

The Causal Content of Game-Theoretic Models

Robert J. Carroll*

Abstract

Political scientists routinely use game-theoretic models to make causal claims, but without a principled account of when those claims are empirically grounded. I argue that Dawid (2000)'s decision-theoretic framework for causal inference provides that account. The extensive form game tree is structurally isomorphic to Dawid's augmented directed acyclic graph: chance nodes map to exogenous random variables, decision nodes to intervention nodes, information sets to observable covariate histories, and—crucially—strategies (complete contingency plans) to treatment regimes. A strategy profile together with a defensible solution concept constitutes a *determining concomitant* in Dawid's sense, licensing causes-of-effects inference when the game is correctly specified. This gives political scientists a precise criterion: effects-of-causes queries in strategic settings are tractable (*sheep*); causes-of-effects queries are tractable only when the game provides a determining concomitant, and intractable in one-shot games where equilibrium paths never visit the relevant node (*goats*). I demonstrate with an application to candidate entry and incumbency advantage.

*Department of Political Science, University of Illinois at Urbana-Champaign, rjc@illinois.edu

1 Introduction

When political scientists explain why conflicts do not happen, why incumbents go unchallenged, why coercive demands are met without resistance, they reach for game-theoretic logic. The explanation runs through equilibrium: the outcome we observe is sustained by what players *would* do if the equilibrium were disturbed. A potential aggressor does not attack because the defender would respond. A quality challenger does not enter because the incumbent would campaign aggressively. A target complies because the sender would follow through on the threat. The causal force in each explanation is located in unobserved counterfactual behavior — the defender’s response to an attack that did not happen, the incumbent’s campaign against a challenger who stayed out, the sender’s action upon defiance that was never tested.

This structure pervades the study of international conflict. The theoretical literature on deterrence (Schelling 1960; Fearon 1995; Powell 1999) explains war and peace through the strategic calculations of states anticipating each other’s responses. The empirical literature (Huth 1988; Signorino and Tarar 2006) tests whether those calculations obtain in the historical record, using statistical models that embed the game-theoretic structure directly (Signorino 1999, 2002). Both literatures make causal claims: nuclear deterrence kept the Cold War cold; extended commitments prevent or invite conventional conflict; audience costs make threats credible. The causal content of these claims rests on the game-theoretic logic, but the foundations of that causal content have not been examined carefully. Fearon (1991) established that game-theoretic models are unusual in political science precisely because they make the counterfactual structure explicit — off-equilibrium-path behavior is the counterfactual, and the model specifies it. But Fearon had no formal apparatus for evaluating whether a given specification is epistemically adequate. What kind of causal inference does a game-theoretic model license? Under what conditions is it legitimate? The literature has proceeded largely without answers.

The same structure appears in the study of electoral competition. The strategic politicians hypothesis (Jacobson and Kernell 1983) holds that quality challengers enter when conditions favor them and stay out when incumbents deter them through resource accumulation. Whether a given incumbent’s war chest *caused* a quality challenger to stay out is structurally identical to whether a given deterrent *caused* an adversary not to attack. In

both cases the explaining event is an equilibrium non-event; in both cases the causal mechanism runs through an off-equilibrium-path response that is never observed; in both cases the analyst faces the same epistemological problem. The entry deterrence game is not merely an analogy across these two literatures. It is the same formal object, applied to different substantive domains.

The causal inference problem embedded in these questions has a precise structure. The credibility revolution in political science — built on randomized experiments, regression discontinuities, instrumental variables, and difference-in-differences designs — has produced rigorous tools for identifying causal effects. For a long time these tools required accepting the stable unit treatment value assumption (SUTVA; Rubin 1974): a unit’s outcome depends only on its own treatment, not on others’. Recent work has partially relaxed this constraint, developing potential-outcomes estimators for settings with interference and spillovers (Sobel 2006; Hudgens and Halloran 2008). But even these extensions treat cross-unit dependence as a statistical nuisance — a complication for identification that careful design or weighting can bound or correct. They do not model the mechanism that generates the dependence. In a game, by contrast, strategic interaction is not a nuisance but the causal mechanism itself. The defender’s response to an attack that never happened, the incumbent’s campaign against a challenger who stayed out: these are not sources of bias to be eliminated but the content of the causal explanation. Recovering them requires engaging with the strategic structure, not conditioning it away. The interference literature’s tools do not do this, and the rich literature on causal inference using game-theoretic models has grown up largely outside their framework (Fearon 1991; Pearl 2009).

A further complication runs deeper. Even granting a well-specified game, it is not obvious what a counterfactual means within it. Lewis (1973)’s possible-worlds analysis requires, for any counterfactual claim, a set of possible worlds and a similarity metric on that set: the counterfactual “had the challenger entered” is evaluated at the closest world in which the challenger enters. Fearon (1991) identifies this minimal-rewrite requirement as the standard of quality for counterfactuals in political science; game-theoretic models are supposed to satisfy it by specifying off-path behavior explicitly. But in a complete-information game with a unique equilibrium, a rational player with those utilities simply does not enter. The Lewis set — the set of worlds in which the antecedent holds — is empty under strict rationality, rendering the

counterfactual vacuously uninformative. Making it non-empty requires relaxing something: the payoffs, the rationality assumption, or the game form itself. I show that different solution concepts are implicitly different specifications of this Lewis metric. Selten’s trembling-hand perfection populates the Lewis set with worlds in which players’ hands tremble; quantal response equilibrium (McKelvey and Palfrey 1995) parametrizes it by a rationality coefficient; Harsanyi’s purification theorem populates it with payoff perturbations. The choice of solution concept is not merely a technical device for equilibrium selection but a substantive epistemological commitment about which counterfactual worlds are close — and therefore about what is being claimed when a game-theoretic causal argument is made.

The two literatures that bear most directly on this problem have developed without formal contact. Fearon (1991) established that game-theoretic models are distinguished by making off-equilibrium behavior explicit, and Signorino’s structural estimation program showed how to use that structure empirically (Signorino 1999, 2002). But neither framework asks the epistemological question: *which* causal claims does a given game-theoretic specification license, and under *what conditions*? On the causal inference side, Dawid (2000)’s decision-theoretic framework provides exactly this kind of accounting — distinguishing when causal inference relies only on what the data can supply from when it requires metaphysical assumptions, and classifying which queries are tractable from the available structure — but it has not been applied to multi-agent strategic settings. The connection is latent in both bodies of work: game-theoretic expected-utility calculations and Dawid’s regime-based inference share the same decision-theoretic foundations, and extensive form games already possess the structure Dawid’s augmented DAGs require. What has been missing is the formal statement.

This paper supplies that formal account, drawing on Dawid (2000, 2021) rather than Pearl or Rubin. Dawid’s approach is decision-analytic rather than counterfactual: causal inference is the comparison of outcome distributions under alternative *treatment regimes* — decision rules assigning treatment as a function of observed covariates — rather than the recovery of potential outcomes. This framework shares its decision-theoretic foundations with game theory in a way that the potential outcomes framework does not. A game’s expected utility calculations are the same kind of object as a statistician’s expected loss; a strategy is a decision rule; a payoff function is a mechanism. The connection is not analogical but structural.

I make it precise. I prove that finite extensive form games of perfect re-

call are augmented directed acyclic graphs in Dawid’s sense, with strategies corresponding to treatment regimes and the correspondence arising naturally from the game’s information structure rather than requiring additional construction. I show that equilibrium concepts — under conditions I state as a proposition — constitute *determining concomitants*: the mechanistic variables that ground causal attribution by specifying what would happen under any treatment realization, including those never observed. And I use Dawid (2000)’s distinction between *sheep* (causal claims grounded in the physical model) and *goats* (claims requiring untestable metaphysical commitments) to classify the game-theoretic causal claims that political scientists routinely make.

The paper proceeds as follows. Section 2 develops Dawid’s framework and introduces the entry deterrence game as the running example. Section 3 proves the isomorphism and states the main proposition on equilibrium as determining concomitant. Section 4 characterizes sheep and goats in strategic settings and derives the deterrence corollary. Section 5 applies the framework to candidate entry, connecting Signorino-style structural estimation to the Lee–Caughey–Sekhon regression discontinuity literature and demonstrating both a sheep and a goat. Section 6 discusses implications for the relationship between structural modeling and the credibility revolution.

2 Dawid’s Framework

The modern approach to causal inference in political science inherits a specific vocabulary: potential outcomes $Y_i(1)$ and $Y_i(0)$, the stable unit treatment value assumption, average treatment effects. This vocabulary, developed by Rubin (1974) and extended by many others, has proved enormously productive. But it rests on a metaphysical commitment that is easy to overlook: the assumption that each unit carries, as a pre-existing attribute, a pair of potential outcomes $(Y_i(1), Y_i(0))$ whose joint distribution is a meaningful object of inference. Dawid (2000) calls this *fatalism* — the treatment of counterfactual outcomes as properties of units rather than as features of decision problems — and argues that it introduces a source of arbitrariness into causal inference that is never resolved by data.¹

1. Rubin’s respondents in the same discussion contest the “metaphysical” characterization, treating potential outcomes as censored observables rather than metaphysical stipulations — attributes of units that exist whether observed or not (see Dawid 2000,

This section presents Dawid’s alternative framework, which grounds causal inference in the decision-analytic tradition rather than in counterfactual metaphysics. I introduce the framework’s central concepts and close by translating a simple entry game into the framework’s language, establishing the probability model that will carry through the rest of the paper.

