SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Causal Inference Frameworks For Platform Integrity: Moving Beyond Correlation In Fraud And Abuse Detection

Sanchit Suman

Independent Researcher, USA ORCID ID: 0009-0005-7545-9945

Abstract
Detecting fraudulent activity across large-scale digital platforms has long relied on correlation-driven classifiers built from labeled transaction records. While effective under stable conditions, such models degrade when policies shift, adversaries evolve, or label quality deteriorates through operational feedback cycles. This article contends that causal inference offers a more rigorous foundation for platform integrity work, allowing practitioners to evaluate the effects of interventions rather than merely surface statistical associations. Analytical methods covered span DAG specification, propensity score modeling, doubly robust procedures, and heterogeneous treatment effect estimation via causal forests, each situated within dispute resolution workflows and fraud governance practice. Publicly available evidence of cross-market performance variation anchors the empirical claims. Consumer fraud losses reported by the Federal Trade Commission exceeded ten billion US dollars in 2023; industry benchmarks place global e-commerce fraud losses at roughly 2.9 percent of revenue for that same year. A production readiness checklist closes the article, accompanied by discussion of unresolved problems in outcome labeling latency, instrument validity, and federated estimation across market boundaries.

1. Introduction

Platform integrity teams at large consumer technology companies face a persistent measurement problem. The ground truth for fraud is never directly observed. What analysts observe is a combination of user behavior, policy-triggered actions, and downstream dispute filings, each of which is partially determined by prior model decisions. A transaction that a model classifies as low-risk is less likely to be reviewed and therefore less likely to generate a fraud label. This circularity is not merely inconvenient. It means that labeled datasets carry the imprint of prior policies, and models trained on those datasets will partially learn to replicate past decisions rather than identify true fraud.

Correlation-based machine learning methods, including gradient-boosted trees and deep neural networks, have achieved strong performance on standard fraud benchmarks. Card payment fraud is heavily concentrated in the United States: the Nilson Report (2025) documents that the country's 25.29 percent share of global card volume corresponds to 42.32 percent of global fraud losses. That gap between spend share and loss share has direct consequences for the rigor required of detection systems operating at this scale. At such scale, even marginal improvements in precision have measurable economic consequences. However, high benchmark performance does not translate automatically to policy-relevant predictions.

The core difficulty is confounding. When an analyst asks whether a new velocity limit policy would reduce fraud rates in a given market, the answer requires estimating the counterfactual outcome under a policy that has not yet been applied. Correlation-based models cannot answer this question because they have no representation of intervention. Causal inference methods, by contrast, are designed precisely for this setting. They take as

input a structural model of the data-generating process and produce estimates of the expected outcome under a specified action.

This paper provides a structured review of causal inference methods applicable to platform fraud detection, with particular attention to dispute management, cross-market heterogeneity, and operational deployment constraints. Section 2 compares the assumptions underlying correlation-based and causal models. Section 3 describes the causal reasoning pipeline from DAG construction to effect validation. Section 4 taxonomizes the confounders most commonly encountered in platform fraud data. Section 5 reviews counterfactual estimation methods. Section 6 presents an adaptive intervention framework for cross-market deployment. The paper's penultimate section walks through an operational readiness framework, while the final section draws together the key conclusions.

2. Correlation-Based vs. Causal Model Assumptions

The distinction between correlation-based and causal models is not merely technical. It reflects a difference in the question being asked. Correlation-based models ask: given that this transaction has these features, how likely is it to be fraudulent? Causal models ask: if we apply this policy to this transaction, what will happen to the outcome? These are different questions, and they require different frameworks.

Table 1 summarizes the key differences across five dimensions: objective, treatment of confounding, generalization scope, model output type, and characteristic failure mode. The practical implications of each dimension are listed in the rightmost column.

