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, and prove that the resulting minimal counterfactual interventions decompose into three disjoint, interpretable classes: evidence, clause, and protocol counterfactuals. We establish a precise correspondence between counterfactual identification and adversarial robustness and 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., 2025, Li et al., 2025] 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.

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; 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 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.

Information economics and strategic disclosure. The distinction between factual and disclosure counterfactuals (Section 6) draws on the information economics of adverse selection [Akerlof, 1970, Stiglitz and Weiss, 1981]. In the agentic setting, a counterparty can optimise its disclosure strategy in real time. Myerson [1979]'s revelation principle 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, 2025], where adaptive aggregation without parametric attack assumptions informs our counterfactual confidence measures.

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]. The key insight we build on (Section 3.2) is that the adversarial distance and the counterfactual margin $\mathrm{cost}(\delta^*)$ are instances of the same optimisation with different cost functions and feasible sets. In particualr, theoretical and computational results for adversarial robustness carry over to counterfactual audit.

Verifiable causality analysis. Song et al. [2025] 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 3.4.

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. 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

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

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

We require cost to be *separable* 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 $\mathrm{cost}(\delta) = \mathrm{cost}(\delta_A) + \mathrm{cost}(\delta_B)$. Since costs are non-negative, separability implies the monotonicity property $\mathrm{cost}(\delta_A \cup \delta_B) \geq \max\big(\mathrm{cost}(\delta_A), \mathrm{cost}(\delta_B)\big)$, which is essential for the layer-wise decomposition in Proposition 3.7. All cost functions considered in this paper (weighted ℓ_p norms, combinatorial per-variable costs) are separable.

The cost function $\mathrm{cost}(\delta)$ must be chosen to produce meaningful counterfactuals; its design is domain-specific and discussed in detail in Section 3.2. We first illustrate both definitions with a toy example.

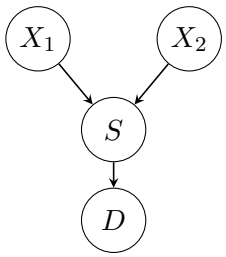
Example 3.3 (A simple SCM for a loan decision)**.** Consider a minimal lending scenario with four endogenous variables: *income* $X_1 \in \mathbb{R}_+$, *debt* $X_2 \in \mathbb{R}_+$, *credit score* $S \in [0, 1]$, and *decision* $D \in \{\text{approve}, \text{reject}\}$. The structural equations are:

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

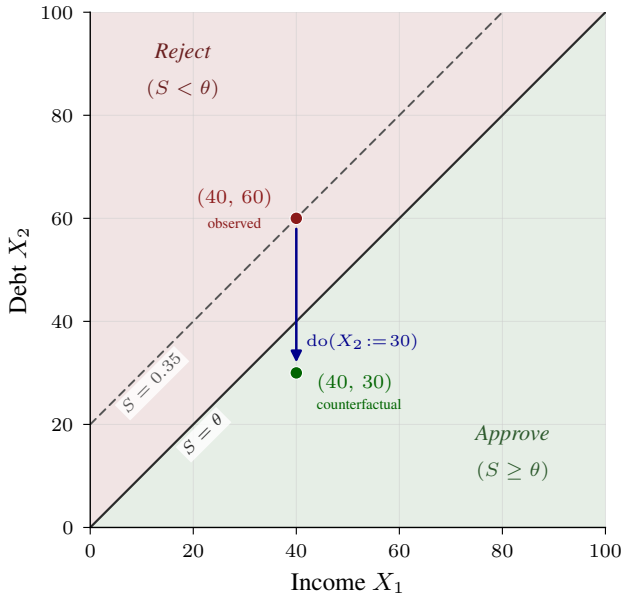
where σ is the sigmoid function, $\alpha, \beta > 0$ are fixed weights, θ is the approval threshold, and U_1, U_2, U_S are independent exogenous variables. Here U_S captures all determinants of the credit score beyond income and debt (payment history, credit utilisation, length of credit history, and

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 $\alpha = \beta = 1, U_S = 0$) are shown in Figure 1.

Suppose we observe a rejected applicant with $X_1 = 40$, $X_2 = 60$, $S = 0.35$, $D =$ reject (with $\theta = 0.5$). 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.