2.1 Treatment Regimes and the Physical Model

The basic unit of analysis in Dawid’s framework is not the potential outcome but the *treatment regime*: a decision rule g specifying how treatment T is assigned, possibly as a function of observed covariates K available at the time of decision. Formally, a regime is a (possibly stochastic) function $g : \mathcal{K} \rightarrow \Delta(\mathcal{T})$, mapping covariate histories to distributions over treatments.

Under regime g , treatment is assigned according to g and the outcome Y follows its conditional distribution given the assigned treatment. The resulting observable distribution $P_g(Y)$ — the distribution of Y we would observe if we implemented g — is the fundamental empirical object. Crucially, $P_g(Y)$ is identifiable from data: if we can implement g (through an experiment, a natural experiment, or a credible structural model), we can in principle learn P_g .

An *effects-of-causes* (EoC) query is any comparison of the distributions P_{g_1} and P_{g_0} for two regimes g_1 and g_0 . This is Dawid’s forward-looking causal question: if we were to implement g_1 rather than g_0 , how would the distribution of outcomes change? The relevant object is the full predictive distribution under each regime, from which any summary — the difference in means, a quantile comparison, the probability that Y exceeds some threshold, the variance of outcomes — can be derived. Writing the EoC quantity as a difference of expectations,

$$\mathbb{E}_{g_1}[Y] - \mathbb{E}_{g_0}[Y], \tag{1}$$

is one choice among many, and often not the most informative one. In the entry game, for instance, a risk-averse challenger cares not just about expected payoff but about the probability of the ruinous outcome $X = -c$ —

Discussion). The present paper’s argument does not turn on the existence question. What matters is the non-identification claim: the joint distribution $P(Y_t, Y_c)$ is never identified from the physical model alone, regardless of whether it “exists.” Sheep and goats are an epistemic distinction, not a metaphysical one.

a feature of P_{g_1} that the mean alone does not capture. No joint distribution over potential outcomes is required for any of these queries; only the two marginal predictive distributions P_{g_1} and P_{g_0} are needed.

The *physical model* is the full collection $\{P_g : g \in \mathcal{G}\}$ of observable regime-outcome distributions. This is everything about the causal structure that is, even in principle, empirically accessible.

The physical model is identified from observational data through an *extended conditional independence* (ECI) condition (Dawid 1979, 2021): the distribution of Y under regime g , given the observable covariate history K , equals the distribution we would observe in a subpopulation in which g was actually implemented. In Dawid’s (1979) notation, $(Y \perp\!\!\!\perp G \mid K)_{\text{ext}}$. This plays the same role that ignorability — “no unmeasured confounders” — plays in the potential-outcomes literature: it licenses the identification of $P_g(Y \mid K)$ from observational data. In game-theoretic settings, each player’s information set is the formal expression of K : it specifies exactly what the player observes before acting. ECI is therefore not an additional assumption layered onto the game; it is instantiated by the game’s information structure. Section 3 makes this correspondence precise through the isomorphism between information sets and observable parent sets in Dawid’s augmented directed acyclic graph.

2.2 Sheep and Goats

Not all causal inferences are created equal. Some depend only on the physical model and can in principle be identified from data; others require a *metaphysical model* — specifically, a joint distribution $P(Y_t, Y_c)$ over potential outcomes under treatment and control simultaneously — that is never identified by any experiment or observational study. Dawid calls the former *sheep* and the latter *goats*.

The sheep/goat distinction turns on a principle of empirical testability: an inference is a sheep if and only if it depends on the data only through quantities that are identifiable from the physical model. The average treatment effect $\mathbb{E}[Y_t - Y_c]$, for instance, is a sheep — it equals $\mathbb{E}_{g_1}[Y] - \mathbb{E}_{g_0}[Y]$ under standard assumptions, which is identifiable. But the *probability of causation* — the probability that treatment caused the observed outcome for a specific unit — is a goat. It requires the joint distribution $P(Y_t, Y_c)$, which

depends on the correlation

$$\rho \equiv \text{cor}(Y_t, Y_c) \tag{2}$$

that is never identified from the physical model alone. Different values of ρ are consistent with the same physical model but yield different answers to the attribution question. Assumptions like the Trivial Uniform Assumption (TUA: $Y_t \perp Y_c$) or monotonicity ($Y_t \geq Y_c$ always) resolve this indeterminacy, but they are untestable metaphysical commitments, not empirical findings.

The *causes-of-effects* (CoE) query — “did treatment cause this outcome?” — is the paradigmatic goat. It is backward-looking attribution: given that unit i received treatment $T_i = t$ and exhibited outcome $Y_i = y$, what is the probability that Y_i would have been different under the counterfactual treatment? This requires $P(Y_c \mid Y_t = y, T = t)$, which is not identified without strong assumptions on the joint distribution.

2.3 Determining Concomitants

Dawid’s framework is not simply a prohibition on CoE analysis. It identifies the conditions under which CoE analysis is legitimate: the existence of a *determining concomitant*.

Definition 1 (Dawid 2000, 2021). *A pre-treatment variable D is a determining concomitant for the causal effect of T on Y if there exists a function f such that $Y = f(T, D)$ holds with probability one.*

When a determining concomitant exists and is observed, both potential outcomes are determined: $Y_t = f(t, D)$ and $Y_c = f(c, D)$. The joint distribution $P(Y_t, Y_c \mid D)$ collapses to a point mass, and the correlation ρ ceases to matter. CoE inference becomes tractable because the counterfactual is no longer a metaphysical stipulation but a physical fact about the mechanism: we know what f is, and D is observed.

The practical requirement is demanding. It asks for knowledge of the mechanism — the function f that maps treatment and background conditions to outcomes. In most observational settings, this function is unknown and cannot be recovered from data alone. But in certain scientific contexts, where a theoretical model specifies the mechanism with sufficient precision, a determining concomitant may be available.

Game-theoretic models are unusual in political science precisely because they sometimes provide this structure explicitly. A strategy profile specifies,

for every player at every information set — including nodes never reached in equilibrium — what each player would do. This is exactly the form of a determining concomitant: a pre-treatment specification of the mechanism that, if correct, pins down both potential outcomes for any unit. The question is under what conditions the strategy profile actually *functions* as a determining concomitant rather than merely having its form. Section 3 makes this precise.

2.4 A Probability Model for the Entry Game

The running example that carries through the paper is a simple two-player entry deterrence game. A Challenger (C) decides whether to enter an electoral contest; an Incumbent (I) decides, conditional on entry, whether to fight or accommodate. The game has the following random variables:

$A \in \{Enter, Stay\ Out\}$	Challenger's action
$R \in \{Fight, Accommodate, \emptyset\}$	Incumbent's response
$X \in \mathbb{R}$	Challenger's payoff

where $R = \emptyset$ when $A = Stay\ Out$ (the incumbent has no move), and

$$X = \begin{cases} -c & \text{if } A = Enter, R = Fight \\ \pi_C & \text{if } A = Enter, R = Accommodate \\ 0 & \text{if } A = Stay\ Out \end{cases} \quad (3)$$

with $\pi_C > 0 > -c$. The incumbent's payoffs are $-d < \pi_I < M$, where M is the benefit from an uncontested seat, π_I the benefit from a contested race under accommodation, and $-d$ the cost of actively fighting entry.

The observable distribution is $P(A, R, X)$, which since X is a deterministic function of (A, R) reduces to $P(A, R)$. The two treatment regimes of interest are g_1 (always enter) and g_0 (always stay out). The EoC query — should E enter? — compares

$$\mathbb{E}_{g_1}[X] = \pi_C \cdot P(Acc | Enter) - c \cdot P(Fight | Enter) \quad \text{vs.} \quad \mathbb{E}_{g_0}[X] = 0. \quad (4)$$

This is identified from $P(A, R)$: it requires only the conditional distribution of the incumbent's response given entry, which is in principle observable. It is a sheep.

The CoE query is different. Suppose C stayed out. Was it because I would have fought? This requires $P(X < 0 \mid A = \textit{Stay Out})$ — a counterfactual distribution over what X would have been had E entered. This is not identified from $P(A, R)$: when E stays out, the incumbent’s response is \emptyset , and we learn nothing about what I would have done. The determining concomitant needed to resolve this ambiguity is I ’s *strategy* — a complete specification of I ’s intended behavior at every information set, including the off-path set that is never reached when E stays out.² Whether such a concomitant is available is precisely the question to which game-theoretic structure speaks. I turn to that question next.

