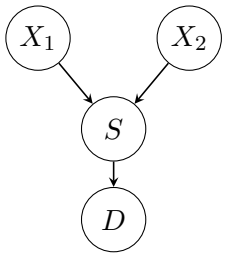
In the audit setting, a more natural cost function is the *weighted* ℓ_∞ *norm*

$$\|\delta\|_{w,\infty} = \max_i\ w_i|\delta_i|,$$

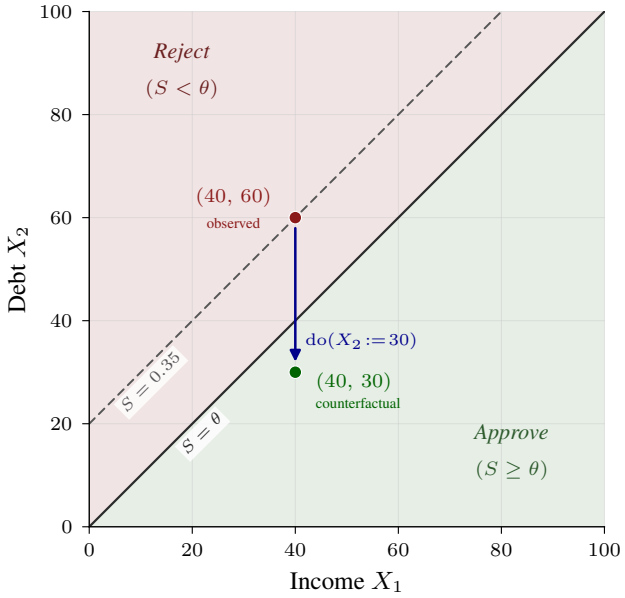
where the weights $w_i > 0$ reflect the *auditability* of each variable: a variable that is cheap to verify (e.g., a bank balance) receives a small weight, while one requiring extensive due diligence (e.g., a revenue projection) receives a large weight. The norm measures the worst-case weighted change to any single variable.

To solve (3.2) with this norm, we introduce a Lagrange multiplier $\lambda \in \mathbb{R}$ for the hyperplane constraint and form the Lagrangian

$$\mathcal{L}(\delta, \lambda) = \|\delta\|_{w,\infty} + \lambda\big(\mathbf{a}^T\mathbf{x} + b + \mathbf{a}^T\delta\big).$$



(a) Causal graph



(b) Decision boundary and counterfactual

FIGURE 1. Toy SCM of Example 3.3. (a) Causal graph: X_1 (income) and X_2 (debt) determine the credit score S, which determines the decision D. (b) Decision boundary with $\alpha = \beta = 1$, $U_S = 0$, $T = 32.3$. The solid line is $S = \theta$, i.e. $X_1 = X_2$; the dashed line is the level set $S = 0.35$, i.e. $X_1 - X_2 = -20$, which passes through the observed applicant $(40, 60)$. The intervention $\mathrm{do}(X_2 := 30)$ moves the applicant from $(40, 60)$ in the reject region to $(40, 30)$ in the approve region.

To compute the dual, we minimise over δ. The key concept here is the *Fenchel conjugate* of a norm: for any norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$,

$$\inf_{\delta}\big[\|\delta\| + \mathbf{c}^T\delta\big] = \begin{cases} 0 & \text{if } \|\mathbf{c}\|_* \leq 1, \\ -\infty & \text{otherwise.} \end{cases}$$

The dual of the weighted ℓ_∞ norm is the weighted ℓ_1 norm with reciprocal weights:

$$\|\cdot\|_{w,\infty}^* = \|\cdot\|_{w^{-1},1}, \qquad \|\mathbf{y}\|_{w^{-1},1} = \sum_i \frac{|y_i|}{w_i}. \tag{3.4}$$

Applying this with $\mathbf{c} = \lambda\mathbf{a}$, the infimum over δ is zero whenever $|\lambda| \cdot \|\mathbf{a}\|_{w^{-1},1} \leq 1$, and $-\infty$ otherwise. The dual problem is therefore

$$\max_{\lambda}\ \lambda\big(\mathbf{a}^T\mathbf{x} + b\big) \quad \text{s.t.} \quad |\lambda| \leq \frac{1}{\|\mathbf{a}\|_{w^{-1},1}}. \tag{3.5}$$

Since $\mathbf{a}^T\mathbf{x} + b < 0$ (the observed point is on the reject side), the optimum is $\lambda^* = -1/\|\mathbf{a}\|_{w^{-1},1}$, yielding the closed-form **counterfactual margin**:

$$d_{w,\infty}(\mathbf{x}, H) = \frac{|\mathbf{a}^T\mathbf{x} + b|}{\displaystyle\sum_i |a_i|/w_i}. \tag{3.6}$$

This is the weighted-ℓ_∞ analogue of the Euclidean formula (3.3): the numerator is the same (the signed distance in the "decision space"), but the denominator is the dual norm $\|\mathbf{a}\|_{w^{-1},1}$.

The optimal δ^* is obtained from the KKT condition $\mathbf{0} \in \partial\|\delta^*\|_{w,\infty} + \lambda^*\mathbf{a}$, where the subdifferential of the weighted ℓ_∞ norm at δ is

$$\partial\|\delta\|_{w,\infty} = \operatorname{conv}\big\{w_j \operatorname{sign}(\delta_j)\, \mathbf{e}_j : j \in \arg\max_i\, w_i|\delta_i|\big\},$$

where $\mathbf{e}_j$ is the j-th standard basis vector and conv denotes the convex hull. The KKT condition requires $-\lambda^*\mathbf{a}$ to lie in this subdifferential, which determines the direction and support of the optimal perturbation. The optimal perturbation is unique precisely when $a_i \neq 0$ for every coordinate i: equality in the chain $\mathbf{a}^T\delta \leq \sum_i (|a_i|/w_i)\, w_i|\delta_i| \leq \|\delta\|_{w,\infty} \sum_i |a_i|/w_i$ forces $w_i|\delta_i^*| = d_{w,\infty}$ and $\operatorname{sign}(\delta_i^*) = \operatorname{sign}(a_i)$ on every coordinate with $a_i \neq 0$, so that $\delta_i^* = \operatorname{sign}(a_i)\, d_{w,\infty}/w_i$. Coordinates with $a_i = 0$ do not affect the constraint and are free within $|\delta_i| \leq d_{w,\infty}/w_i$; the canonical minimiser sets them to zero.

Example 3.4 (Counterfactual margin in the toy SCM)**.** Returning to Example 3.3 with $\alpha = \beta = 1$, $U_S = 0$, $\theta = 0.5$, so that $\mathbf{a} = (1, -1)^T$, $b = 0$ (the scale T drops out since $\sigma^{-1}(0.5) = 0$), and $\mathbf{x} = (40, 60)^T$.

Equal weights ($w_1 = w_2 = 1$):

$$d_{w,\infty} = \frac{|40 - 60|}{1/1 + 1/1} = \frac{20}{2} = 10.$$

The optimal perturbation is $\delta^* = (10, -10)$, moving to $(50, 50)$ on the boundary. Both variables change by the same amount.

Asymmetric weights ($w_1 = 2$, $w_2 = 1$; income is harder to change than debt):

$$d_{w,\infty} = \frac{20}{1/2 + 1/1} = \frac{20}{3/2} = \tfrac{40}{3} \approx 13.3.$$

The margin increases because the cheaper variable (X_2, debt) must absorb more of the perturbation to compensate for the expensive one (X_1, income). The optimal perturbation has $w_1|\delta_1^*| = w_2|\delta_2^*|$, so $\delta_1^* = 40/6 \approx 6.7$ and $\delta_2^* = -40/3 \approx -13.3$, moving to approximately $(46.7, 46.7)$.