(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 $\alpha = \beta = 1$, $U_S = 0$. The solid line is $S = \theta$; the dashed line is the level set $S = 0.35$. The intervention $\mathrm{do}(X_2 := 30)$ moves the applicant from $(40, 60)$ in the reject region to $(40, 30)$ in the approve region.

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 $S = \sigma(\mathbf{a}^T\mathbf{x} + U_S)$ 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 $b = U_S - \sigma^{-1}(\theta)$. 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

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

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 $\|\cdot\|_*$,

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

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 $|\lambda| \cdot \|\mathbf{a}\|_{w^{-1},1} \le 1$, and $-\infty$ otherwise. The dual problem is therefore

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

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 sparsity pattern of the optimal perturbation. When all active coordinates (those achieving the maximum in $\|\delta^*\|_{w,\infty}$) have distinct weights, the perturbation is unique.

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 $\mathbf{a} = (1, -1)^T$, $b = 0$, 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 has several consequences.

The first consequenceis that the fundamental limits on adversarial robustness carry over. In the classification setting, Fawzi et al. [2018] showed that if the data generating process $g\colon \mathbb{R}^m \to \mathbb{R}^d$ is L-Lipschitz, then for any classifier the probability of being within distance ϵ of the decision boundary satisfies $\mathbb{P}\{\hat{\Delta}(X) \leq \epsilon\} \geq 1 - \sqrt{\pi/2}\, e^{-\epsilon^2/2L^2}$. When L is large relative to ϵ—as it is for expressive models such as deep networks—most points lie near a boundary, and adversarial perturbations are unavoidable. The analogue for the audit setting is that when the decision pipeline is composed of high-capacity learned components (as in LLM-based agents), the counterfactual margin will generically be small, 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.

3.1. **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.

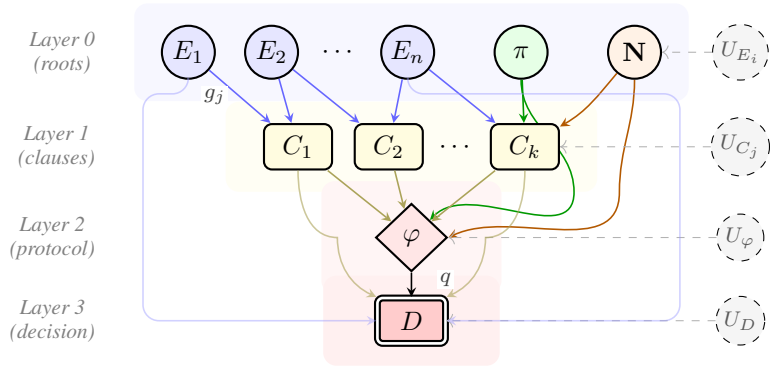


FIGURE 2. Causal graph for the agent coordination pipeline. Solid nodes are endogenous variables; dashed nodes are exogenous (latent). The graph is layered: evidence, policy, and negotiation at Layer 0; clause verdicts at Layer 1; protocol selection at Layer 2; and the terminal decision at Layer 3. Node 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.

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.

3.2. **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 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.

Remark 3.6 (The challenge of mixed variable modalities.)**.** Unlike the toy 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 by separating the optimisation into independent subproblems, each involving variables of a single modality.

3.3. **Decomposition of Counterfactual Interventions.** The layered structure of the causal graph (Figure 2) enables a decomposition of the optimisation problem (3.11) into independent subproblems, one per layer.

Proposition 3.7 (Decomposition of Counterfactual Interventions)**.** *Under the causal graph defined by Equations* (3.7)*–*(3.9)*, a minimal counterfactual intervention δ^* can be decomposed into three disjoint classes:*

(1) ***Evidence counterfactuals****: $\delta_E \subseteq \{E_1, \ldots, E_n\}$ — changes to the evidence package that would alter clause satisfaction;*
(2) ***Clause counterfactuals****: $\delta_C \subseteq \{C_1, \ldots, C_k\}$ — direct changes to policy clause outcomes (representing policy modifications);*
(3) ***Protocol counterfactuals****: $\delta_\varphi = \{\varphi\}$ — a different protocol selection that would change the decision under the same clauses.*

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]."