3 Games as Causal Models

The probability model introduced in Section 2 has a structural feature that deserves emphasis: the variables A , R , and X are not symmetrically related. A is a choice; R is a response to that choice; X is determined by both. The causal structure runs in one direction, and it is encoded in the order of the game. This is not incidental. Extensive form game trees are devices for representing exactly the kind of sequential, conditional structure that Dawid’s augmented directed acyclic graphs are designed to capture. In this section I make the correspondence precise, show what it implies about strategies as treatment regimes, and state the conditions under which game-theoretic structure licenses the causes-of-effects analysis that Section 2 identifies as the hard problem.

2. Dawid (2021) distinguishes the *intended* treatment (the regime g , specifying what the decision rule assigns) from the *applied* treatment (what actually occurs, which may deviate through non-compliance or stochastic implementation). The intention-to-treat quantity compares P_{g_1} to P_{g_0} without requiring the applied treatment to equal the intended regime. Game-theoretic models map naturally onto this distinction: a player’s strategy is the intended regime; the realized action at any node is the applied treatment. Trembling-hand and quantal-response equilibria incorporate this gap explicitly, treating the realized action as a stochastic perturbation of the intended strategy. Section 3 formalizes this: the equilibrium concept selects among intended regimes, and the EoC query compares them. CoE analysis requires knowing not only which regime was intended but whether the specific observed path is consistent with it.

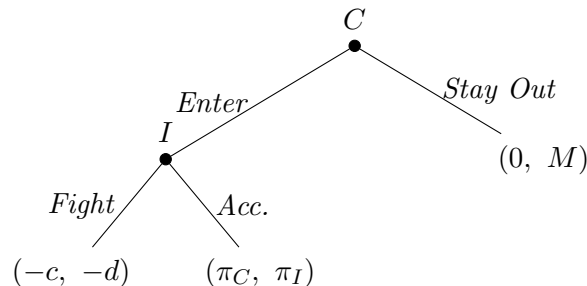


Figure 1: The entry deterrence game in extensive form. Payoffs are listed as (Challenger, Incumbent). The subgame perfect equilibrium has the Challenger enter and the Incumbent accommodate, since fighting is dominated ($\pi_I > -d$). But a second Nash equilibrium exists in which the Incumbent commits to Fight and the Challenger stays out, leaving the Incumbent’s decision node off the equilibrium path and permanently unobserved. It is in this deterrence equilibrium that the Incumbent’s strategy at the Fight/Accommodate node — the would-be response to entry — is the determining concomitant that licenses causes-of-effects inference. Section 3 shows that this node corresponds to an intervention node in Dawid’s augmented directed acyclic graph.

3.1 The Structural Isomorphism

A finite extensive form game Γ is a rooted tree (T, P) — a finite set of nodes T with a distinguished root and predecessor function $P : T \setminus \{r\} \rightarrow T$ — in which each non-terminal node is assigned a *decider*: either a member of the player set N or Nature. For each player $i \in N$, the set of i ’s decision nodes is partitioned into *information sets* \mathcal{H}_i : player i observes only which information set they are in, not the specific node within it. Edges out of each decision node are the available actions. Terminal nodes $Z \subset T$ carry payoff vectors $u = (u_i)_{i \in N}$. Nature’s moves carry a probability distribution ρ .

Dawid’s augmented DAG specifies a set of variables, directed acyclic edges encoding causal dependencies, a partition of variables into chance nodes, decision nodes, and outcome nodes, and a probability model over the graph. The correspondence between these two objects, summarized in Table 1, is canonical.

Proposition 1 (Game–DAG Isomorphism). *Let Γ be a finite extensive form game of perfect recall. There exists a canonical augmented DAG $\mathcal{D}(\Gamma)$ in*

Table 1: Structural isomorphism between extensive form games and Dawid’s augmented directed acyclic graphs.

Extensive form game	Dawid’s augmented DAG
Nature’s move at node t	Exogenous chance node ω_t
Decision node for player i at h	Intervention node $D_{i,h}$
Information set $h \in \mathcal{H}_i$	Observable parent set $\text{Pa}^{\text{obs}}(D_{i,h})$
Behavioral strategy $b_i : \mathcal{H}_i \rightarrow \Delta(A(h))$	Treatment regime $g_i : \mathcal{K}_i \rightarrow \Delta(\mathcal{T})$
Strategy profile (b_i, b_{-i})	Joint treatment regime $g = (g_i, g_{-i})$
Terminal node $z \in Z$	Outcome node Y_i
Payoff function $u_i : Z \rightarrow \mathbb{R}$	Loss function $\ell_i : \mathcal{Y}_i \rightarrow \mathbb{R}$

Dawid’s sense such that:

- (a) every information set $h \in \mathcal{H}_i$ corresponds to the observable parent set of the decision node $D_{i,h}$ in $\mathcal{D}(\Gamma)$;
- (b) every behavioral strategy b_i for player i corresponds to a treatment regime g_i for the same player in $\mathcal{D}(\Gamma)$;
- (c) for any strategy profile b , the distribution P_b over terminal payoffs in Γ equals the regime distribution P_g over outcomes in $\mathcal{D}(\Gamma)$.

The proof is given in the Appendix. The construction is injective and invertible on its image: given an augmented DAG produced by this construction, the game Γ can be recovered uniquely. The term *isomorphism* is therefore used in the sense of a structure-preserving embedding of games into augmented DAGs, not a claim that every augmented DAG arises from a game — general DAGs admit nodes with multiple parents and need not have tree structure.

The construction is closely related to the multi-agent influence diagram (MAID) framework of Koller and Milch (2003), who establish the representational equivalence between game trees and influence diagrams (their Proposition 4.1) and use it to derive strategic relevance from the graph structure via s -reachability and to decompose games for efficient equilibrium computation. The present construction builds on the same game-to-graph correspondence but puts it to a different use: not solving the game but evaluating the epistemological status of causal claims made within it. A feature of the Dawidian

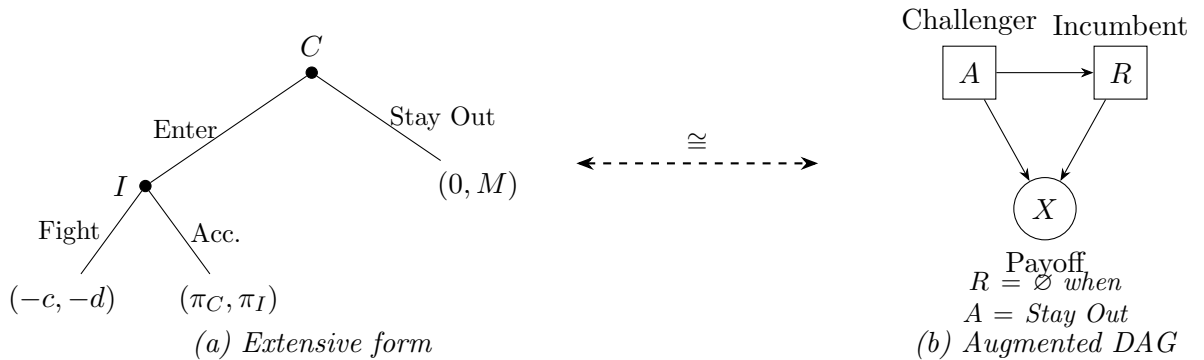


Figure 2: The entry deterrence game as an extensive form tree (a) and as Dawid's augmented directed acyclic graph (b). Squares denote intervention (decision) nodes; circles denote chance nodes. The causal structure is identical: the Challenger's action A causes the Incumbent's response R (by determining whether I moves at all), and both cause the payoff X . The information set of each player corresponds to the set of parent nodes in the DAG that are observable at the time of decision.

construction that makes this application possible is its canonicity: the information sets of the game tree directly and completely determine the observable parent sets of the DAG, so that no additional modeling choices are required beyond the game form itself. Perfect recall is the operative condition: it guarantees that information sets are well-nested, so that the observable parent set is consistently defined at each decision node and the DAG is uniquely determined by Γ .

Figure 2 makes the correspondence visual for the entry deterrence game. The game tree on the left is Figure 1; the augmented DAG on the right represents the same causal structure in Dawid's notation, with squares denoting decision nodes and circles denoting chance and outcome nodes. The directed edges encode causal dependence: A causes R (the challenger's action determines whether the incumbent moves at all) and both A and R cause X . There are no cycles. The DAG is the game tree with its sequential structure projected into a directed graph, with no information discarded and no structure added.

3.2 Strategies as Treatment Regimes

The correspondence in Table 1 is not merely notational. The row mapping strategies to treatment regimes carries substantive content.

A *strategy* s_i in an extensive form game is a complete contingency plan: a function from every information set $h \in \mathcal{H}_i$ to a distribution over i 's available actions at h . The emphasis on *complete* is essential. Player i 's strategy specifies behavior not only at information sets that are reached in equilibrium but at every information set in the game tree, including those that are off the equilibrium path — nodes that, given the strategies of the other players, will never actually be visited. A strategy is not a description of what a player *does*; it is a description of what a player *would do* in every possible circumstance.

This is exactly what Dawid requires of a treatment regime. A regime g specifies how treatment would be assigned under every possible covariate history $k \in \mathcal{K}$, including covariate histories that are never realized under the actual data-generating process. The regime is a counterfactual object: it specifies the treatment that *would* be applied in hypothetical circumstances that may not obtain. Off-equilibrium-path strategies in a game and off-support treatment regimes in Dawid's framework are the same kind of object.