Remark 3.5 (Adversarial robustness)**.** Our optimization problem is structurally similar to the adversarial perturbation problem in classification. Both problems seek the minimal perturbation that flips a decision; this correspondence between counterfactual explanations and adversarial examples is well documented in the explanation literature [Freiesleben, 2022, Pawelczyk et al., 2022]. It has several consequences for the audit setting.

The first consequence is that the fundamental limits on adversarial robustness carry over. In the classification setting, Fawzi et al. [2018] showed that if the data is generated by an L-Lipschitz map $g\colon \mathbb{R}^m \to \mathbb{R}^d$ from a latent Gaussian space, then for any classifier the distance $\hat{\Delta}(X)$ from a sample X to the decision boundary satisfies $\mathbb{P}\{\hat{\Delta}(X) \leq \epsilon\} \geq 1 - \sqrt{\pi/2}\, e^{-\epsilon^2/2L^2}$. The bound is informative when ϵ is comparable to or larger than L: most points then lie within distance ϵ of a decision boundary, so no classifier over such data can have margins much larger than the natural scale L of the data-generating process, and adversarial perturbations of that magnitude are unavoidable. The analogue for the audit setting is that the counterfactual margin of a decision pipeline evaluated on naturally generated evidence will generically be no larger than the scale of ordinary variation in that evidence, making counterfactual identification both feasible and necessary.

Secondly, the algorithms developed for adversarial perturbation search apply, with modifications, to counterfactual generation. The DeepFool algorithm [Moosavi-Dezfooli et al., 2016] iteratively linearises the decision boundary and projects onto the nearest class boundary, a strategy that extends to a layered causal graph by applying linearisation layer by layer. More broadly, any constrained optimisation method for finding adversarial examples (projected gradient descent, Carlini–Wagner attacks [Carlini and Wagner, 2017]) can be adapted to the counterfactual problem by replacing the norm constraint with the causal cost function and restricting the feasible set to respect the graph structure.

Ultimately, from a geometric perspective, both the problem of counterfactual generation and the problem of adversarial perturbations can be framed as the problem of deciding whether a given

point is within a given ε-neighbourhood of a geometric object. When we take a randomized point of view and ask for the probability that a typical point is within a given distance from a decision boundary becomes a question about the volume of tubular neighbourhoods. For general Pfaffian decision boundaries, and more specifically for one-layer sigmoid neural networks (such as the function presented in Example 3.3), bounds on the probability of being within a certain distance from the boundary were derived in [Lezeau and Lotz, 2026] using tools from algebraic geometry.

3.2. **Causal Graph Structure for Agent Coordination.** We now apply this framework to the richer structure of agent coordination. The endogenous variables partition naturally into four groups:

- *Facts* F_i: observable evidence presented during the interaction (e.g., financial metrics, verification documents, bank feed data);
- *Clauses* C_j: contract and policy clauses active during negotiation (e.g., disclosure requirements, exposure limits, escalation rules);
- *Protocol states* P_ℓ: intermediate states of the coordination protocol (e.g., selected negotiation pattern, verification method, fallback trigger);
- *Decision* D: the terminal decision (approve, reject, counter, refer).

The exogenous variables $\mathbf{U}$ capture hidden counterparty state, unobserved risk factors, and noise.

To make the graph concrete, we introduce the following notation. Let:

- $\mathbf{E} = (E_1, \ldots, E_n)$ denote the evidence package submitted by the counterparty (financial statements, verification documents, etc.);
- π denote the local agent's policy state (hard constraints, soft preferences, escalation rules);
- $\mathbf{N} = (N_1, \ldots, N_r)$ denote the negotiation transcript (proposals, counter-proposals, clarification requests);
- $\varphi \in \Phi$ denote the selected protocol from the protocol library;
- $\mathbf{C} = (C_1, \ldots, C_k)$ denote the clause satisfaction verdicts;
- D denote the terminal decision.

The structural equations take the form:

$$C_j = g_j(\mathbf{E}, \pi, \mathbf{N}; U_{C_j}), \quad j = 1, \ldots, k, \tag{3.7}$$

$$\varphi = h(\pi, \mathbf{C}, \mathbf{N}; U_\varphi), \tag{3.8}$$

$$D = q(\mathbf{C}, \varphi, \mathbf{E}; U_D), \tag{3.9}$$

where g_j checks whether clause j is satisfied given the evidence and negotiation context, h selects the protocol given the policy and clause status, and q produces the terminal decision.

The resulting causal graph has a layered structure, illustrated in Figure 2. The causal graph $\mathcal{G}$ encodes the dependency structure of the pipeline: a disclosure fact F_i causally influences a clause check C_j, which in turn influences the protocol selection P_ℓ and ultimately the decision D. Crucially, the graph also encodes the influence of the counterparty's actions through the shared negotiation state.

In practice, the causal graph is populated from an *interaction trace* $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T, d)$, where s_t is the shared state at step t, a_t is the action taken (by either agent), and $d \in \mathcal{D}$ is the terminal decision. The trace provides the observed values of the endogenous variables; the structural equations (3.7)–(3.9) describe how these values were generated. Together, the trace and the structural equations constitute the SCM $\mathcal{M}$ on which counterfactual queries operate.

3.3. **Minimal Counterfactual Identification in the Agent Setting.** We now instantiate the general minimal counterfactual problem (Definition 3.2) on the agent coordination graph. The causal graph determines both the *feasible set* and the *propagation rule* for the optimisation. A counterfactual intervention δ fixes the values of the intervened variables and recomputes all downstream variables by forward-propagating through the structural equations in topological order; variables that are not

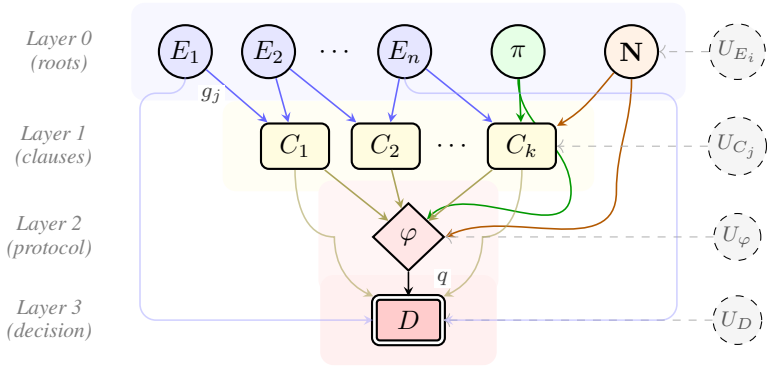


FIGURE 2. Causal graph for the agent coordination pipeline. Solid nodes are endogenous variables; dashed nodes are exogenous (latent). The graph is layered: evidence, policy, and negotiation at Layer 0; clause verdicts at Layer 1; protocol selection at Layer 2; and the terminal decision at Layer 3. Node shapes encode variable type: circles for root inputs, rounded rectangles for clause verdicts, a diamond for the protocol, and a double-bordered rectangle for the terminal decision. Colours reinforce the grouping: blue (evidence), green (policy), orange (negotiation), yellow (clauses), red (protocol/decision). Edge labels g_j and q correspond to the structural equations (3.7) and (3.9). A counterfactual intervention δ sets one or more endogenous variables to new values and propagates downstream through the structural equations.