Dimension	Correlation-Based Models	Causal Models	Practical Impact
Objective	Predict outcome from features	Estimate effect of intervention	Determines whether model can support policy decisions
Confounding	Not explicitly addressed	Identified and blocked via DAG	Affects whether model outputs can be acted upon
Generalization	Within training distribution	Across intervention conditions	Determines deployment stability under policy changes
Model output	Score or classification	Treatment effect estimate	Affects how model results are used operationally
Failure mode	High false positive rate under distribution shift	Requires correct DAG specification	Different risks in production fraud systems

The failure mode column in Table 1 is particularly important for production systems. A correlation-based model will typically exhibit degraded precision under distribution shift, generating an elevated false positive rate when user behavior changes due to a new policy or an adversarial campaign. A causal model, by contrast, fails if the specified DAG is incorrect. Neither failure is invisible, but they require different monitoring strategies. Distribution shift can be detected through feature drift metrics and precision-recall tracking. DAG misspecification requires domain expert review and sensitivity analysis, which are less commonly automated.

The implication for platform integrity teams is that the choice of model type should be driven by the intended use of the output. If the goal is to score transactions for real-time review, a well-calibrated classifier may be sufficient. If the goal is to estimate the effect of a policy intervention or to allocate enforcement resources across markets, a causal model is required.

3. The Causal Reasoning Pipeline

Implementing causal inference in a production fraud system requires a structured workflow that begins before any data is collected and continues through post-deployment monitoring. The following flowchart represents the six-stage pipeline used in causal fraud analysis.

Step	Description
1. Define intervention	Specify the fraud policy action to evaluate (e.g., account suspension, velocity limit)
2. Construct DAG	Map causal relationships between features, fraud labels, and outcomes
3. Identify confounders	Apply do-calculus to identify adjustment sets
4. Estimate propensity	Fit propensity model to reweight observational data
5. Estimate causal effect	Apply doubly robust or CATE estimator to treatment assignment
6. Validate and monitor	Test counterfactual predictions on holdout; monitor for distribution shift

The first stage, defining the intervention, is often treated as obvious but is frequently underspecified in practice. A velocity limit policy, for example, may apply at the account level, the device level, or the card level. Each of these targeting choices implies a different estimand and a different identification strategy. Analysts who skip this step often produce causal estimates that answer a question no stakeholder was asking.

DAG construction in the second stage requires collaboration between data scientists and domain experts who understand the operational process. In fraud settings, the DAG must encode the relationship between user behavior, model scores, review decisions, dispute filings, and fraud labels. Every directed edge in a DAG encodes a posited causal pathway, while every absent edge asserts that no direct effect exists between the corresponding nodes. These assumptions are falsifiable in principle but require careful specification.

The do-calculus step in stage three translates the DAG into an identification strategy. If the DAG permits identification of the causal effect through backdoor adjustment, the analyst constructs an adjustment set and proceeds to propensity estimation. If identification requires an instrumental variable or front-door adjustment, the required data structures must be confirmed to exist before proceeding. Stage four and five then apply the chosen estimator, and stage six validates the result on held-out data and establishes monitoring protocols for production.