Signorino (2002) comes closest to making this explicit among the existing literature on strategic statistical modeling: “The ‘path play’ in strategic situations results in the observed sample. *Everything else off the path of play can be thought of as the counterfactuals that comprise the rest of the strategic model.*” What Signorino identifies as a statistical observation — that off-path behavior defines the counterfactual structure of strategic data — is, in Dawid's terms, the claim that a strategy profile constitutes the joint treatment regime that determines the physical model.

The practical consequence is that off-path behavior is not a nuisance to be managed through solution concept selection; it is the structural content of the causal model. A political scientist who estimates a strategic statistical model (in the manner of Signorino 1999; Signorino and Tarar 2006) is, implicitly, specifying a joint treatment regime and computing effects-of-causes quantities from it. The causal content of the model is in the strategies, not in the reduced-form coefficients.

3.3 Equilibrium as Determining Concomitant

The mapping from strategies to treatment regimes raises an immediate difficulty. In a single-player decision problem, the treatment regime fully determines the distribution of outcomes: the agent controls the treatment, the treatment causes the outcome, and there is no further dependence to account for. In a game, player i controls only their own strategy s_i , not the full profile $s = (s_i, s_{-i})$. The outcome u_i depends on what all players do. Player i 's strategy is a treatment regime, but it is not a *determining* concomitant: it does not determine the outcome by itself.

The solution concept fills this gap. An equilibrium concept C — Nash, subgame perfect, quantal response, trembling hand — specifies, for each strategy s_i that player i might adopt, a prediction about what the other players will do in response. It pins down s_{-i} as a function of s_i and the game's parameters. The solution concept is therefore not merely a device for selecting among multiple equilibria; it is the component of the causal model that converts i 's strategy into a determining concomitant by closing the loop on the other players' counterfactual behavior.

Proposition 2. *A strategy profile s^* selected by solution concept C constitutes a determining concomitant for player i 's outcomes if: (a) the game form is correctly specified; (b) C uniquely selects s^* ; and (c) payoffs u_i are known. Under these conditions, $u_i = f(s_i, s_{-i}^*, \theta)$ is a physical relationship in Dawid's sense: the counterfactual outcome under any deviation by i is determined by s_{-i}^* and the payoff function, not by an untestable assumption about the correlation $\rho(u_i^t, u_i^c)$ between potential outcomes.*

Condition (a) requires that the game tree correctly represents the strategic situation: the players, their action sets, the order of moves, the information structure, and the payoff-relevant state of the world are all correctly specified. This is a strong requirement, but it is the same requirement that any structural model places on its researcher. Condition (c) — that payoffs are known — is similarly strong; in practice, payoffs are typically estimated alongside other parameters, introducing uncertainty that the Proposition brackets. Condition (b) is the deepest requirement, and I examine it in the following subsection.

3.4 Solution Concepts as Lewis Metrics

Proposition 2 requires that solution concept C uniquely select a strategy profile s^* . This condition deserves closer examination, because it is doing more work than equilibrium selection. It is specifying the *Lewis metric* for the game-theoretic counterfactual — the answer to Fearon (1991)’s demand that counterfactuals in political science specify the closest possible world in which the antecedent holds.

Lewis (1973)’s possible-worlds analysis of counterfactuals requires, for any conditional “had X obtained, Y would have followed,” a set of possible worlds and a similarity metric: the counterfactual is evaluated at the closest world in which the antecedent holds. In a complete-information game with a unique Nash equilibrium, however, the Lewis set for an off-equilibrium antecedent — the set of worlds in which a dominated action is taken — is empty under strict rationality. A rational player with those utilities and beliefs simply cannot play the dominated action. The counterfactual has no nearest world in which to be evaluated; it is vacuously uninformative.

Making the Lewis set non-empty requires relaxing something. Different solution concepts correspond to different ways of doing so, and hence to different similarity metrics:

Trembling hand (Selten 1975). The closest world in which E enters is one in which E ’s hand trembled with probability $\varepsilon > 0$. The Lewis set is the family of ε -perturbed games, and the metric is ε itself. Off-equilibrium behavior is defined as I ’s best response given the possibility of trembles, in the limit $\varepsilon \rightarrow 0$. Selten’s refinement is, in this light, a specification of Lewis’s closest-world semantics for game-theoretic counterfactuals — a connection that, to our knowledge, the formal literatures have not drawn explicitly.

Quantal response equilibrium (McKelvey and Palfrey 1995). The closest world in which E enters is a world with rationality parameter $\lambda < \infty$. The Lewis set is the one-parameter family of QRE distributions, and the metric is $|\lambda - \infty|$. For any finite λ , every action is played with positive probability, every information set is reached, and off-path behavior is directly observable in repeated play. The variance term in a QRE is not merely a model of bounded rationality; it is a parametric Lewis metric that determines how far we must travel from the actual world to reach one in which the antecedent holds. This is the

solution concept underlying Signorino’s (1999) strategic probit, and his motivation for it — that SPE assigns zero probability to off-equilibrium outcomes, making the likelihood degenerate — is a statistical restatement of the empty-Lewis-set problem.

Utility perturbation (Harsanyi 1973). The closest world in which E enters is one in which E ’s payoffs are perturbed by the minimum amount necessary to make entry weakly optimal. The Lewis set is a neighborhood in utility space, and the metric is Euclidean distance between utility vectors. Harsanyi’s purification theorem — that any mixed-strategy Nash equilibrium is the limit of pure-strategy equilibria under small utility perturbations — can be read as a theorem about the structure of Lewis’s nearest worlds for mixed-strategy counterfactuals.

The three solution concepts above vary along a single dimension: how rational the players are. But rationality is not the only thing an analyst can relax. There are two other sources of Lewis distance in game-theoretic counterfactuals, and they are worth naming explicitly because they correspond exactly to the other two conditions of Proposition 2.

The first is *payoff distance*. The closest world in which E enters need not be one in which E is less rational; it could be one in which E ’s payoffs are slightly different — entry is more attractive, or the cost of fighting is higher for I . Harsanyi’s purification theorem already deploys this metric in its specific application to mixed-strategy equilibria, but the underlying idea is more general. Any analyst who says “perhaps the challenger expected a better outcome from entry than we estimated” is implicitly using a payoff metric: the closest world is the one that minimizes the distance in payoff space between the estimated game and the game in which entry is rational. This is exactly what it means for condition (c) of Proposition 2 — that payoffs are known — to fail. An analyst uncertain about payoffs is uncertain about which payoff world she is in; the Lewis metric is the structure of that uncertainty.

The second is *structural distance*: edit distance on the game tree itself. The closest world in which E enters might be one in which the game has a different information structure, a different action set, or a different order of moves. Perhaps the incumbent’s commitment is not credible in the way the tree assumes; perhaps there is a signaling stage that the model omits; perhaps the challenger can observe something the model says she cannot.

Each of these is a minimal rewrite of the game form rather than of the players' rationality or payoffs. The structural metric asks: how many edits to the game tree are required to reach a world where the antecedent holds? This is the content of condition (a) — correct game specification — expressed as a Lewis distance. When the game form is misspecified, the closest possible world in which the counterfactual antecedent holds may not be the world the analyst thinks she is evaluating.

The full taxonomy of Lewis metrics in game-theoretic counterfactuals therefore has three dimensions, each corresponding to one of the proposition's conditions:

Relaxed dimension	di-	Metric	Condition
Player rationality	rational-	ε -trembles, λ -rationality, payoff noise	(b) unique selection
Payoff values		Distance in utility space	(c) known payoffs
Game form		Edit distance on the tree	(a) correct specification

This taxonomy clarifies what it means to make a game-theoretic counterfactual claim. Every such claim implicitly holds two of these dimensions fixed while varying the third. A trembling-hand argument holds the game form and the payoffs fixed, and relaxes rationality. A sensitivity analysis over payoff parameters holds the game form and the solution concept fixed, and relaxes payoff certainty. An argument that the game was misspecified — that the relevant world is one with a different information structure — holds rationality and payoffs fixed, and relaxes the structure. In each case, the analyst is making a Lewis metric choice, and that choice determines the content of the counterfactual.

The practical implication is that condition (b) of Proposition 2 — that C uniquely selects s^* — is a commitment to a specific metric along one of these three dimensions. Two analysts who agree on the game form and payoffs but disagree on the solution concept are not merely making different technical choices about equilibrium selection; they are disagreeing about which worlds are close and will therefore arrive at different counterfactual densities even from identical observed data. The determining concomitant that licenses CoE inference is well-defined only relative to a choice of metric along all three dimensions.

This is a genuine limitation, but it is also a clarification. The arbitrariness is not new — it was always present in game-theoretic causal claims. What the framework makes explicit is that this arbitrariness has a precise location (the Lewis metric, equivalently the conditions of the proposition) and a precise consequence (different off-path strategies, different determining concomitants, different CoE answers). A political scientist who wants to claim that deterrence caused non-escalation must commit, at minimum, to a solution concept, a set of estimated payoffs, and a correctly specified game form — and each of those commitments should be stated and defended, not left implicit.