descendants of the intervention set in $\mathcal{G}$ remain unchanged. The counterfactual decision is

$$d_{\mathcal{M}}(\delta) = q\big(\mathbf{C}^{(\delta)},\, \varphi^{(\delta)},\, \mathbf{E}^{(\delta)};\, U_D\big), \tag{3.10}$$

where $\mathbf{E}^{(\delta)}$, $\mathbf{C}^{(\delta)}$, and $\varphi^{(\delta)}$ are obtained by applying δ and propagating. The minimal counterfactual problem (3.1) becomes a constrained optimisation over the graph:

$$\delta^* = \arg\min_{\delta \in \Delta} \ \mathrm{cost}(\delta) \quad \text{s.t.} \quad q\big(\mathbf{C}^{(\delta)}, \varphi^{(\delta)}, \mathbf{E}^{(\delta)}; U_D\big) \neq d, \tag{3.11}$$

where the feasible set Δ encodes actionability constraints (e.g., the policy π may be fixed, or certain evidence items may be unmodifiable). The graph structure enters in two ways: through the propagation rule (3.10), which determines how each candidate intervention affects the decision, and through the layer structure, which enables the decomposition we establish in the next subsection.

The design of the cost function determines what the resulting counterfactuals mean. In the audit setting we weight individual variables by their *auditability*: a fact that is cheap to verify (a bank balance) receives a low weight, while one requiring extensive due diligence (a revenue projection) receives a high weight, so that the reported counterfactual concentrates on changes an auditor could actually check. Regularisation terms can further shape the solution: an ℓ_1 penalty on the continuous components promotes sparse, plausible counterfactual scenarios over diffuse, adversarial-looking ones. For the discrete components, per-variable flip costs play the same role and can encode that overriding a clause verdict (a policy question) is not commensurable with revising an evidence item (a factual question).

Remark 3.6 (The challenge of mixed variable modalities)**.** Unlike Example 3.3, where all variables are continuous scalars and the decision boundary is a hyperplane, the agent coordination graph involves fundamentally heterogeneous variable types. The evidence variables $\mathbf{E}$ mix continuous quantities (financial metrics, risk scores) with discrete items (identity verification outcomes, document presence flags) and structured objects (natural-language negotiation transcripts). The clause verdicts $\mathbf{C}$ are binary or categorical. The protocol selection φ is a discrete choice from a finite library Φ. This heterogeneity has three consequences for the optimisation (3.11). First, the cost

function $\mathrm{cost}(\delta)$ must handle mixed types: a weighted norm is well-defined for the continuous components, but interventions on discrete variables (e.g., flipping a clause verdict or switching a protocol) require a combinatorial cost that is not naturally captured by a norm. Second, gradient-based methods (DeepFool, projected gradient descent) apply only to the continuous subspace; the discrete components require enumeration or relaxation. Third, the feasible set Δ has a product structure — continuous perturbations on $\mathbf{E}$, combinatorial choices on $\mathbf{C}$ and φ — that the layer decomposition (Proposition 3.7) exploits: the pure-layer subproblems each involve variables of a single modality, and they bound (or, under layer-local pivotality, determine) the global optimum.

3.4. **Decomposition of Counterfactual Interventions.** The layered structure of the causal graph (Figure 2) organises the optimisation problem (3.11) by layer. Partition the endogenous variables as $L_0 = \{\mathbf{E}, \pi, \mathbf{N}\}$ (roots), $L_1 = \{C_1, \ldots, C_k\}$ (clauses), $L_2 = \{\varphi\}$ (protocol), and $L_3 = \{D\}$ (decision); we never intervene on D itself, as D is the quantity we wish to change. Every intervention $\delta \in \Delta$ then decomposes uniquely as a disjoint union $\delta = \delta_0 \cup \delta_1 \cup \delta_2$, where δ_i acts on variables in L_i. We call δ *pure* if only one component is non-empty, and *mixed* otherwise. Pure interventions carry the interpretable audit labels:

(1) **Evidence counterfactuals**: δ_0 acting on the evidence package $\{E_1, \ldots, E_n\}$ — changes to the facts that would alter clause satisfaction. (Interventions on the policy π, which also lives in L_0, are the policy counterfactuals of Section 7 and are typically excluded from Δ in a factual audit; interventions on the transcript $\mathbf{N}$ are handled analogously to evidence.)
(2) **Clause counterfactuals**: $\delta_1 \subseteq \{C_1, \ldots, C_k\}$ — direct overrides of policy clause outcomes (representing policy modifications);
(3) **Protocol counterfactuals**: $\delta_2 = \{\varphi\}$ — a different protocol selection that would change the decision under the same clauses.

For each layer i, define the *layer-restricted margin*

$$m_i = \min_{\delta_i \in \Delta_i} \mathrm{cost}(\delta_i) \quad \text{s.t.} \quad d_{\mathcal{M}}(\delta_i) \neq d, \tag{3.12}$$

where $\Delta_i \subseteq \Delta$ denotes the feasible pure layer-i interventions; we set $m_i = +\infty$ if no pure layer-i intervention flips the decision, and write δ_i^* for a minimiser when one exists.

Proposition 3.7 (Layer decomposition of counterfactual interventions)**.** *Let the causal graph be given by Equations* (3.7)–(3.9)*, and let the cost function be monotone in the sense of Definition* 3.2*. Then:*

(i) ***(Pure-layer upper bound.)*** *The minimal counterfactual cost satisfies*

$$\mathrm{cost}(\delta^*) \;\leq\; \min(m_0, m_1, m_2).$$

(ii) ***(Exactness under layer-local pivotality.)*** *Say the instance satisfies* layer-local pivotality *if every decision-changing intervention $\delta = \delta_0 \cup \delta_1 \cup \delta_2 \in \Delta$ has at least one component δ_i that is decision-changing on its own (holding the other layers at their factual values). Under this condition,*

$$\mathrm{cost}(\delta^*) \;=\; \min(m_0, m_1, m_2),$$

and the minimum is attained by a pure intervention: the minimal counterfactual can be taken to be an evidence, clause, or protocol counterfactual.

When exactness holds, the audit statement reports the class of δ^*, providing a structured explanation: "the decision would have changed if [evidence E_i had value e_i'] / [clause C_j had been satisfied] / [protocol φ' had been selected]." In general, the audit statement reports the per-layer components of δ^* together with the pure-layer margins (m_0, m_1, m_2).

Proof. (i) Each Δ_i is a subset of Δ, and the constraint in (3.12) is the constraint of (3.11) restricted to Δ_i. Hence every feasible point of a layer-restricted problem is feasible for the global problem, and the global minimum is at most $\min(m_0, m_1, m_2)$.

(ii) Let $\delta \in \Delta$ be any decision-changing intervention and let δ_i be a decision-changing component, which exists by layer-local pivotality. Then $\delta_i \in \Delta_i$ is feasible for the layer-i problem (3.12), so $\mathrm{cost}(\delta_i) \geq m_i \geq \min(m_0, m_1, m_2)$. By monotonicity of the cost, $\mathrm{cost}(\delta) \geq \mathrm{cost}(\delta_i)$. Taking the infimum over decision-changing δ gives $\mathrm{cost}(\delta^*) \geq \min(m_0, m_1, m_2)$; combined with (i), equality holds, and the minimum is attained by the cheapest layer-restricted minimiser δ_i^*, which is pure.