Proof. We exploit the layered structure of the causal graph. Partition the endogenous variables by layer: $L_0 = \{\mathbf{E}, \pi, \mathbf{N}\}$ (roots), $L_1 = \{C_1, \ldots, C_k\}$ (clauses), $L_2 = \{\varphi\}$ (protocol), $L_3 = \{D\}$ (decision). Since the graph is acyclic with edges only from L_i to L_j for $i < j$, every intervention $\delta \in \Delta$ can be written as a disjoint union $\delta = \delta_0 \cup \delta_1 \cup \delta_2$, where δ_i acts on variables in L_i. (We do not intervene on D itself, as D is the quantity we wish to change.)

For each layer i, define the *layer-restricted* optimisation:

$$\delta_i^* = \arg\min_{\delta_i \in \Delta \cap L_i} \mathrm{cost}(\delta_i) \quad \text{s.t.} \quad d_{\mathcal{M}}(\delta_i) \neq d. \tag{3.12}$$

Step 1: Independence across layers. Consider an intervention $\delta = \delta_0 \cup \delta_1$. The variables in L_1 are deterministic functions of L_0 via the structural equations (3.7). If δ_0 changes some evidence variables $\mathbf{E}' \subset \mathbf{E}$, the clause values update to $C_j^{(\delta_0)} = g_j(\mathbf{E}^{(\delta_0)}, \pi, \mathbf{N}; U_{C_j})$. If δ_1

additionally sets $C_j := c'_j$ for some j, this *overrides* the value propagated from δ_0. Therefore, for any mixed intervention $\delta_0 \cup \delta_1$ with $d_{\mathcal{M}}(\delta_0 \cup \delta_1) \neq d$, either δ_0 alone already flips the decision (and is at least as cheap), or δ_1 alone flips it when applied to the *original* clause values (and is at least as cheap as the override). Formally, by separability of the cost function (Definition 3.2), $\mathrm{cost}(\delta_0 \cup \delta_1) = \mathrm{cost}(\delta_0) + \mathrm{cost}(\delta_1) \geq \max\big(\mathrm{cost}(\delta_0),\, \mathrm{cost}(\delta_1)\big)$, so the global optimum is attained by a single-layer intervention.

Step 2: Layer-by-layer solution. By Step 1, we solve three independent problems (3.12) for $i = 0, 1, 2$ and take the cheapest:

$$\delta^* = \arg\min\big\{\mathrm{cost}(\delta_0^*),\, \mathrm{cost}(\delta_1^*),\, \mathrm{cost}(\delta_2^*)\big\}.$$

Each δ_i^* is, by construction, a pure evidence, clause, or protocol counterfactual.

Step 3: Counterfactual evaluation via truncated factorisation. For each layer-restricted intervention δ_i, we need to evaluate the counterfactual decision $d_{\mathcal{M}}(\delta_i)$. 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. This is the *truncated factorisation* — the joint distribution over the non-intervened variables factors as a product of the original conditional distributions, with the factors corresponding to the intervened variables removed and replaced by point masses. 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, recompute φ and D. For protocol counterfactuals ($i = 2$), we fix $\varphi^{(\delta_2)}$ and recompute D only. □

Remark 3.8 (Adversarial robustness per layer)**.** The decomposition yields a *per-layer robustness margin*: define

$$\rho_i = \mathrm{cost}(\delta_i^*), \qquad i \in \{E, C, \varphi\},$$

as the cost of the cheapest evidence, clause, and protocol counterfactual, respectively. The overall counterfactual margin is $\rho = \min(\rho_E, \rho_C, \rho_\varphi)$, but the triple $(\rho_E, \rho_C, \rho_\varphi)$ is more informative: a decision that is robust to evidence changes (ρ_E large) but fragile with respect to protocol selection (ρ_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.

3.4. **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. [2025] 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 [Amit et al., 2024]. In the GKR protocol [Goldreich et al., 1991], a prover who evaluates a layered circuit can produce a proof of correctness that the verifier checks in time proportional to the *depth* of the circuit (number of layers) rather than 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) [Amit et al., 2024], 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: since evidence, clause, and protocol counterfactuals are independent, the verifier only needs to check the layers *downstream* of the intervention. A protocol counterfactual (δ_φ) requires verifying only the decision function q; an evidence counterfactual (δ_E) 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. Specifically, if V_i is d-separated from D given the remaining variables, it 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:

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

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 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:

$$\mathbf{E} = \sigma(\mathbf{E}^{\mathrm{true}}, \pi^{\mathrm{cp}}), \tag{6.1}$$

where $\mathbf{E}^{\mathrm{true}}$ 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 $\mathbf{E}^{\mathrm{true}}$);
- *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.