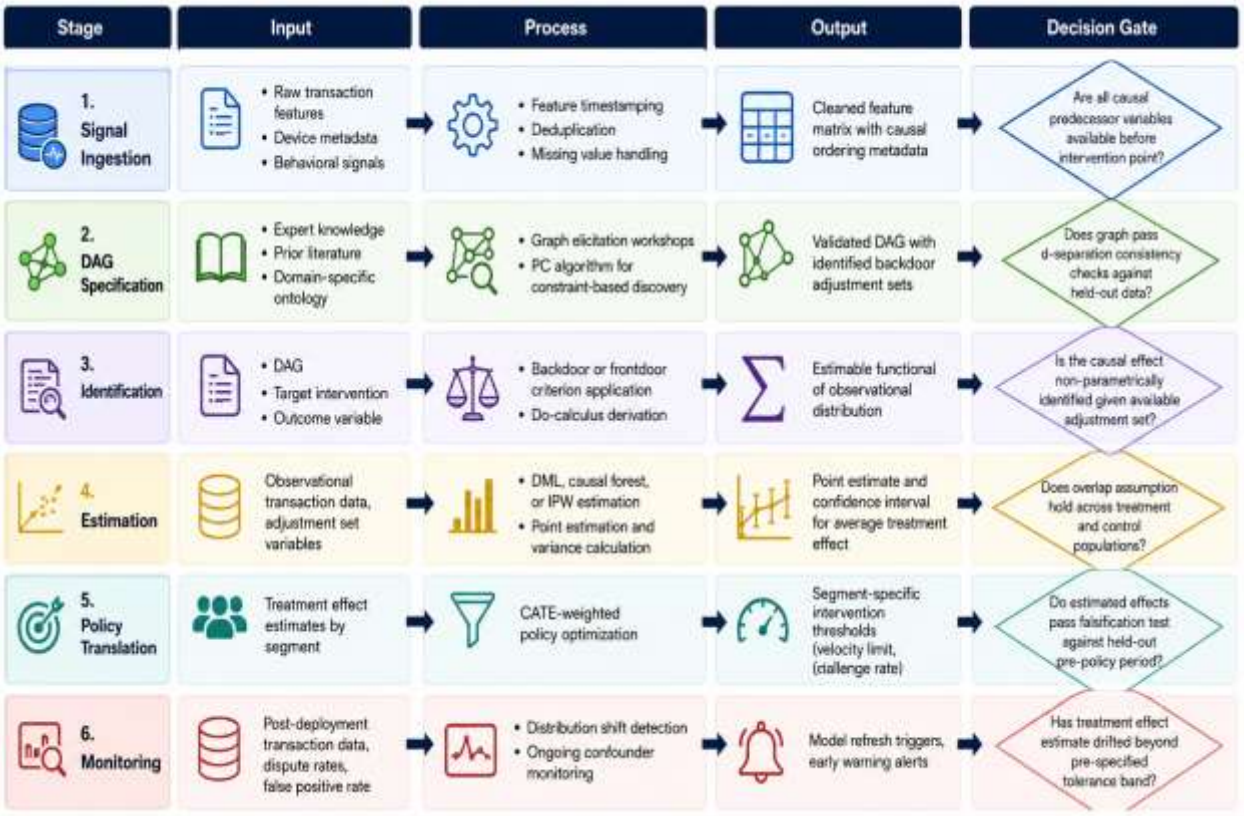

Figure 1. Causal Reasoning Pipeline for Platform Fraud Policy

4. Confounder Taxonomy in Platform Fraud Data

Confounders in platform fraud data are more varied and more operationally specific than those typically discussed in the causal inference literature. The following taxonomy covers the five categories most frequently encountered in dispute management and fraud label generation.

Confounder Type	Example in Fraud Context	Mitigation Strategy
Selection bias	High-value accounts flagged more often	Propensity score weighting
Temporal confounding	Fraud patterns change seasonally	Time-stratified analysis
Geographic confounding	Regional payment norms affect fraud rates	Market-fixed effects
Label confounding	Disputes filed only when users engage	Survival model adjustment
Feedback loop	Past model decisions affect future fraud labels	Counterfactual correction

A systematic imbalance emerges in fraud datasets when investigative capacity is limited. Accounts of higher value attract disproportionate review attention, accumulating fraud labels at a rate that reflects operational prioritization rather than true risk. When a model ingests this data without adjusting for unequal review exposure, it conflates resource allocation patterns with genuine fraud signals. Propensity score weighting addresses this distortion by rescaling each observation according to the reciprocal likelihood of its having been selected for review.

Temporal confounding is ubiquitous in fraud data because fraud patterns are not stationary. Seasonal promotions, regulatory changes, and adversarial campaign cycles all produce temporal variation in fraud rates that is unrelated to any specific intervention. Time-stratified analysis, in which causal effects are estimated separately within narrow time windows, reduces the influence of these seasonal patterns on the estimated treatment effect.

Geographic confounding reflects the fact that payment norms, user behavior, and regulatory environments vary substantially across markets. MRC/Cybersource survey data for 2023 recorded a 4.6 percent fraud rate across Latin American platforms, compared with 2.4 percent in North America; the same survey found Latin American operators devoting roughly 19 percent of revenue to fraud prevention, close to double the proportion seen among North American and European peers. These differences are not random. They reflect differences in card infrastructure, chargeback policy, and consumer protection law. Market-fixed effects absorb this variation to the extent that it is stable within market over time.

Label confounding arises from the structure of the dispute process. Fraud labels in most platform systems are generated by user-initiated dispute filings. Users who disengage from the platform after a fraudulent transaction never file a dispute, producing a censored label that is systematically related to user behavior. Survival models that account for the time-to-dispute distribution can correct for this censoring.