3.5 What the Proposition Relocates

The gain from Proposition 2 is real but carefully bounded. In Dawid’s framework without game structure, causes-of-effects inference requires an untestable assumption about $\rho(Y_t, Y_c)$ — a number that is never identified from any data, regardless of sample size or experimental design. The proposition replaces this with conditions (a)–(c) plus a choice of Lewis metric. These are stronger assumptions in one sense — they require correct game specification, unique equilibrium selection, and known payoffs — but they are different in kind. They are *substantive claims about the strategic environment* that are, at least in principle, subject to empirical scrutiny. A game form can be misspecified in ways that leave observable traces. A solution concept can be evaluated against data. Payoffs can be estimated. None of this is possible for ρ , which is metaphysical all the way down.

This relocation of arbitrariness — from an unidentified correlation to a set of testable structural assumptions — is precisely what Signorino’s program of structural statistical modeling delivers in practice, without having had a causal inference framework to explain why. The strategic probit model (Signorino 1999; Signorino and Yilmaz 2003) is an effects-of-causes engine: it specifies the joint treatment regime (the strategy profile under QRE), uses it as the determining concomitant, and computes $P_g(Y)$ for any hypothetical intervention on player i ’s strategy. The causal content is in the structure of the model, not in an assertion about ρ .

The Signorino program is highlighted here because it makes the connection explicit: the motivation for QRE is precisely the need to assign positive probability to off-equilibrium paths, and Signorino’s commentary identifies off-path behavior as the counterfactual structure of the data. But the frame-

work is not specific to the QRE approach. Any researcher who specifies a game form, applies a solution concept, and estimates payoff parameters from data is implicitly specifying a joint treatment regime and computing EoC quantities from it — whether the application is crisis bargaining, legislative coalition formation, or electoral entry. Sartori (2003) demonstrates this in a different register: her estimator for binary-outcome selection models without exclusion restrictions achieves identification through the game structure itself — the same variables drive both strategic entry and subsequent outcomes, and the game form supplies the identifying content that a standard Heckman approach would import from outside. The Dawidian framework makes explicit the causal content already present in this broader practice of structural game-theoretic estimation.

The relationship to existing causal games literature is worth making precise. Koller and Milch (2003) show that game trees and influence diagrams are representationally equivalent and exploit this computationally: their MAID framework derives strategic relevance from the graph via s -reachability and decomposes games for efficient equilibrium computation. Hammond et al. (2023) build on Koller and Milch’s representational foundation to bring Pearl’s full causal hierarchy into games. Their “mechanised MAIDs” lift influence diagrams to all three levels of Pearl’s ladder — predictions, interventions (both hard and soft), and counterfactuals — computing causal queries via structural causal models. The present paper asks a different question from either: not how to represent or solve games graphically (Koller and Milch), nor how to compute causal quantities within them (Hammond et al.), but which causal claims a game-theoretic model licenses given what is empirically identifiable — the sheep/goats distinction. This epistemological question is orthogonal to Pearl’s causal hierarchy; it cross-cuts all three levels.

The complementary gap is sharpest at the counterfactual level. Hammond et al. compute counterfactuals via abduction-action-prediction within a structural causal model, which requires specifying the full functional form and noise distributions. Dawid’s framework asks whether the resulting counterfactual is identified from the physical model or depends on untestable structural assumptions. Both questions are legitimate; they are asked from different sides of the same problem. The game structure that Hammond et al. use to compute a counterfactual may or may not provide a determining concomitant in Dawid’s sense — and whether it does is precisely what distinguishes a sheep from a goat.

A sharper point concerns what the two frameworks treat as the primitive intervention. Pearl (2022) charges that Dawid’s regime indicator is simply $do(X)$ in different notation — that the structural commitments are the same, only concealed. The charge misidentifies the level of abstraction. Pearl’s $do(x)$ is an atomic intervention: it sets a single variable to a single value and severs its incoming edges. Dawid’s regime $g : \mathcal{K} \rightarrow \Delta(\mathcal{T})$ is a policy: a complete function specifying treatment assignment for every possible covariate history, including histories never realized in the data. In games, the natural causal question is never “what if variable X had been set to value x ?” but always “what if the player had played *strategy* g ?” — a comparison of policies, not of values. The regime is not a notational variant of do ; it is the operationally correct object for multi-agent sequential settings (see also Dawid 2022; Pearl 2022). Games were always Dawidian, but the do -calculus is not the right reduction.

Remark 1 (Two interventional primitives). *Two distinct interventional objects appear in the causal games literature and should not be conflated.*

Atomic intervention $do(x)$ (Pearl 2009). *Sets variable X to value x , severing incoming edges. Identifies interventional distributions $P(Y \mid do(x))$ via d -separation in the mutilated graph. Hammond et al. (2023) extend this to soft interventions on decision rules, allowing policies and counterfactuals to be computed within Pearl’s structural causal model framework. Asks: what is the effect of forcing a variable (or a decision rule) to a specific value?*

Treatment regime g (Dawid 2000, 2021). *A policy function $g : \mathcal{K} \rightarrow \Delta(\mathcal{T})$ specifying treatment for every covariate history, including off-support histories. Corresponds directly to a strategy. The equilibrium concept selects which regime obtains; deviating from it corresponds to switching to a different regime, not to a different value of a single variable.*

The regime is the correct primitive in strategic settings because: (a) equilibria are strategy profiles, not variable-value assignments; (b) off-equilibrium-path behavior is an off-support region of the regime, handled naturally by the policy function and awkwardly by do ; and (c) the equilibrium concept selects the regime that functions as the determining concomitant, a role that atomic do -interventions cannot play.

4 Sheep and Goats in Strategic Settings

The framework developed in Sections 2 and 3 yields two criteria — one permissive, one demanding — for evaluating causal claims in strategic settings. Fearon (1991) established that game-theoretic models are distinguished by making off-equilibrium-path behavior explicit, which is necessary for any counterfactual argument; the present framework formalizes the epistemological consequence of that observation. The permissive criterion covers effects-of-causes queries and is satisfied whenever the game provides a well-specified predictive model of opponents’ responses. The demanding criterion covers causes-of-effects queries and requires the full conditions of Proposition 2. In Dawid’s language: the former produces sheep; the latter, when its conditions fail, produces goats.

4.1 The Sheep Criterion

An effects-of-causes query in a game asks: what is the distribution of outcomes under a specified treatment regime for player i ? Concretely, for the entry game: what is $P_{g_1}(X)$ — the distribution of payoffs when the challenger always enters?

This query is tractable whenever the game provides a predictive model of opponents’ responses. The solution concept, together with the payoff specification, supplies a distribution over s_{-i} for any strategy s_i that player i might adopt. This is all the causal model needs: it converts player i ’s treatment regime into a regime over outcomes by integrating out opponents’ responses. No joint distribution over potential outcomes is required. Comparing P_{g_1} to P_{g_0} — or extracting any statistic from either distribution: a mean, a quantile, a tail probability, the probability of a specific terminal node — involves only the physical model.

The sheep criterion in strategic settings is therefore: *the EoC query is a sheep if the game and solution concept jointly identify the distribution of outcomes under the counterfactual regime*. This is weaker than the full conditions of Proposition 2: it does not require unique equilibrium selection or perfectly known payoffs, only that the predictive distribution $P_g(u_i)$ is well-defined under the regime in question. In a QRE model with estimated λ and estimated payoff parameters, the EoC quantity is a function of estimated parameters — uncertain, but identified. This is the mode of inference in Signorino’s strategic probit: given the estimated model, compute the predicted

probability of war under alternative strategy profiles. The causal interpretation is EoC throughout.

4.2 The Goat Criterion

A causes-of-effects query asks a different question: given that a specific outcome was observed, what caused it? In the entry game: given that the challenger stayed out, was it the incumbent’s threat that caused the non-entry? In the deterrence literature: given that war did not occur, was it the alliance, the nuclear capability, or the balance of forces that caused the peace? This is the *probability of causation* (PoC) — a quantity that Dawid and Musio (2016) show is unidentifiable from statistical data alone in precisely the contexts (pharmaceutical litigation, tort attribution) where it is most consequentially demanded. The game-theoretic setting is the same structure.

These queries require the joint distribution over potential outcomes — what would the outcome have been under the counterfactual treatment? — and this in turn requires the determining concomitant of Proposition 2. The concomitant must fill in the missing potential outcome: not the distribution of outcomes across a population of cases under a hypothetical regime, but the specific counterfactual outcome for *this* case. The conditions of the proposition must all hold: the game must be correctly specified, the solution concept must uniquely select the strategy profile, and the payoffs must be known. When any of these fails, the CoE query cannot be answered from the data, regardless of how much data is available.

Three specific failure modes are common in political science applications:

One-shot games. When the game is played once, off-equilibrium nodes are never visited and the opponents’ strategies at those nodes are never revealed. The determining concomitant exists in principle — the opponent has a strategy — but it is unobservable from a single realization. The CoE query requires knowing what the opponent *would have done*, and this is unidentified from a single outcome.