7. 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 all other dossier fields fixed. 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 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

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

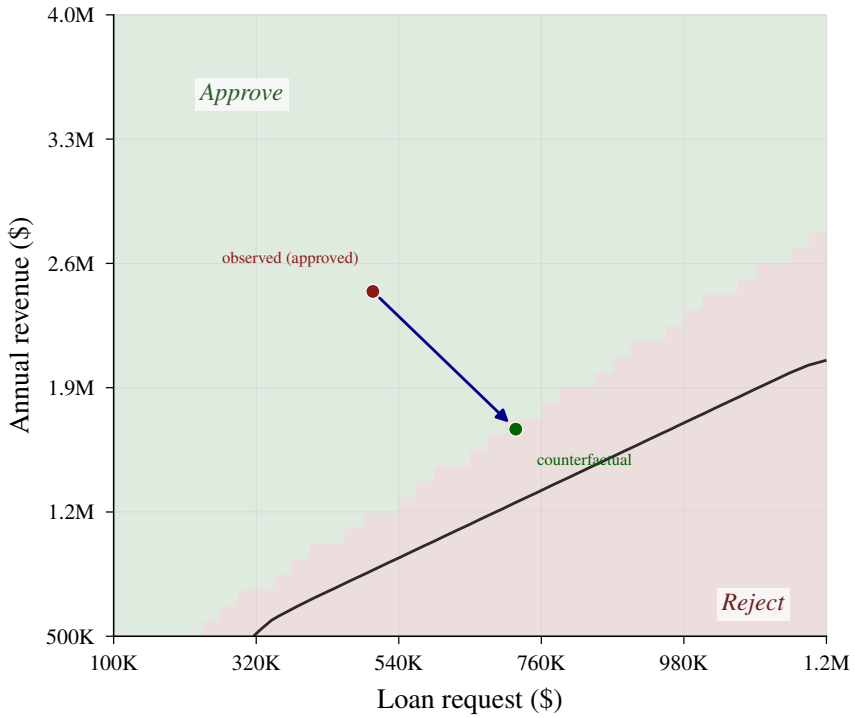


FIGURE 3. Decision boundary from the Lending Simulator, sweeping loan request amount (horizontal) against annual revenue (vertical) for borrower BRW-001. The green region indicates approval; the red region indicates rejection. The boundary is non-linear: it follows a hyperbolic curve reflecting the loan-to-revenue 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.

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 would produce different per-variable counterfactual margins ρ_{E_1} and ρ_{E_2} in the notation of Remark 3.8, providing 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. When the same experiment is run with an LLM-based lender, the boundary becomes smoother but less predictable: the LLM integrates soft signals from the narrative description and sector context that the rule-based evaluator ignores, and its decision surface may shift between model versions. This non-stationarity reinforces the need for the robustness measures developed in Section 5: a counterfactual that is valid under one model checkpoint may become invalid under the next, and the counterfactual confidence score $\rho(\delta)$ quantifies this fragility.

8. 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 decompose counterfactuals into three interpretable classes: evidence, clause, and protocol counterfactuals.

Several aspects of this work merit emphasis. First, the duality between counterfactual identification and adversarial robustness is not merely an analogy: 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 verifiable counterfactual audit construction (Section 3.4) 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.

Noga Amit, Shafi Goldwasser, Orr Paradise, and Guy N Rothblum. Models that prove their own correctness. *arXiv preprint arXiv:2405.15722*, 2024.

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*, 2025.

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

Oded Goldreich, Silvio Micali, and Avi Wigderson. Proofs that yield nothing but their validity, or, all languages in NP have zero-knowledge proof systems. In *Journal of the ACM*, volume 38, pages 691–729, 1991.

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.

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

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*, 2025.

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

Seunghwa Lee et al. vCNN: Verifiable convolutional neural network inference. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024. See also zkML frameworks: EZKL, Giza, Modulus Labs.

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

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

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.

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

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*, 2025.

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.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. *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: martin.lotz@warwick.ac.uk

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

SEA.DEV, LONDON, UK

SEA.DEV, LONDON, UK