Feedback loop confounding is perhaps the most structurally challenging. When a model assigns a high fraud score to a transaction, that transaction is more likely to be reviewed, blocked, or reversed. The subsequent absence of a fraud label reflects the model's intervention, not the underlying fraud rate. Over time, this creates a dataset in which the model's prior predictions are embedded in the labels used to retrain it. Counterfactual correction, using the known prior policy scores to reweight the training distribution, is the standard remediation strategy.

5. Counterfactual Estimation Approaches for Dispute Management

Once the DAG has been specified and the adjustment set identified, the analyst must choose an estimator. The choice depends on sample size, data dimensionality, the assumed smoothness of the treatment effect surface, and computational constraints. Table 3 summarizes the five most commonly used approaches in platform fraud settings.

Method	Key Assumption	Fraud Application	Limitation
Propensity score matching	No unmeasured confounders	Match disputed vs. non-disputed transactions	Requires large sample; sensitive to model misspecification
Inverse probability weighting	Correct propensity model specification	Reweight observed fraud label distribution	Unstable with extreme propensity scores
Doubly robust estimation	Either outcome or propensity model correct	Estimate counterfactual dispute rate	Computationally intensive at scale
Causal forests (CATE)	Unconfoundedness; overlap	Estimate heterogeneous treatment effects by segment	Requires sufficient data per market segment
Instrumental variables	Valid instrument exists	Isolate exogenous variation in fraud label arrival	Difficult to find valid instruments in platform data

Propensity score matching is the simplest of the five methods and remains widely used in industry settings where interpretability is required. In dispute management, the method matches disputed transactions to non-disputed transactions with similar propensity scores, creating a pseudo-experimental comparison. The matched sample is then analyzed using standard regression methods. The method's main limitation is sensitivity to propensity model misspecification, which becomes acute when the feature space is high-dimensional or when the propensity score distribution is highly skewed.

Inverse probability weighting (IPW) avoids the discretization implicit in matching by reweighting the entire sample. Each observation is weighted by the inverse of its probability of receiving the treatment, so that the reweighted sample approximates a randomized experiment. IPW estimates are unbiased under correct propensity model specification but can be highly variable when estimated propensities are near zero or one. Trimming or truncation of extreme weights is standard practice.

Doubly robust estimation combines an outcome model with a propensity model in a way that produces a consistent estimate if either model is correctly specified. This robustness property is particularly valuable in platform settings where neither the fraud outcome model nor the review propensity model can be verified with certainty. Empirical evaluations have documented substantial reductions in false positives using causal methods: one study at ECML PKDD 2018 reported a 54 percent reduction in false positives and savings of 190,000 euros across 1.852 million transactions using a dispute-aware causal correction.

Causal forests, implemented through the generalized random forest framework, extend the doubly robust approach to estimation of conditional average treatment effects (CATE). Rather than reporting a single population-level effect, causal forests produce an individualized treatment effect for each observation. In fraud settings, this enables the identification of market segments, user populations, or transaction types where a given policy has the strongest effect. When decision tree classifiers are deployed within intervention-aware fraud frameworks, savings against no-action baselines can reach 71.72 percent, as documented in JRFM (2026)—a figure that underlines the real-world payoff from accurate individualized treatment effect estimation.

When unobserved confounders resist adjustment through standard backdoor methods, instrumental variables (IV) offer an alternative identification route. A qualifying instrument must shift the treatment probability without any pathway to the outcome that bypasses the treatment itself. Within platform fraud contexts, candidate instruments include jurisdiction-specific regulatory amendments that alter review policy selectively, or intermittent infrastructure disruptions that introduce quasi-random variation in review capacity. Instruments satisfying the exclusion restriction are scarce in practice, making formal sensitivity analysis a prerequisite before IV-based findings are used to underpin policy changes.

6. Market-Adaptive Intervention Framework

Cross-market deployment introduces heterogeneity that a single global model cannot adequately capture. Average-level performance may appear acceptable while accuracy in individual markets—where fraud patterns, user behavior, or regulatory requirements diverge sharply from the global baseline—remains poor.

The market-adaptive framework responds by estimating causal effects at the market level and consolidating those estimates through a federated meta-learner.