It remains to record how the constraint $d_{\mathcal{M}}(\delta) \neq d$ is evaluated. Since the structural equations are known (or learned), we apply Pearl's submodel construction [Pearl, 2009, Definition 7.1.2]: the intervention replaces the structural equations of the intervened variables with the prescribed values, and all downstream variables are recomputed by forward propagation through the remaining (unmodified) equations, with the exogenous variables held at their abducted values. For evidence counterfactuals ($i = 0$), we set $\mathbf{E}^{(\delta_0)}$, recompute $\mathbf{C}^{(\delta_0)}$ via (3.7), then $\varphi^{(\delta_0)}$ via (3.8), and finally $D^{(\delta_0)}$ via (3.9). For clause counterfactuals ($i = 1$), we fix $\mathbf{C}^{(\delta_1)}$ directly and recompute φ and D. For protocol counterfactuals ($i = 2$), we fix $\varphi^{(\delta_2)}$ and recompute D only. □

Remark 3.8 (When is layer-local pivotality plausible?)**.** The condition holds trivially when the feasible set Δ contains only pure interventions, and, more interestingly, whenever the decision flips as soon as a *single* downstream condition toggles. A common instance is a veto-style decision rule — approval requires every clause to pass — audited in the approve-to-reject direction: any decision-changing intervention must make some clause fail in the counterfactual world, and that clause fails either because δ_1 overrides it (so δ_1 alone flips the decision) or because the propagated evidence change makes it fail (so δ_0 alone flips the decision).

The condition is not vacuous, and exactness genuinely fails without it. Consider a *conjunctive* flip: the decision changes only if $C_1 = C_2 = 1$, with observed values $C_1 = C_2 = 0$. Suppose C_1 can be raised by a cheap evidence change (cost 1), while C_2 depends only on π and $\mathbf{N}$, which are not actionable, so the only way to set $C_2 := 1$ is a clause override (cost 10 per override, additively separable cost). Then no pure evidence intervention flips the decision ($m_0 = +\infty$), the cheapest pure clause intervention costs $m_1 = 20$ (both overrides), but the mixed intervention — the evidence change plus a single override of C_2 — flips the decision at cost $11 < 20$. The pure-layer bound of Proposition 3.7(i) is strict, and the minimal counterfactual is irreducibly mixed. The audit statement must then report both components; collapsing it to a single class would misattribute the decision's fragility.

Remark 3.9 (Adversarial robustness per layer)**.** The decomposition yields a *per-layer robustness margin*: write $m_E = m_0$, $m_C = m_1$, and $m_\varphi = m_2$ for the layer-restricted margins (3.12) of the evidence, clause, and protocol layers. The overall counterfactual margin satisfies $\mathrm{cost}(\delta^*) \leq \min(m_E, m_C, m_\varphi)$, with equality under the layer-local pivotality condition of Proposition 3.7(ii), but the triple (m_E, m_C, m_φ) is more informative than any single number: a decision that is robust to evidence changes (m_E large) but fragile with respect to protocol selection (m_φ small) has a qualitatively different risk profile from one that is fragile at the evidence layer. This structured robustness profile has no direct analogue in standard adversarial robustness, where the perturbation space is unstructured.

3.5. **Towards Verifiable Counterfactual Audit.** A counterfactual audit statement—"the decision would have changed if E_i had taken value e_i'"—is only useful if a third party can verify that (i) the interaction trace is authentic, (ii) the causal graph was correctly constructed from the trace, and (iii) the counterfactual propagation was computed correctly. Without verifiability, the audit is only as trustworthy as the auditor itself, which is inadequate in adversarial multi-agent settings where any party may have an incentive to produce misleading explanations.

Recent work on verifiable causality analysis provides a starting point. Song et al. [2026] introduce VCAUSE, a system for verifiable causality analysis over provenance graphs in cloud-based endpoint auditing. Their approach rests on two authenticated data structures: a *graph accumulator* built from hierarchical indexed Merkle trees, which provides efficient cryptographic proofs for individual node queries; and a *verifiable versioned provenance graph*, which computes *incoming* and *outgoing*

path digests for each node, cryptographically committing to the node's full set of causal ancestors and descendants. A key result is that this construction achieves *unforgeability*: no polynomial-time adversary can produce a forged causality analysis result that passes verification (under standard cryptographic assumptions on the signature scheme and hash function).

The structural parallel to our setting is direct. The VCAUSE provenance graph, where nodes represent system entities and edges represent causal dependencies derived from event logs, corresponds to our causal graph $\mathcal{G}$, where nodes are trace variables (E_i, C_j, φ, D) and edges are the structural equations (3.7)–(3.9). Several of their techniques adapt naturally:

(1) **Trace commitment via graph accumulator.** Each step of the interaction trace can be committed to a Merkle-tree-based accumulator as it is recorded. The layered structure of our graph (Figure 2) maps to VCAUSE's hierarchical tree organisation: each layer corresponds to a level in the accumulator, and the layer-by-layer propagation in our proof of Proposition 3.7 corresponds to traversing the tree from leaf (evidence) to root (decision).

(2) **Path digests for causal ancestry.** The incoming path digest Π_I in VCAUSE—which commits to all backward causally related components of a node—can be applied to commit to the causal ancestry of the decision D. A verifier can then check that the counterfactual intervention δ^* targets variables that are indeed causal ancestors of D in the committed graph, preventing the auditor from fabricating causal relationships.

(3) **Segmented digests for layered verification.** VCAUSE introduces segmented outgoing path digests to reduce update overhead from exponential to linear in the graph depth. In our setting, the natural segmentation boundary is the layer boundary: each layer's digest commits to the variables and structural equations within that layer. The per-layer decomposition (Proposition 3.7) then enables *per-layer verification*: a verifier can check an evidence counterfactual by validating only the Layer 0 and Layer 1 digests, without recomputing the full graph.

What VCAUSE does not address is the correctness of the counterfactual propagation itself. VCAUSE verifies that causality analysis was performed on an authentic, untampered graph; it does not verify that a hypothetical intervention was correctly evaluated. In our setting, this amounts to proving that the forward propagation in (3.10) was computed correctly given the intervention δ^* and the committed structural equations.

The layered structure of our causal graph suggests a natural approach. The propagation from intervention to counterfactual decision is a composition of functions (one per layer) applied in sequence:

$$\mathbf{E}^{(\delta)} \xrightarrow{g_1,\dots,g_k} \mathbf{C}^{(\delta)} \xrightarrow{h} \varphi^{(\delta)} \xrightarrow{q} D^{(\delta)}.$$

The resulting structure is that of a *layered arithmetic circuit*, for which efficient interactive proof protocols exist [Goldwasser et al., 2015]. In the GKR protocol [Goldwasser et al., 2015], a prover who evaluates a layered circuit can produce a proof of correctness that the verifier checks in time that grows with the input size and the *depth* of the circuit (number of layers), rather than with its total size. Our four-layer graph would require only four rounds of interaction, regardless of the number of evidence variables or clause checks.