Incomplete information. When players have private information (types), the determining concomitant includes the opponent’s type, which is not directly observable. This failure mode is a scope condition rather than a universal complaint: it binds only when the proposed equilibrium is

pooling or semi-separating, so that the observed action does not reveal the type. In a fully separating equilibrium the action is fully revealing, the type is identified from the equilibrium path, and the complaint does not arise. The epistemological status of any CoE attribution in a Bayesian game therefore turns on whether the equilibrium separates — a question the signaling literature on deterrence already treats as central, here recast as a precondition for causal inference.

Non-visiting equilibria. Even in a repeated game with complete information, if the equilibrium never visits the relevant decision node, the concomitant at that node remains unobservable. This is the deterrence case: a successful deterrence equilibrium is one in which the potential aggressor never initiates, so the defender’s response to initiation is never tested.

The third failure mode has a particular character that warrants separate statement.

Corollary 1 (Deterrence Corollary). *Successful deterrence is epistemically self-undermining: when the equilibrium outcome is non-entry or non-escalation, the node at which the potential responder acts is never reached, so the determining concomitant — the responder’s actual strategy at that node — is never revealed by the observed outcome. The game provides the form of a determining concomitant but not its substance.*

The corollary says that the very success of deterrence destroys the evidence that would be needed to confirm that deterrence worked. A peaceful outcome is consistent with the deterrer having a credible threat that the potential aggressor correctly anticipated, and equally consistent with the potential aggressor having decided not to challenge for reasons entirely unrelated to the deterrent. The game structure — the strategy space, the payoff function, the solution concept — tells us what we *would* need to know: the challenger’s type and threshold, and the defender’s off-path strategy. But a single peaceful outcome provides none of this. The CoE query is a goat.

This is not merely a statistical power problem, resolvable with more data or better instruments. It is a structural identification problem: the equilibrium path never visits the node that contains the identifying information. More observations of peace do not help, because they are all generated by the same off-path node. Repeated play helps only if the equilibrium changes —

if the challenger sometimes enters and we observe the defender’s response — but a defender whose response is always tested is not, in the relevant sense, a deterrer.

5 Application: Candidate Entry

The framework’s practical content is best illustrated by an application that is familiar, strategically structured, and rich enough in data to make the identification questions concrete. The strategic politicians hypothesis (Jacobson and Kernell 1983) provides all three. I use it to demonstrate a sheep and a goat: one causal query that the game structure licenses, and one that it does not — not because the strategic reasoning is wrong, but because the structural conditions of Proposition 2 are not met.

5.1 The Strategic Entry Game

The strategic politicians hypothesis holds that quality challengers are strategic actors: they enter electoral contests when conditions favor them and stay out when they do not. The decision to challenge depends on the probability of winning, which depends in turn on the incumbent’s perceived vulnerability — their electoral margin, their war chest, their position-taking, the national political environment. Incumbents anticipate this calculation and invest in deterrence: they accumulate campaign funds, claim credit, avoid positions that invite challenge. The result is a strategic equilibrium in which many incumbents are never seriously challenged, not because they are universally popular, but because they have successfully altered the entry calculus.

This is the entry deterrence game of Section 2, now fully specified. Formally, let w denote the incumbent’s pre-electoral investment (war chest, position-taking), v the challenger’s quality (name recognition, fundraising capacity, electoral experience — private information held by the challenger), $\pi_C(v)$ the challenger’s expected payoff from entry as a function of their quality, and $c(w)$ the cost that the incumbent’s investment imposes on a challenger of any type. The challenger enters if and only if

$$\pi_C(v) > c(w). \tag{5}$$

The left side is increasing in v : higher-quality challengers expect larger returns from entry. The right side is increasing in w : heavier incumbent invest-

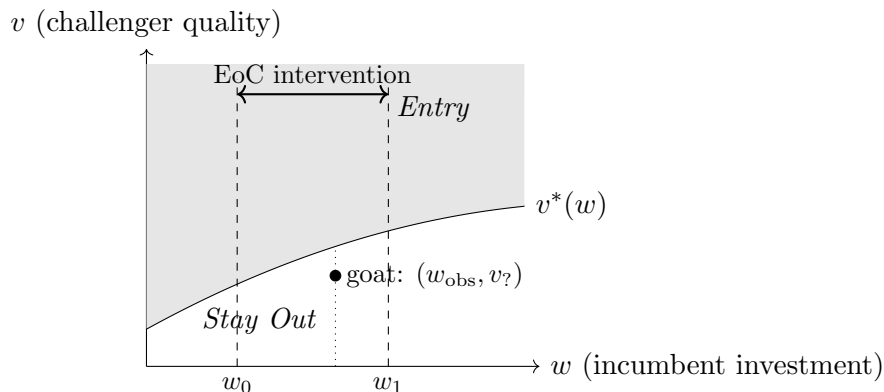


Figure 3: The cutpoint equilibrium in the strategic entry game. Challengers with quality v above the cutpoint $v^*(w)$ enter; those below stay out. An effects-of-causes (EoC) intervention shifts the incumbent’s investment from w_0 to w_1 , raising the cutpoint and shifting the distribution of entering challengers toward higher quality. The goat query asks whether a specific observed non-entry (the filled point) was caused by the incumbent’s war chest: the challenger’s type v is unobserved, so we cannot determine whether it fell below $v^*(w_{\text{obs}})$ due to the war chest or would have been below the threshold regardless.

ment raises the cost of entry. There is therefore a cutpoint $v^*(w)$, implicitly defined by $\pi_C(v^*) = c(w)$, such that challengers with $v > v^*(w)$ enter and challengers with $v \leq v^*(w)$ stay out. The cutpoint is increasing in w : a better-resourced incumbent deters a higher-quality challenger.

The incumbent’s strategy is an investment level w ; the challenger’s strategy is a threshold rule on their own quality. The game has a cutpoint equilibrium that is the electoral analog of the subgame perfect equilibrium in Figure 1. The strategy profile $(w^*, v^*(w^*))$ together with the payoff functions π_C and c is the determining concomitant — if it is correctly specified and uniquely selected.

5.2 A Sheep: Incumbency Advantage and the Distribution of Challengers

Consider the following EoC query: *what happens to the distribution of challenger quality when we intervene to increase incumbent electoral margin?*

This is a regime comparison. The treatment is the incumbent’s electoral environment; the outcome is the quality of the challenger who enters (or whether any challenger enters at all). The game specifies the mapping from treatment to outcome: a wider margin raises the incumbent’s capacity for pre-electoral investment, shifts the cutpoint $v^*(w)$ upward, and reduces both the probability of entry and the expected quality of entering challengers. The predictive distribution P_{g_1} (challenger quality) under a regime of high incumbent margin differs from P_{g_0} (challenger quality) under a regime of low margin in ways that the game makes precise. No joint distribution over potential outcomes is required; only the two marginal predictive distributions, identified from the game’s equilibrium, are needed. This is a sheep.

The regression discontinuity literature on incumbency advantage (Lee 2008; Caughey and Sekhon 2011) provides empirical identification of exactly this quantity. Close elections generate near-random variation in incumbency status: an incumbent who won by a few hundred votes is, in the relevant respects, similar to a challenger who lost by a few hundred votes but now holds the out-party position. The RD estimate — the discontinuous jump in subsequent electoral outcomes at the zero vote-share threshold — identifies the effect of incumbency status on future vote share, candidate quality, and entry decisions. This is, within the game-theoretic framework, an estimate of the difference $P_{g_1}(Y) - P_{g_0}(Y)$ for the two regimes defined by the electoral outcome: the distribution of challenger responses to the incumbency treatment.

Signorino’s strategic probit (Signorino 1999) identifies the same EoC quantity by a different route. The QRE model specifies $P_g(Y)$ directly from the game’s payoff structure and rationality parameter λ : the predicted probability of any terminal node is the product of LQRE action probabilities along the path. Varying the incumbent’s investment w shifts the regime g and changes the predicted distribution of entering challengers — exactly $P_{g_1}(Y) - P_{g_0}(Y)$ derived structurally rather than quasi-experimentally. The two methods estimate the *same sheep* with different identifying assumptions: the RD requires near-random variation near the electoral threshold; the structural model requires correct game specification and consistent λ estimation. They are complementary identification strategies for the same causal object, and the Dawid framework explains why.

The game-theoretic framework provides the structural interpretation that the reduced-form RD estimate lacks. The RD tells us that incumbency causes a shift in the electoral environment; the Jacobson-Kernell model tells us

through what mechanism: the incumbent invests in deterrence, the cutpoint rises, the challenger distribution shifts left. The two approaches are not in competition. The reduced-form estimate identifies the EoC quantity; the structural model gives it causal content.

Caughey and Sekhon (2011)’s finding that very close House elections show evidence of incumbent manipulation of the running variable is, in the framework’s terms, a violation of condition (a) of Proposition 2: the game form is misspecified if incumbents can precisely control their vote margins near the threshold. A correctly specified game must include the incumbent’s margin manipulation as a strategic action, not treat the margin as exogenous noise. The RD design’s identifying assumption is, equivalently, the assumption that the game does not include margin manipulation at the threshold. This is a testable claim about the game form — exactly the kind of structural accountability that Dawid’s framework demands.