Market Segment	Adaptive Intervention Parameter
Latin America (4.6% fraud rate)	Higher velocity thresholds; regional dispute lookup; CATE estimated on local transaction graph
North America (2.4% fraud rate)	Standard velocity rules; propensity-matched holdout controls
Europe (GDPR-constrained)	Feature set restricted to non-PII; survival model for label delay
Asia-Pacific	Time-zone-adjusted label arrival windows; geographically stratified CATE
Global cross-market	Federated CATE aggregation; meta-learner trained on pooled market embeddings

Latin America provides a concrete illustration of why market-specific calibration is necessary. Latin American platforms face a fraud rate of 4.6 percent and channel 19 percent of revenue into prevention efforts, placing them in a fundamentally different risk position from North American or European operators. Velocity controls calibrated for North American transaction patterns may impose excessive friction in Latin America, where payment infrastructure and usage norms produce distinct transaction cadences; market-level tuning resolves this mismatch. The CATE estimator deployed in this region draws on the local transaction graph rather than the global training corpus, allowing it to reflect confounding structures specific to that market.

The European market presents a different set of constraints. GDPR restrictions limit the feature set available for modeling, excluding many identifiers that would otherwise be informative. The survival model for label delay is particularly important in Europe because consumer protection law creates long windows during which disputes can be filed, generating substantial label censoring. A naive model trained on observed dispute rates would underestimate fraud exposure in markets with long statutory dispute windows.

The Asia-Pacific market introduces time-zone variation in label arrival patterns. Transactions processed in the Asia-Pacific region generate dispute filings that arrive during business hours in that region, which may be overnight for the fraud operations team. A model that uses label arrival time as a feature will inadvertently encode this time-zone effect as a risk signal. Time-zone-adjusted label arrival windows correct for this by normalizing arrival times to a common reference frame before feature extraction.

Global cross-market aggregation uses a federated approach in which market-level CATE estimates are aggregated using a meta-learner trained on pooled market embeddings. This preserves market-specific calibration while enabling the detection of global patterns that no single market has sufficient data to identify independently. The federated approach also provides a natural mechanism for markets with limited sample sizes to borrow strength from larger markets without directly pooling data, which may be prohibited by data localization requirements.

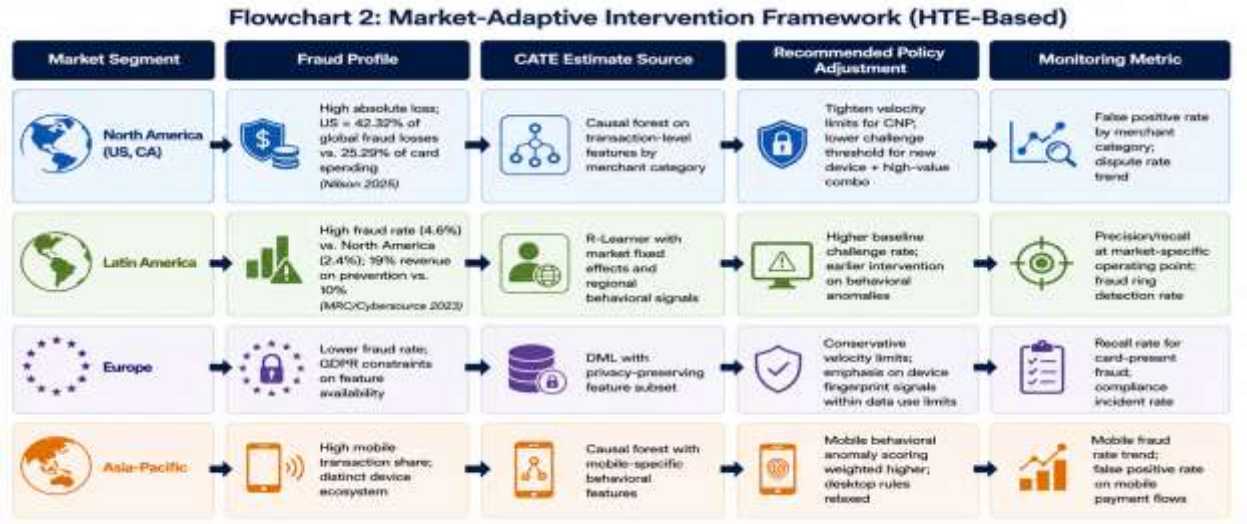

Figure 2. Market-Adaptive Intervention Framework (HTE-Based)

7. Operational Readiness for Causal Inference Deployment

Deploying a causal inference system in a production fraud environment requires readiness across six dimensions. Table 4 provides a checklist that organizations can use to assess their readiness before deployment and to establish the monitoring requirements that must be met after deployment.