In the full agent coordination setting, the structural equations are more complex: clause checks g_j may involve rule evaluation or learned classifiers, protocol selection h may involve combinatorial optimisation over the library, and the decision function q may involve an LLM call. This motivates a *hybrid verification* strategy that matches the proof technique to the complexity of each layer:

- **Deterministic layers** (threshold checks, rule-based clause evaluation, protocol selection from a finite library): these can be expressed as small arithmetic circuits and verified via succinct non-interactive arguments of knowledge (SNARKs) [Parno et al., 2013, Groth, 2016], or simply re-executed by the verifier if the computation is cheap.
- **Learned layers** (neural network–based clause checks, LLM decision functions): verifiable inference for neural networks [Lee et al., 2024] can produce proofs that a committed model was evaluated correctly on given inputs. This is computationally expensive but feasible for

the moderate-sized surrogate models that would typically serve as structural equations in the SCM (the full LLM is not the structural equation; a distilled decision model is).

- **Opaque layers** (full LLM calls that cannot be expressed as circuits): the auditor commits to the input–output pair and provides a *replay guarantee*: the verifier can re-query the same model with the same input and check that the output matches. This is weaker than a cryptographic proof but sufficient when the model is deterministic (or when the auditor commits to the random seed).

The per-layer decomposition (Proposition 3.7) is essential here: the verifier only needs to check the layers *downstream* of the intervened variables (for a mixed intervention, downstream of all of its components). A protocol counterfactual (δ_2) requires verifying only the decision function q; an evidence counterfactual (δ_0) requires verifying all three downstream layers, but the cost is still linear in the depth rather than exponential in the graph size.

Even without full cryptographic verification, the combination of VCAUSE-style trace commitment (ensuring the graph is authentic) with replay-based propagation checks (ensuring the counterfactual was computed on the committed graph with the committed intervention) already rules out a significant class of attacks in which the auditor manipulates the trace, the graph structure, or the propagation to produce a desired counterfactual conclusion.

4. COMPUTATIONAL APPROACH

We propose two complementary computational strategies for solving the minimal counterfactual problem.

Strategy 1: Replay-based counterfactual search. Given a recorded trace τ, we re-execute the coordination pipeline with modified inputs. For each candidate intervention δ, we replay the Negotiation Safety Core and Security Auditor with the perturbed evidence, clauses, or protocol choice, and observe whether the decision changes. This approach is exact (no approximation) but may be expensive for large intervention spaces. We propose to make it tractable via:

(1) **Gradient-guided search**: When the clause-checking and decision functions are differentiable (or admit differentiable surrogates), use gradient descent on $\mathrm{cost}(\delta)$ subject to the constraint $d_{\mathcal{M}}(\delta) \neq d$. This follows the approach of Mothilal et al. [2020].

(2) **Causal pruning**: Use the causal graph to restrict the search to ancestors of D in $\mathcal{G}$, avoiding interventions on variables that provably cannot change the decision: an intervention $\mathrm{do}(V_i := v_i')$ can affect only the descendants of V_i, so any V_i that is not an ancestor of D can be excluded from Δ [Pearl, 2009].

(3) **Importance sampling over traces**: For stochastic coordination protocols, use importance sampling to estimate the probability that a given intervention changes the decision, prioritising high-probability interventions.

Strategy 2: Learned counterfactual generator. Train a generative model G_θ that, given a trace τ and decision d, directly produces a minimal counterfactual intervention δ. The training objective combines:

$$\mathcal{L}(\theta) = \underbrace{\mathbb{E}_{\tau\sim\mathcal{T}}\left[\mathrm{cost}(G_\theta(\tau))\right]}_{\text{minimality}} + \mu \cdot \underbrace{\mathbb{E}_{\tau\sim\mathcal{T}}\left[\mathbf{1}[d_{\mathcal{M}}(G_\theta(\tau)) = d]\right]}_{\text{validity penalty}}, \tag{4.1}$$

where $\mathcal{T}$ is the distribution of recorded traces from the evaluation environment. The generator G_θ can be trained on traces from the Lending Simulator and then fine-tuned on traces from the transfer domains (procurement, insurance), testing whether the counterfactual structure generalises.

5. COUNTERFACTUAL CONFIDENCE AND ROBUSTNESS

A critical risk is that counterfactual audit outputs may become post-hoc rationalisations. We address this with three formal safeguards:

(1) **Replay validation**: For each reported counterfactual δ^*, the Security Auditor re-executes the full coordination pipeline with the intervention applied and verifies that the decision indeed changes. This produces a binary *replay-valid* flag.
(2) **Counterfactual confidence score**: Define the *counterfactual robustness* of an intervention δ as the probability that it remains decision-changing under perturbations to the exogenous variables:

$$\rho(\delta) = \Pr_{U \sim P(U|\tau)} [d_{\mathcal{M}}(\delta; U) \neq d] \,, \tag{5.1}$$

where $d_{\mathcal{M}}(\delta; U)$ denotes the counterfactual outcome of Definition 3.1 computed with the exogenous variables set to U, and $P(U \mid \tau)$ is the abduction posterior given the trace. (When abduction point-identifies U, as in Definition 3.1, $\rho(\delta) \in \{0, 1\}$; the score is informative precisely when the trace leaves residual uncertainty about the exogenous variables.) A counterfactual with $\rho(\delta) < \rho_{\min}$ (e.g., $\rho_{\min} = 0.8$) is flagged as *fragile* in the audit statement, indicating that the decision change depends on specific assumptions about hidden state.
(3) **Multiplicity reporting**: When multiple distinct minimal counterfactuals exist, the audit statement reports the full set (or a representative subset), avoiding the impression that there is a single "reason" for the decision. This follows the Rashomon-set approach applied to explanations [Mothilal et al., 2020].

6. Counterfactual Identifiability via Dynamic Optimal Transport

The framework developed so far computes counterfactuals by forward-propagating interventions through the structural equations (Section 3). This presumes the equations are known. For the deterministic layers of the coordination graph — clause checks against explicit policy, protocol selection from a finite library — they are, and counterfactuals are computable exactly. But the terminal decision function q of an LLM-based agent, and a fortiori the counterparty's disclosure strategy σ, are *opaque*: we observe interaction traces (the joint distribution over evidence, clause verdicts, protocol states and decisions) without access to the mechanism that generated them. A counterfactual computed through a fitted surrogate of q is then only as trustworthy as the surrogate, and two different surrogates that fit the observed distribution equally well may disagree about the counterfactual — precisely the ambiguity an audit statement cannot afford.

This is the classical problem of *counterfactual identifiability*: which counterfactual queries are determined by the observational (or interventional) distribution alone, without knowledge of the structural equations? For bijective or noise-monotone mechanisms, exact counterfactual identifiability is available [Nasr-Esfahany et al., 2023]; for discrete mechanisms, Oberst and Sontag [2019] make the non-identifiability explicit by fixing a particular counterfactual coupling (the Gumbel-max SCM) — a modelling choice rather than an identified object. Recent work by De Sousa Ribeiro et al. [2025] gives a constructive answer with a distinctly transport-theoretic flavour: under regularity conditions, the Benamou–Brenier formulation of dynamic optimal transport singles out the *unique monotone* transport map from exogenous noise to observed outcomes, and the counterfactuals computed through this map are provably consistent with every SCM *in the monotone class* compatible with the observed data — identification is relative to the class of noise-monotone mechanisms, not to all mechanisms. Monotonicity here plays the role that the abduction step plays when the equations are known: it pins down how the latent noise is held fixed across the factual and counterfactual worlds.

Application to the layered audit graph. The layered structure of Section 3.2 lets us apply this result surgically rather than globally. We propose a three-tier epistemic regime, refining the hybrid verification strategy of Section 3.5:

(1) *Known layers* (clause checks g_j, protocol selection h): counterfactuals computed exactly through the declared equations; identifiability is not at issue.