5.3 A Simulation: Identification and Its Limits

The distinction between sheep and goat is not merely conceptual. A simple simulation makes it precise and directly demonstrates the paper’s central claim: the same data-generating structure that licenses the EoC query cannot license the CoE query, and more data does not close the gap.

Panel (a) of Figure 4 simulates 600 congressional districts with exogenous variation in incumbent war chest — the analog of a regression discontinuity design. Challenger entry is determined by whether the challenger’s type v exceeds the equilibrium cutpoint $v^*(w) = 0.2 + 0.1w$, with v drawn uniformly. The estimated probability of entry as a function of incumbent investment, together with its 95% confidence band, traces the EoC quantity directly: how does the investment level shift the entry probability? The band narrows with sample size. The EoC quantity is identified; residual uncertainty is finite-sample estimation error, nothing more. This is a sheep.

Panel (b) turns to a specific district where no challenger entered, with observed war chest $w_{\text{obs}} = 4$, implying cutpoint $v^*(w_{\text{obs}}) = 0.6$. The counterfactual asks what would have happened if the incumbent had invested less: $w_{\text{small}} = 1$, implying $v^*(w_{\text{small}}) = 0.3$. Two type distributions are shown. Both are consistent with the observed 40% entry rate: they agree on $F(v^*(w_{\text{obs}})) = 0.60$. They disagree on $F(v^*(w_{\text{small}}))$ — the mass of challenger types that would have entered under the counterfactual war chest. Under F_1 , most non-entrants have low latent quality, concentrated below $v^*(w_{\text{small}})$;

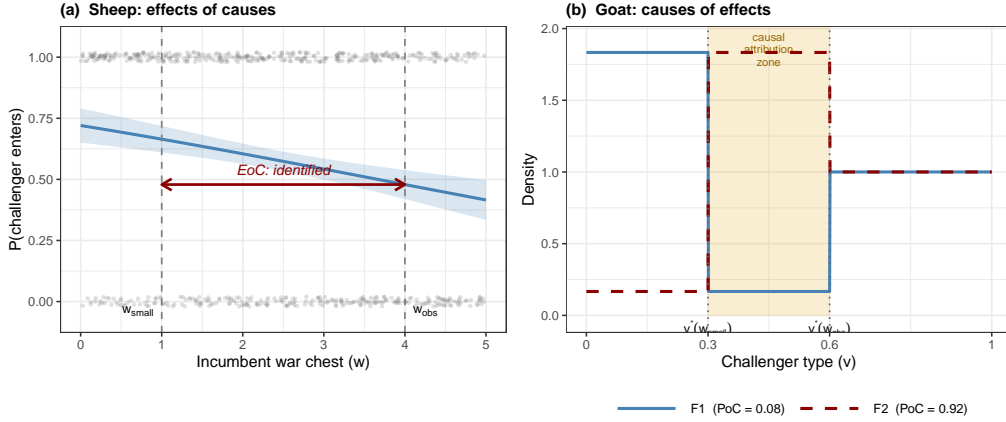


Figure 4: Panel (a) illustrates the EoC quantity (sheep): with exogenous variation in incumbent war chest the estimated entry probability and its confidence band are identified and the band narrows with sample size. The double-headed arrow marks the estimable shift from w_{small} to w_{obs} . Panel (b) illustrates CoE non-identification (goat): two challenger type distributions, F_1 and F_2 , both imply the same 40% entry rate at w_{obs} but yield probabilities of causation of 0.08 and 0.92 respectively. The orange shaded band is the causal attribution zone $[v^*(w_{small}), v^*(w_{obs})]$; the mass each distribution places in this zone determines the PoC. The distributions are observationally equivalent at the equilibrium war chest and cannot be distinguished by additional data at w_{obs} .

they would have stayed out regardless of the incumbent's investment, and the probability of causation is 0.08. Under F_2 , most non-entrants have moderate quality, clustered in the causal attribution zone $[v^*(w_{small}), v^*(w_{obs})]$; they were just barely deterred, and the probability of causation is 0.92. Neither distribution is falsified by the data, because the identifying variation — what entry rates look like at w_{small} — is never generated by the equilibrium. The incumbent always invests at w^* ; the off-equilibrium war chest is never tested. Adding observations at w_{obs} does not help because it refines only $F(v^*(w_{obs}))$, which is already identified. The CoE query is a goat.

5.4 A Goat: Attributing a Specific Non-Entry

Now consider a different query: *did this incumbent's war chest cause this specific challenger to stay out?*

This is a CoE query. It is backward-looking, it concerns a specific unit, and its answer requires a counterfactual: what would this challenger have done if the incumbent's war chest had been smaller? To answer it, we need the determining concomitant — the challenger's type v and the cutpoint $v^*(w)$ — evaluated at the specific observed case.

The problem is visible in Figure 3. The observed case (the filled point) sits in the Stay Out region. We observe w_{obs} (the incumbent's war chest) but not v (the challenger's quality). To determine whether the war chest caused the non-entry, we need to know whether v fell above or below $v^*(w)$ for the counterfactual war chest. If v was already well below $v^*(w_{\text{obs}})$, the challenger would have stayed out regardless of how much the incumbent spent; the war chest had no causal role. If v was just below $v^*(w_{\text{obs}})$ and just above $v^*(w_{\text{small}})$, then a smaller war chest would have induced entry and the incumbent's spending was decisive. We cannot distinguish these cases from the observed outcome alone. The challenger's type is private information — exactly the incomplete information failure mode identified in Section 4.

This is a goat not because the game-theoretic reasoning is wrong — the cutpoint model is a reasonable account of the strategic situation — but because the determining concomitant cannot be filled in from the observed non-event. A challenger who did not run left no record of their type. The war chest, the electoral environment, and the equilibrium cutpoint are all in principle observable; the challenger's private assessment of their own prospects is not. The CoE query requires precisely the variable that the equilibrium hides.

The non-event structure of the goat is characteristic of deterrence arguments generally, and the candidate entry case inherits this structure directly. The parallel to international deterrence is not coincidental: both involve an equilibrium in which the threat is never tested, the off-path node is never reached, and the counterfactual — what would have happened if the challenger had entered, if the aggressor had attacked — is exactly the information destroyed by the equilibrium's success.

5.5 The Repeated Game Case

The goat is not permanent. In repeated electoral settings — the same incumbent facing overlapping pools of potential challengers across multiple election cycles — the off-path behavior becomes observable over time, and the determining concomitant becomes estimable.

Consider an incumbent who serves five terms. In some cycles, challengers enter; in others, they do not. Across cycles, we accumulate observations of: which challengers entered at which levels of incumbent investment, what their quality was (estimated from their subsequent electoral performance or professional background), and how the incumbent responded when challenged. This is the data needed to estimate the cutpoint function $v^*(w)$ and the challenger type distribution $F(v)$. With these estimated, the determining concomitant is empirically grounded and CoE analysis becomes, at least in principle, tractable: we can ask whether a specific non-entry in cycle t was caused by the incumbent’s investment, conditional on the estimated type distribution and cutpoint.

Two caveats apply. First, this analysis yields population-level attribution (*what fraction of non-entries in cycles like this one were caused by deterrence?*) more naturally than individual attribution (*was this specific non-entry caused by deterrence?*). The individual CoE query still requires the unobserved type v for the specific challenger in question; the population CoE query requires only the estimated type distribution. The former remains hard; the latter is recoverable from the repeated game data.

Second, the repeated game rescue requires stability of the game form across cycles. If the payoff function $\pi_C(v)$ or the cost function $c(w)$ shifts across cycles — due to redistricting, changes in the national environment, or shifts in party strength — the data from different cycles do not all identify the same determining concomitant. Structural stability is condition (a) of Proposition 2 applied across time: the game must be correctly and stably specified for the accumulation of off-path observations to license CoE inference. The sequential decision structure of repeated electoral games maps directly onto the dynamic treatment regime literature (Robins 1986; Murphy 2003; Blackwell 2013): the incumbent’s investment sequence (w_1, w_2, \dots) is a dynamic regime adapted to evolving electoral conditions, and the population CoE query is the analog of an optimal DTR estimation problem. Blackwell (2013) develops exactly this framework for political science, showing that action sequences across a campaign are treatment histories in the dynamic

causal inference sense and that marginal structural models can recover causal effects of sequential strategies under sequential ignorability — the dynamic analog of Dawid’s extended conditional independence condition.

6 Conclusion

Game-theoretic models are unusual among social-scientific frameworks in that they explicitly provide the mechanistic structure Dawid (2000) requires for causal inference. A correctly specified game with a uniquely selected equilibrium gives us a strategy profile, a payoff function, and an information structure: together, these constitute the determining concomitant that grounds causal claims by specifying what would happen under any treatment realization, including those never observed on the equilibrium path. This is not a property of most other models political scientists use. A regression does not specify a mechanism. A matching estimator does not tell us what the untreated units’ outcomes would have been under treatment. A structural VAR does not ground counterfactual attribution. A game does, when it is correctly specified and its equilibrium is well-defined. In this sense, game theory is not merely compatible with serious causal inference — it is, when used carefully, among the best tools available for it.