Readiness Dimension	Assessment Criteria	Status Indicator
DAG specification	Causal graph reviewed by domain experts	Required before deployment
Confounder coverage	All major confounders identified and measured	Required before deployment
Overlap check	Sufficient overlap in treatment/control groups	Required before deployment
Estimation pipeline	CATE estimator tested on holdout data	Required before deployment
Monitoring framework	Concept drift detection active	Required post-deployment
Human review integration	Escalation path defined for low-confidence scores	Required post-deployment

DAG specification and confounder coverage are prerequisites for any causal analysis. A DAG that has not been reviewed by domain experts cannot be trusted to correctly encode the causal structure of the fraud-generating process. Similarly, a confounder set that is incomplete will produce biased estimates regardless of the sophistication of the estimation method. These two requirements are listed as required before deployment rather than merely recommended because their absence invalidates the causal interpretation of any downstream estimates.

The overlap check is a technical prerequisite for propensity-based methods. Causal effect estimation requires that every unit in the sample has a positive probability of receiving either treatment or control. When overlap is violated, because all units in a subgroup receive the same treatment, the counterfactual cannot be estimated from the data without extrapolation. Overlap violations are common in fraud settings because enforcement policies are often nearly deterministic for extreme risk scores.

The estimation pipeline requirement ensures that the CATE estimator has been tested on held-out data before deployment. In fraud settings, held-out testing must account for temporal ordering: the test set should consist of data from a later time period than the training set, not a random split. Random splits allow the model to use future information implicitly through shared seasonal patterns, producing optimistic performance estimates.

After launch, the primary operational imperatives are ongoing distributional monitoring and structured human oversight. Adversarial adaptation by fraudsters causes the feature distribution to evolve continuously, making drift detection an active rather than periodic concern. Cases where the causal effect estimate carries low confidence should route to human analysts for adjudication rather than triggering automatic enforcement—a safeguard that is especially critical during early deployment, before the system's production behavior has been empirically validated.

8. Discussion and Open Challenges

Several open challenges limit the applicability of causal inference methods to platform fraud systems in their current form. The most significant is label delay. In most platform dispute systems, the true fraud label is not available at the time of the transaction. It arrives days or weeks later, after the user has had an opportunity to file a dispute. This delay creates a temporal mismatch between the features observed at transaction time and the label used for training and evaluation. Survival models address this partially, but they require correct specification of the censoring mechanism, which is itself a causal assumption.

Instrument validity is a second open challenge. Instrumental variables methods require instruments that affect the treatment but have no direct effect on the outcome. In platform settings, candidate instruments are rare and often imperfect. Regulatory changes affect both review policy and user behavior. Infrastructure events affect transaction volume as well as review throughput. The sensitivity of IV estimates to instrument validity

violations means that results obtained from imperfect instruments should be interpreted with caution and subjected to formal sensitivity analysis before being used to justify policy changes.

A third challenge is the computational cost of causal methods at platform scale. Doubly robust estimation and causal forests both require fitting multiple models, and CATE estimation over billions of transactions requires distributed computing infrastructure that is not yet standardized. Approximate methods, including subsampling and kernel approximations, can reduce computational cost at the expense of statistical efficiency, but the tradeoffs between cost and accuracy have not been systematically evaluated in production fraud settings.

Finally, the organizational challenge of deploying causal methods should not be underestimated. Causal inference requires a level of domain knowledge investment that is not required for correlation-based methods. DAG construction requires structured collaboration between data scientists, fraud analysts, and product managers. This collaboration is difficult to sustain in organizations where analytical workflows are highly automated. Building the organizational capacity for causal reasoning is as important as building the technical infrastructure for it.

9. Conclusion

This paper has argued that causal inference frameworks offer a principled and operationally valuable alternative to correlation-based models for platform integrity applications. The central claim is that fraud policy decisions require counterfactual reasoning, and that correlation-based models are structurally incapable of supporting such reasoning. DAG-based causal models, combined with doubly robust estimation and CATE methods, provide estimates that are interpretable as policy-relevant quantities rather than mere predictive scores.