(2) *Opaque continuous layers* (an LLM lender's effective decision score, as revealed by the decision-boundary sweeps of Section 8): estimate the monotone transport map from logged traces; the resulting counterfactual is *canonical* — consistent with any monotone-mechanism SCM generating the observed distribution — under the regularity conditions of De Sousa Ribeiro et al. [2025].
(3) *Opaque discrete outcomes* (the terminal approve/reject decision itself): point identification generally fails; the transport framework instead yields *set-valued* counterfactuals — bounds on the set of decisions consistent with the observed data, in the spirit of bounds on probabilities of causation [Tian and Pearl, 2000] — which the audit statement reports as such.

Epistemic labelling of audit statements. This regime upgrades the honesty of the audit output. Every reported counterfactual carries, alongside its confidence score $\rho(\delta)$ (Section 5), an *identifiability status*: *identified* (computed through known equations, or through the canonical monotone transport), *bounded* (set-valued, with the bounds stated), or *model-dependent* (computed through a fitted surrogate whose choice matters, with the surrogate declared). A decision-flip claim labelled *identified* holds for every data-consistent monotone model of the counterparty; one labelled *model-dependent* is an artefact of a modelling choice until strengthened. For adversarial audit settings — where the party being audited has every incentive to dispute the auditor's modelling choices — this distinction determines which audit findings are contestable.
Caveats. Monotonicity is a substantive assumption for LLM decision mechanisms: it must hold in the chosen representation (e.g. a scalar risk score revealed by boundary sweeps, not the raw token-level mechanism), can fail across model versions (Section 8 documents boundary drift between rule-based and LLM-based lenders), and requires care for the mixed continuous/discrete variables of Section 3. We regard the transport route not as a universal identification device but as the principled fallback for exactly the layers where surrogate modelling would otherwise smuggle unexamined assumptions into audit conclusions.

7. Counterfactual Analysis Under Strategic Behaviour

In adversarial settings, the counterparty may strategically withhold or misrepresent evidence to manipulate the decision. A naïve counterfactual analysis that assumes the evidence package is exogenous will produce misleading results. We extend the framework to account for strategic evidence provision by treating the counterparty's disclosure strategy σ as an additional endogenous variable:

$$\mathbf{E} = \sigma(\mathbf{E}^{\text{true}}, \pi^{\text{cp}}), \tag{7.1}$$

where $\mathbf{E}^{\text{true}}$ is the true state and π^{cp} is the counterparty's policy (which is only partially observable). The counterfactual analysis then distinguishes between:

- *Factual counterfactuals*: "if the true state had been different" (intervention on $\mathbf{E}^{\text{true}}$);
- *Disclosure counterfactuals*: "if the counterparty had disclosed differently" (intervention on σ);
- *Policy counterfactuals*: "if our policy had been different" (intervention on π).

This decomposition is particularly valuable for the lending proving ground, where the distinction between "the borrower was truly risky" and "the borrower failed to disclose material information" is decision-critical for audit purposes.
Legitimate versus manipulation-induced policy change. The decomposition also yields an operational handle on a subtler audit question. The local policy state may change *during* an interaction through sanctioned channels — escalation clauses, renegotiation moves defined in the protocol library, protocol-mediated updates. Write $\pi \to \pi'$ for a policy update reachable through this sanctioned transition set. We call the update *legitimate* if it survives the disclosure counterfactual $\text{do}(\sigma := \sigma^{\text{truthful}})$ — that is, the same update would have occurred had the counterparty

disclosed truthfully — and *manipulation-induced* otherwise: the update is then an artefact of the counterparty's strategic disclosure rather than of the underlying facts. This is the operational-policy analogue of distinguishing legitimate from illegitimate value change in AI systems, and it equips the auditor with a machine-checkable test: replay the sanctioned transition under the truthful-disclosure counterfactual and compare outcomes.

8. Illustrative Experiment: Decision Boundaries in a Lending Simulator

To ground the framework in a concrete multi-agent setting, we apply the counterfactual analysis to decisions produced by a lending simulator,[1] a competitive multi-agent environment in which LLM-based models act as autonomous loan underwriters evaluating borrower dossiers. Each dossier contains continuous financial variables (annual revenue, net income, monthly cash flow), discrete attributes (sector, years in business), and structured data (bank statements, quarterly income reports). The lender agent produces a binary decision (approve or reject) based on standard underwriting criteria: debt-service coverage, net margin, loan-to-revenue ratio, and profitability.

Experiment design. We take a single borrower (BRW-001, a logistics company with \$2.4M revenue, 27% net margin, requesting a \$500k loan) and sweep two continuous variables — the loan request amount and the annual revenue — over a grid while holding the expense base and all other dossier fields fixed. Crucially, annual expenses are held at BRW-001's base level (\$1.78M) rather than scaled proportionally with revenue; this means that lowering revenue genuinely compresses the net margin rather than preserving it, producing a realistic variation in borrower strength. For each grid point, the lender agent evaluates the modified dossier and produces an approve/reject decision. This traces out a two-dimensional cross-section of the decision surface, analogous to fixing U_S and varying (X_1, X_2) in Example 3.3.

Observations. The resulting decision boundary (Figure 3) exhibits several features that the toy linear example of Section 3.1 cannot capture.

First, the boundary is *non-linear*. It follows a hyperbolic curve driven by the loan-to-revenue ratio constraint (rejection when the ratio exceeds ≈ 0.55), combined with absolute floors on revenue and debt-service coverage that create the steep rise at low revenue values. This is precisely the setting where the weighted-ℓ_∞ closed-form margin (3.6) does not apply, and the full constrained optimisation (3.11) is needed.

Second, the *counterfactual direction* is economically interpretable. The arrow from the observed point to the nearest reject point indicates that the decision is most sensitive to simultaneous increases in the loan amount and decreases in revenue — the leverage ratio. An auditor can read this as: "the approval would have been reversed if the borrower had requested \$250k more while earning \$750k less."

Third, the boundary is *asymmetric* in the two variables. A horizontal cross-section at fixed revenue shows a sharp approval cliff as the loan amount increases, while a vertical cross-section at fixed loan amount shows a more gradual transition. This asymmetry means the cheapest decision-changing perturbation depends strongly on which variable absorbs it: the evidence-layer margin m_E of Remark 3.9 conceals a directional structure, and reporting the per-variable minimal flips alongside m_E provides the auditor with a structured robustness profile rather than a single scalar margin.

From rule-based to LLM-based boundaries. The boundary in Figure 3 was generated using a rule-based underwriting evaluator that applies standard financial ratio thresholds. We repeat the experiment with an LLM-based lender (DeepSeek-V3-0324 via OpenRouter, temperature 0, 20×20 grid, 400 evaluations). The system prompt instructs the model to apply the same three criteria—DSCR ≥ 1.25, net margin $\geq 10\%$, and loan-to-revenue ratio ≤ 0.55—and to return a binary approve/reject decision.

The results (Figure 4) reveal three phenomena absent from the rule-based boundary.

[1] Available at https://github.com/seadotdev/Simulator-Loanville.