The qualification matters. The conditions of Proposition 2 are demanding. Correct specification is harder than it sounds: it requires not only the right action space and information structure, but payoff functions that are identified from data rather than assumed. Unique equilibrium selection is often unavailable: many games of political interest admit multiple equilibria, and the solution concept does not select among them without additional assumptions. Observable information sets are a strong requirement in settings with private information, incomplete information, or unobservable actions. When any of these conditions fails, the game provides the *form* of a determining concomitant without its substance: the strategy space and equilibrium concept carve out the relevant counterfactual territory, but they cannot be filled in from the observable data. The game-theoretic causal claim then runs ahead of its epistemic warrant.

The deterrence success problem is the most vivid instance of this gap. When the equilibrium outcome is non-entry or non-escalation, the off-path node at which the potential responder acts is never reached. We observe the peace but not the mechanism that produced it. The game tells us what we

would need to know to attribute the peace to the deterrent: the potential aggressor’s type, the defender’s actual response strategy at the unreached node, the payoff gradient near the equilibrium. But a single peaceful outcome — a single realization of non-entry, non-escalation, non-attack — tells us none of this. More observations of peace do not help. The structural identification problem is not a power problem. It is inherent in the equilibrium’s success.

The practical implication is a discipline political scientists should apply when presenting game-theoretic causal arguments. *Effects-of-causes* queries — what would happen if we intervened to change the incumbent’s margin, the balance of forces, the information available to one player — are tractable whenever the game provides a well-specified predictive model. Signorino-style strategic estimation identifies these quantities directly: the estimated model computes predicted outcome distributions under alternative strategy profiles, and the causal interpretation is EoC throughout. These are sheep, and they should be presented as such: conditional on the model, here is the predicted effect of the regime change. *Causes-of-effects* queries — did this war chest cause this non-entry, did this alliance cause this peace — require the additional conditions of Proposition 2 and should be presented as conditional on those assumptions being met. When they are not met, the analysis should stop at the EoC question.

This framework clarifies the relationship between structural modeling and the credibility revolution — a relationship more complementary than competitive. Regression discontinuities, natural experiments, and quasi-experimental designs identify EoC quantities: they estimate the distribution of outcomes under a shift in the treatment regime, without requiring a structural model. Game-theoretic models provide structural interpretation of what the identified quantity means — through what mechanism the treatment operates, what the equilibrium logic is, why the effect has the sign and magnitude it does. The RD estimate of incumbency advantage and the Jacobson-Kernell model are not competing accounts of the same phenomenon; they are complementary accounts operating at different levels of the causal story. Credibility revolution tools identify the sheep; structural models explain why it is a sheep.

Four extensions merit future work. First, the incomplete-information case: when players have private types, those types are candidate determining concomitants. Type revelation through repeated interaction — and the partial identification results of Tamer (2003) on games with multiple equilibria — may offer a path to CoE analysis in Bayesian games that the present

framework does not cover. Second, dynamic games and sequential treatment regimes: subgame-perfect equilibrium strategies are sequential decision rules, and they map directly onto the dynamic treatment regime literature (Robins 1986; Murphy 2003). Blackwell (2013) brings this framework into political science, showing that action sequences across a campaign constitute treatment histories and that marginal structural models can recover causal effects of sequential strategies — the same structure the present framework identifies as the repeated-game path to CoE tractability. The intersection of strategic modeling and optimal dynamic treatment regime estimation is largely unexplored. Third, mechanism design as causal policy analysis: designing a game form to achieve a desired outcome distribution is Dawid’s EoC analysis applied at the level of the mechanism rather than within it. The welfare theorems of mechanism design are, in this reading, structural causal claims about what outcome distributions are achievable under alternative game-form regimes — a reformulation that may clarify both what mechanism design delivers and what assumptions it requires. Fourth, bridging to potential-outcomes notation: Richardson and Robins (2023) establish that Dawid’s augmented DAG framework is formally equivalent to single-world intervention graphs (SWIGs); extending the present isomorphism result to SWIG representations may make the framework’s conclusions available to researchers who prefer the Rubin causal model’s notation, without requiring them to adopt the decision-theoretic vocabulary.

Games were always Dawidian. Political scientists have been doing decision-analytic causal inference in strategic settings for decades, drawing on the same expected-utility foundations that Dawid’s framework formalizes. What has been missing is the explicit epistemological accounting: which queries the framework licenses, which it does not, and what conditions make the difference. This paper provides that accounting. The tools were already there; the sheep just needed sorting from the goats.

References

- Blackwell, Matthew. 2013. “A Framework for Dynamic Causal Inference in Political Science.” *American Journal of Political Science* 57 (2): 504–519.

- Caughey, Devin, and Jasjeet S. Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942–2008." *Political Analysis* 19 (4): 385–408.
- Dawid, A. Philip. 1979. "Conditional Independence in Statistical Theory." *Journal of the Royal Statistical Society, Series B* 41 (1): 1–31.
- . 2000. "Causal Inference Without Counterfactuals." *Journal of the American Statistical Association* 95 (450): 407–424.
- . 2021. "Decision Theoretic Foundations for Statistical Causality." *Journal of Causal Inference* 9 (1): 1–27.
- . 2022. "Decision Theoretic Foundations for Statistical Causality: Response to Shpitser." *Journal of Causal Inference* 10 (1).
- Dawid, A. Philip, and Monica Musio. 2016. "Statistical Causality from a Decision-Theoretic Perspective." *Annual Review of Statistics and Its Application* 3:273–303.
- Fearon, James D. 1991. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43 (2): 169–195.
- . 1995. "Rationalist Explanations for War." *International Organization* 49 (3): 379–414.
- Hammond, Lewis, James Fox, Tom Everitt, Alessandro Abate, Michael Wooldridge, and Nicholas Jennings. 2023. "Reasoning about Causality in Games." *Artificial Intelligence* 320:103919.
- Harsanyi, John C. 1973. "Games with Randomly Disturbed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points." *International Journal of Game Theory* 2 (1): 1–23.
- Hudgens, Michael G., and M. Elizabeth Halloran. 2008. "Toward Causal Inference with Interference." *Journal of the American Statistical Association* 103 (482): 832–842.
- Huth, Paul K. 1988. *Extended Deterrence and the Prevention of War*. Yale University Press.
- Jacobson, Gary C., and Samuel Kernell. 1983. *Strategy and Choice in Congressional Elections*. Yale University Press.

- Koller, Daphne, and Brian Milch. 2003. “Multi-Agent Influence Diagrams for Representing and Solving Games.” *Games and Economic Behavior* 45 (1): 181–221.
- Lee, David S. 2008. “Randomized Experiments from Non-random Selection in U.S. House Elections.” *Journal of Econometrics* 142 (2): 675–697.
- Lewis, David. 1973. *Counterfactuals*. Harvard University Press.
- McKelvey, Richard D., and Thomas R. Palfrey. 1995. “Quantal Response Equilibria for Normal Form Games.” *Games and Economic Behavior* 10 (1): 6–38.
- Murphy, Susan A. 2003. “Optimal Dynamic Treatment Regimes.” *Journal of the Royal Statistical Society, Series B* 65 (2): 331–355.
- Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. 2nd. Cambridge University Press.
- . 2022. “Causation and Decision: Decision Theoretic Foundation for Statistical Causality.” *Journal of Causal Inference* 10 (1): 265–270.
- Powell, Robert. 1999. *In the Shadow of Power: States and Strategies in International Politics*. Princeton University Press.
- Richardson, Thomas S., and James M. Robins. 2023. “Potential Outcome and Decision Theoretic Foundations for Statistical Causality.” *arXiv:2302.03899*.
- Robins, James M. 1986. “A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period.” *Mathematical Modelling* 7:1393–1512.
- Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66 (5): 688–701.
- Sartori, Anne E. 2003. “An Estimator for Some Binary-Outcome Selection Models without Exclusion Restrictions.” *Political Analysis* 11 (2): 111–138.
- Schelling, Thomas C. 1960. *The Strategy of Conflict*. Harvard University Press.

- Selten, Reinhard. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4 (1): 25–55.
- Signorino, Curtis S. 1999. "Strategic Interaction and the Statistical Analysis of International Conflict." *American Political Science Review* 93 (2): 279–297.
- . 2002. "Strategy and Selection in International Relations." *International Interactions* 28 (1): 93–119.
- Signorino, Curtis S., and Ahmer Tarar. 2006. "A Unified Theory and Test of Extended Immediate Deterrence." *American Journal of Political Science* 50 (3): 586–605.
- Signorino, Curtis S., and Kuzey Yilmaz. 2003. "Strategic Misspecification in Regression Models." *American Journal of Political Science* 47 (3): 551–566.
- Sobel, Michael E. 2006. "What Do Randomized Studies of Housing Mobility Demonstrate? Causal Inference in the Face of Interference." *Journal of the American Statistical Association* 101 (476): 1398–1407.
- Tamer, Elie. 2003. "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria." *Review of Economic Studies* 70 (1): 147–165.