The reviewed evidence supports the practical value of this approach. A 54 percent reduction in false positives documented in a controlled evaluation demonstrates that causal correction of label confounding produces measurable operational gains. Cross-market evidence from Latin America, North America, and Europe establishes that adaptive estimation approaches are necessary to accommodate divergent fraud rates, regulatory environments, and label generation dynamics.

Section 7's readiness checklist gives practitioners a structured basis for evaluating whether their organizations are positioned to deploy these methods. The six readiness dimensions reflect the technical, organizational, and monitoring requirements that must be satisfied before causal estimates can be trusted as the basis for policy decisions. Meeting these requirements is non-trivial, but the alternative, continuing to make policy decisions based on associative predictions that cannot answer causal questions, carries its own risks in an environment where consumer fraud losses exceeded 10 billion dollars in a single year.

References

[1] J. Pearl, "Causal diagrams for empirical research," Biometrika, vol. 82, no. 4, pp. 669-688, 1995.
[2] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," Biometrika, vol. 70, no. 1, pp. 41-55, 1983.
[3] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," JASA, vol. 113, no. 523, pp. 1228-1242, 2018.
[4] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," The Annals of Statistics, vol. 47, no. 2, 2019.
[5] V. Chernozhukov et al., "Double/debiased machine learning for treatment and structural parameters," The Econometrics Journal, vol. 21, no. 1, pp. C1-C68, 2018.
[6] X. Nie and S. Wager, "Quasi-oracle estimation of heterogeneous treatment effects," Biometrika, vol. 108, no. 2, pp. 299-319, 2021.
[7] G. W. Imbens, "Potential outcome and directed acyclic graph approaches to causality," Journal of Economic Literature, vol. 58, no. 4, pp. 1129-1179, 2020.
[8] H. R. Varian, "Causal inference in economics and marketing," PNAS, vol. 113, no. 27, pp. 7310-7315, 2016.
[9] P. Cui and S. Athey, "Stable learning establishes some common ground between causal inference and machine learning," Nature Machine Intelligence, vol. 4, no. 2, pp. 110-115, 2022.
[10] R. Guo et al., "A survey of learning causality with data," ACM Computing Surveys, vol. 53, no. 4, art. 75, 2020.
[11] J. Pearl, "The seven tools of causal inference," Communications of the ACM, vol. 62, no. 3, pp. 54-60, 2019.

[12] L. Yao et al., "A survey on causal inference," ACM TKDD, vol. 15, no. 5, art. 74, 2021.
[13] S. Milli, L. Belli, and M. Hardt, "Causal inference struggles with agency on online platforms," in Proc. ACM FAccT 2022, pp. 357-365, 2022.
[14] S. Athey and S. Wager, "Policy learning with observational data," Econometrica, vol. 89, no. 1, pp. 133-161, 2021.
[15] Y. Cui et al., "Estimating heterogeneous treatment effects with right-censored data via causal survival forests," JRSS-B, vol. 85, no. 2, pp. 179-211, 2023.
[16] E. Esenogho et al., "A neural network ensemble with feature engineering for improved credit card fraud detection," IEEE Access, vol. 10, pp. 16400-16407, 2022.
[17] X. Li et al., "CausalFD: Causal invariance-based fraud detection against camouflaged preference," Int. J. Machine Learning and Cybernetics, vol. 15, no. 11, pp. 5053-5070, 2024.
[18] J. Debener, V. Heinke, and J. Kriebel, "Detecting insurance fraud using supervised and unsupervised machine learning," Journal of Risk and Insurance, vol. 90, no. 3, pp. 743-768, 2023.
[19] H. Hosseinmardi et al., "Causally estimating the effect of YouTube's recommender system using counterfactual bots," PNAS, vol. 121, no. 8, e2313377121, 2024.
[20] L. Jiao et al., "Causal inference meets deep learning: A comprehensive survey," Research, vol. 7, art. 0467, 2024.
[21] A. Gruzd, F. B. Soares, and P. Mai, "Trust and safety on social media," Social Media + Society, vol. 9, no. 3, 2023.