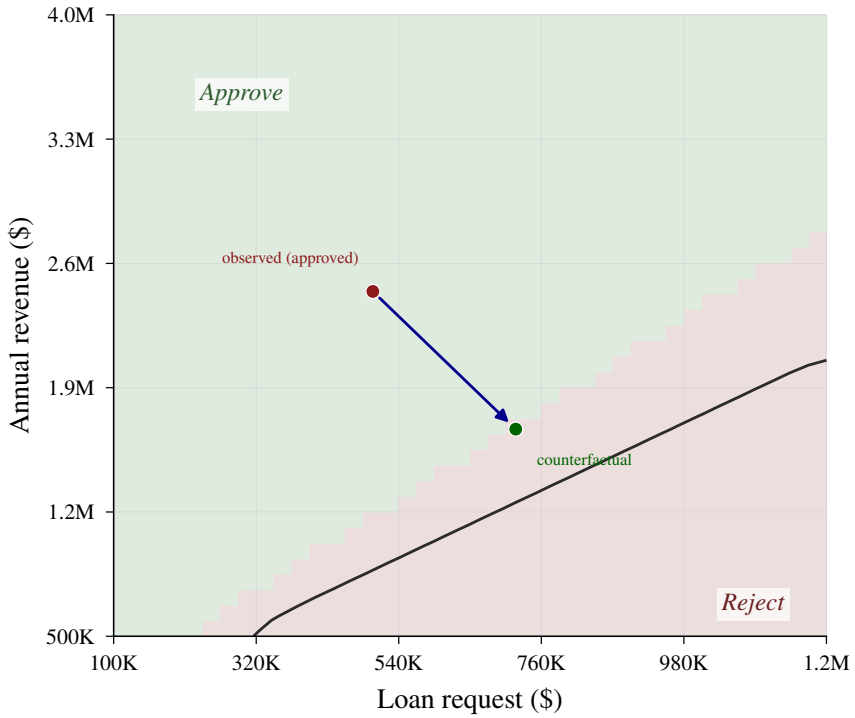


FIGURE 3. Decision boundary from the Lending Simulator, sweeping loan request amount (horizontal) against annual revenue (vertical) for borrower BRW-001. The green region indicates approval; the red region indicates rejection. The boundary is non-linear: it follows a hyperbolic curve reflecting the loan-to-revenue ratio constraint, combined with a debt-service coverage floor that creates the steep rise at low revenue. The observed borrower ($500k loan, $2.4M revenue) lies in the approve region. The counterfactual arrow shows the minimal perturbation — increasing the loan and decreasing the revenue — that would move the borrower across the boundary into the reject region. A 40×40 grid (1,600 evaluations) was used.

First, the LLM boundary is *nearly horizontal*: the model approves 325 of 400 evaluations, and rejection depends almost entirely on annual revenue (equivalently, net margin) rather than loan size. Above $\sim$\$2M revenue (margin $\gtrsim 12\%$), the LLM approves uniformly, even at a loan-to-revenue ratio of 0.75—a ratio that the rule-based evaluator would reject. The model thus effectively ignores the DSCR and leverage constraints it was instructed to apply, relying instead on profitability as a single sufficient signal.

Second, the boundary exhibits *stochastic scatter* in the transition zone. At \$1.89M revenue (margin $\approx 6\%$), the model produces a mixture of approvals and rejections across loan amounts with no monotone pattern. This is not sampling noise (temperature is zero); it reflects sensitivity of the deterministic decode to small changes in the numerical content of the prompt. The implied counterfactual margin in this zone is effectively zero: a borrower near the boundary can be flipped by a negligible perturbation to a single financial variable, and the direction of the perturbation is unpredictable.

Third, the *counterfactual direction differs* between evaluators. For the rule-based boundary, the minimal counterfactual from the observed point involves simultaneously increasing the loan and decreasing revenue (the leverage ratio); the arrow is diagonal. For the LLM boundary, the minimal counterfactual is a pure revenue decrease: the arrow points straight down, because the model is insensitive to the loan dimension. This means that an audit statement generated under the rule-based model ("the decision would have changed if the loan had been \$250k larger *and* revenue \$750k

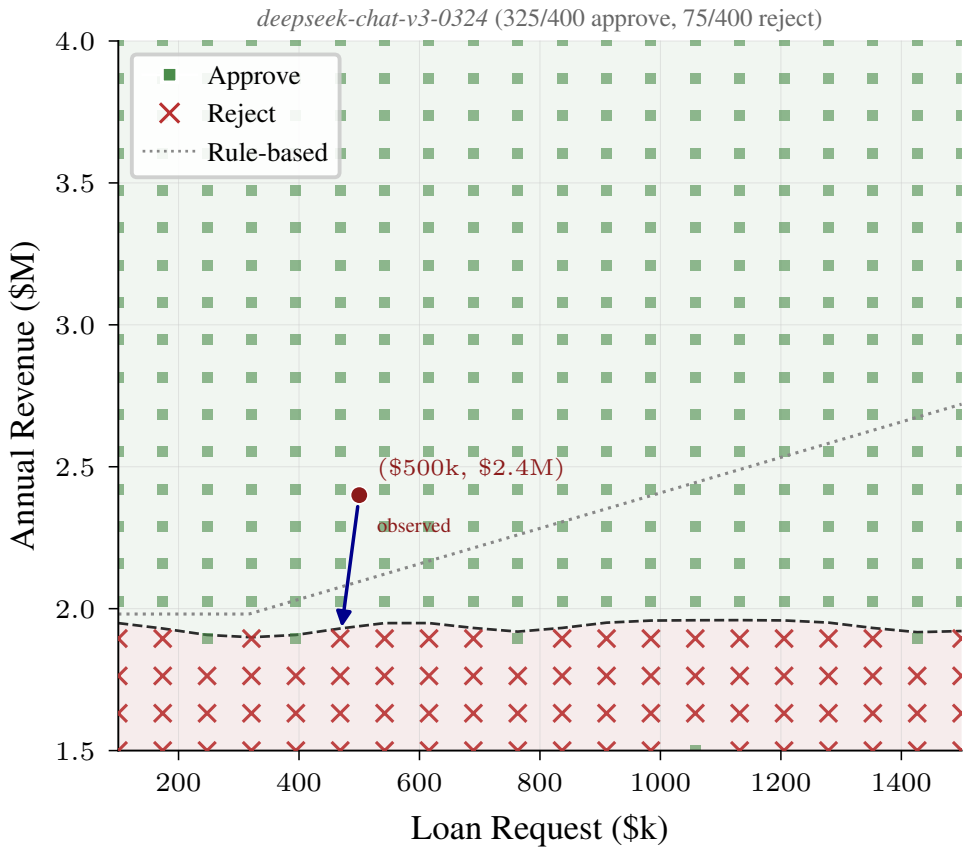


FIGURE 4. LLM-based decision boundary, sweeping loan request amount (horizontal) against annual revenue (vertical) for borrower BRW-001 with fixed expenses. Green squares indicate approval; red crosses indicate rejection. The dashed curve is the smoothed LLM boundary; the dotted curve is the rule-based boundary from Figure 3 (combining the margin floor, DSCR, and loan-to-revenue constraints). The LLM boundary is nearly horizontal, driven almost entirely by net margin: above ∼$2M revenue (margin $\gtrsim$12%), the model approves regardless of loan size. The rule-based boundary, by contrast, rises with the loan amount as the DSCR constraint binds. The noisy transition zone at ∼$1.9M (margin ≈6%) contains scattered approvals among rejections. The counterfactual arrow shows the minimal perturbation from the observed point to the nearest rejection.

lower") would be qualitatively different from one generated under the LLM ("the decision would have changed if revenue had been $450k lower"). The two counterfactuals implicate different risk factors and would lead to different remedial actions—a concrete instance of the non-stationarity that the counterfactual confidence score $\rho(\delta)$ of Section 5 is designed to detect.

9. CONCLUSION

We have developed a formal framework for counterfactual audit analysis in multi-agent negotiation systems. The framework models the coordination pipeline as a structural causal model, defines minimal counterfactual interventions as constrained optimisations over the causal graph, and exploits the layered structure of the graph to organise counterfactuals into three interpretable classes — evidence, clause, and protocol counterfactuals — with the cheapest single-class intervention bounding the counterfactual margin from above, exactly so under the layer-local pivotality condition of Proposition 3.7.

Several aspects of this work merit emphasis. First, the correspondence between counterfactual identification and adversarial robustness — established in the explanation literature [Freiesleben, 2022, Pawelczyk et al., 2022] — becomes a working tool in the audit setting: the counterfactual margin and the adversarial distance are instances of the same optimisation problem, and the closed-form margin formulae we derive via Lagrange duality (for the weighted ℓ_∞ norm natural to audit cost functions) are the direct counterparts of the SVM margin in classification. This connection imports a mature body of algorithms (DeepFool, projected gradient descent, Carlini–Wagner) and fundamental limits (Fawzi's theorem on the inevitability of small margins for expressive models) into the audit setting.

Second, the per-layer decomposition is load-bearing for both interpretability and verification. For interpretability, it tells the auditor *at which level* the decision is fragile: whether a small change to the evidence, the policy clauses, or the protocol selection would have altered the outcome. For verification, it enables layer-by-layer checking of counterfactual claims: a protocol counterfactual requires verifying only the decision function, while an evidence counterfactual requires propagating through all downstream layers, but the cost remains linear in the graph depth.

Third, the extension to strategic counterfactuals — distinguishing between factual, disclosure, and policy interventions — addresses the fundamental challenge that in adversarial coordination, the evidence itself is the output of an optimisation by a self-interested counterparty. Without this distinction, an audit that finds "the decision would have changed if the revenue had been higher" cannot differentiate between a genuinely low-revenue borrower and one who strategically underreported.

Limitations and future work. Several directions remain open. The current framework assumes a known (or partially known) causal graph; learning the graph structure from interaction traces, particularly when the counterparty's mechanism is opaque, is a significant challenge that connects to the causal discovery literature. The transport-based identification of Section 6 addresses the opaque-*mechanism* half of this problem, but assumes the graph itself; joint identification of structure and counterfactuals remains open. The verifiable counterfactual audit construction (Section 3.5) provides integrity guarantees for the trace and causal ancestry, but verifying the correctness of the counterfactual propagation itself, particularly through learned or LLM-based structural equations, remains open and likely requires advances in verifiable computation for neural networks. The robustness measures we introduce (counterfactual confidence scores, multiplicity reporting) are defined but not yet empirically validated at scale; evaluation on real multi-agent coordination traces is a natural next step. Finally, extending the framework from single-episode counterfactuals to sequential settings, where the audit question is "which turn in a multi-round negotiation was decision-critical?", would connect the layered graph structure to the temporal structure of interaction traces, a direction with rich connections to dynamic treatment regimes and path-dependent causal inference.

REFERENCES

George A Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

Reek Das and Biplab Kanti Sen. FedAOT: Dynamic meta-layer aggregation for Byzantine-robust federated learning. *arXiv preprint arXiv:2603.16846*, 2026.

Fabio De Sousa Ribeiro, Ainkaran Santhirasekaram, and Ben Glocker. Counterfactual identifiability via dynamic optimal transport, 2025.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *Advances in Neural Information Processing Systems*, 31, 2018.

Timo Freiesleben. The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, 32(1):77–109, 2022.

Shafi Goldwasser, Yael Tauman Kalai, and Guy N Rothblum. Delegating computation: Interactive proofs for muggles. *Journal of the ACM*, 62(4):27:1–27:64, 2015. Preliminary version in STOC 2008.

Sanford J Grossman. The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3):461–483, 1981.

Jens Groth. On the size of pairing-based non-interactive arguments. In *Advances in Cryptology – EUROCRYPT 2016*, pages 305–326, 2016.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 111–122, 2016.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55: 1–29, 2022.

Juhee Kim, Xiaoyuan Liu, Zhun Wang, Shi Qiu, Bo Li, Wenbo Guo, and Dawn Song. The attack and defense landscape of agentic AI: A comprehensive survey. *arXiv preprint arXiv:2603.11088*, 2026.

Sebastian Krier. Coasean bargaining at scale. Cosmos Institute Blog, 2025. https://blog.cosmos-institute.org/p/coasean-bargaining-at-scale.

Seunghwa Lee, Hankyung Ko, Jihye Kim, and Hyunok Oh. vCNN: Verifiable convolutional neural network based on zk-SNARKs. *IEEE Transactions on Dependable and Secure Computing*, 2024. See also zkML frameworks: EZKL, Giza, Modulus Labs.

Paul Lezeau and Martin Lotz. Tubular neighbourhoods of pfaffian sets and applications to neural networks, 2026. URL https://arxiv.org/abs/2607.08370.

Ninghui Li, Kaiyuan Zhang, Kyle Polley, and Jerry Ma. Security considerations for artificial intelligence agents. *arXiv preprint arXiv:2603.12230*, 2026. Perplexity Response to NIST/CAISI Request for Information 2025-0035.

Martin Lotz. *Mathematics of Machine Learning*. University of Warwick, 2024. Lecture notes, Mathematics Institute.

Paul R Milgrom. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 12(2):380–391, 1981.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

Rawal Kaur Mothilal, Amit Sharma, and Chenhao Tan. DiCE: Diverse counterfactual explanations for machine learning classifiers. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

Roger B Myerson. Incentive compatibility and the bargaining problem. *Econometrica*, 47(1): 61–73, 1979.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Bryan Parno, Jon Howell, Craig Gentry, and Mariana Raykova. Pinocchio: Nearly practical verifiable computation. In *IEEE Symposium on Security and Privacy (SP)*, pages 238–252, 2013.

Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

Juan C Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Qiyang Song, Qihang Zhou, Xiaoqi Jia, Zhenyu Song, Wenbo Jiang, Heqing Huang, Yong Liu, and Dan Meng. vCAUSE: Efficient and verifiable causality analysis for cloud-based endpoint auditing. *arXiv preprint arXiv:2603.15216*, 2026.

Joseph E Stiglitz and Andrew Weiss. Credit rationing in markets with imperfect information. *The American Economic Review*, 71(3):393–410, 1981.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1–4):287–313, 2000.

Andrew Trask, Emma Bluemke, Ben Garfinkel, Claudia Ghezzou Cuervas-Mons, and Allan Dafoe. Beyond privacy trade-offs with structured transparency. *arXiv preprint arXiv:2012.08347*, 2020.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31: 841–887, 2018.

Poom Wangsomcharoen and Piotr Dworczak. The coasean singularity: Demand, supply, and market design for AI agents. In *Economics of Transformative AI*. National Bureau of Economic Research, 2025. Forthcoming.

WARWICK MATHEMATICAL INSTITUTE, UNIVERSITY OF WARWICK, UK
Email address: martin.lotz@warwick.ac.uk

WARWICK MATHEMATICAL INSTITUTE, UNIVERSITY OF WARWICK, UK
Email address: pietro.aluffi@warwick.ac.uk

SEA.DEV, LONDON, UK
Email address: marya@sea.dev

SEA.DEV, LONDON, UK
Email address: matt@sea.dev

SEA.DEV, LONDON, UK