Fighting in the Shadow of Intervention: A Learned-Proxy Analysis

Robert J. Carroll

March 12, 2026

Abstract

Theory predicts that anticipated third-party intervention shapes the calculus of rebellion, yet no measure of expected intervention covers the full population of potential interveners. I construct one using a two-stage learned-proxy design: a machine-learning ensemble predicts the probability and direction of intervention for each directed dyad-year; predictions are aggregated to a country-year shadow disciplined by a Nash fixed-point condition. Expected government-biased intervention deters onset while expected opposition-biased intervention encourages it. The two shadow variables are jointly significant ($p < 0.001$) despite wide individual confidence intervals driven by inter-shadow collinearity, and the directional pattern is robust across specifications, measurement draws, and a country fixed-effects estimator. The shadow subsumes two leading existing proxies and carries out-of-sample predictive content.

Keywords: civil war, military intervention, machine learning, learned proxy, onset
Word count: 7,635

A growing body of theory predicts that *anticipated* third-party intervention shapes civil war onset (Cetinyan, 2002; Cunningham, 2016; Dudley, 2024; Gibilisco and Monteiro, 2022; Kuperman, 2008; Kydd and Straus, 2013; Langø, 2023; Sambanis et al., 2020), yet the quantity has never been measured systematically. No existing measure covers the full population of potential interveners, varies annually with shifting alliances and foreign policy alignments, and treats intervention direction—government-biased versus opposition-biased—as the primary quantity of interest. This paper constructs such a measure and uses it to test the hypothesis that the shadow of intervention deters and emboldens.

The construction is the core contribution. A machine-learning ensemble trained on the Regan (2000) dataset of military interventions (1946–2014) predicts, for each directed dyad in a given year, the probability and direction of intervention. Dyad-level predictions are aggregated across all potential interveners to produce a country-year shadow, disciplined by a Nash fixed-point condition requiring each state's predicted probability to be self-consistent as an input to the others' predictions. The resulting measures are tested in a standard country-year onset framework (Fearon and Laitin, 2003): expected government-biased intervention enters negatively (deterrence), expected opposition-biased intervention enters positively (emboldening), with the directional pattern robust across specifications, measurement draws, and a country fixed-effects estimator.

The closest existing approaches each sacrifice scope for identification. Cunningham (2016) uses the Lake security hierarchy as a proxy for anticipated government-biased intervention—a static, unidirectional measure restricted to the US patron-client channel. The present paper generalizes in three directions: it covers all potential interveners, both sides of intervention, and derives probabilities from a calibrated classifier rather than structural position. Gibilisco and Monteiro (2022) derive equilibrium intervention expectations from a structural game among the P5, gaining clean identification at the cost of restricting the player set to a size that keeps joint-game estimation tractable. In the Regan data, P5 states account for fewer than half of all military interventions; the remaining 53 intervening states—neighbors, coethnics, regional rivals, Cold War proxies—represent the majority of actual deployments and are driven by logics qualitatively different from the global-order interests that animate P5 behavior. Langø (2023) develops a formal model in which the threat of rebel-sided intervention can simultaneously deter onset—by compelling governments to concede—and encourage it—by raising rebels' expected return from fighting; the utility parameters in the present framework can represent these nonmonotone effects.

This paper takes a learned-proxy approach: a flexible first-stage model constructs the latent quantity, and a parametric second-stage model tests the theory. This two-stage design is an increasingly common way to bring machine learning into social science inference (Knox et al., 2022), and arguably the least controversial (Morucci and Spirling, 2024): the ensemble does the measurement work it is suited for, while the onset regression retains the interpretability that causal claims require. The design nests the structural functional form—the unpenalized multinomial logit in the ensemble library directly corresponds to the softmax best-response function of Gibilisco and Monteiro (2022)—so whether that parametric assumption fits the data is answered empirically rather than imposed by construction. (It earns less than one percent of the ensemble weight; the data strongly prefer nonparametric alternatives.)

The learned-proxy approach has well-documented pitfalls. Knox et al. (2022) catalog these systematically: measurement-stage uncertainty is routinely ignored, proxy quality is asserted rather than tested, and the disconnect between prediction and inference is papered over. This paper follows their recommendations closely: Stage 1 performance is validated out-of-fold across 25 imputation draws; measurement uncertainty is propagated into Stage 2 via a two-stage pairs cluster bootstrap that averages across all 25 draws; calibration is verified by reliability diagrams; and a labeled-only robustness check confirms that the onset estimates are not driven by out-of-sample extrapolation. The honest accounting of generated-regressor uncertainty is itself a result: the corrected standard errors are roughly 2.9–3.5 times larger than naive MLE output, with a variance decomposition showing that approximately 88–92% of the total coefficient variance originates in the measurement stage and only 8–12% from primary-analysis sampling. The measurement model, not the regression, is the binding source of uncertainty—precisely the situation Knox et al. (2022) warn about. The two shadow variables are correlated at $r = 0.94$, which inflates individual confidence intervals, but tested jointly they reject the baseline at $p < 0.001$; the directional pattern holds across all 25 measurement draws, ten specifications, and country fixed effects. The shadow subsumes both the Lake security hierarchy (Cunningham, 2016) and the Gibilisco and Monteiro (2022) structural P5 estimates: neither adds explanatory power once the shadow variables are included.

The paper proceeds as follows. Section 1 develops the theoretical frame-

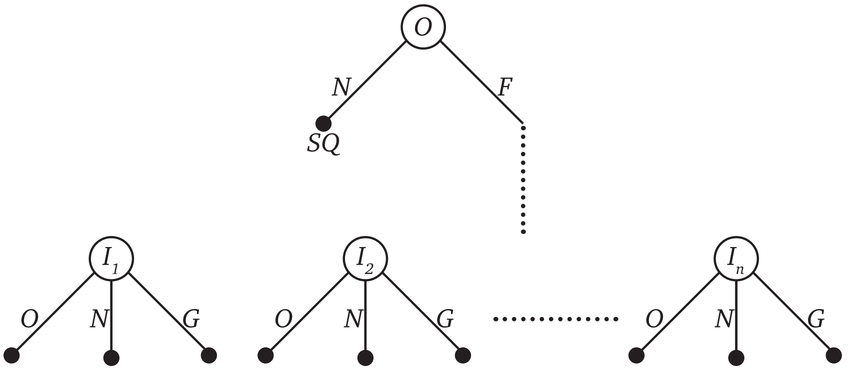

Figure 1: Motivating formal model.

work linking intervention expectations to the onset decision. Section 2 describes the construction and validation of the shadow measure. Section 3 presents the onset analysis, including measurement-corrected inference and heterogeneous-utility extensions.

1 Intervention and the Onset Decision

What does it mean, formally, for an opposition group to act in the shadow of intervention? This section develops a model of the onset decision that derives the functional form that intervention expectations must take to be both theoretically coherent and empirically tractable.

1.1 The Model

The decision problem is depicted in Figure 1. An opposition group O chooses whether to Fight ($s_O = F$) or Not fight ($s_O = N$). Should it not fight, the game ends at the status quo SQ. If it fights, n potential interveners simultaneously choose actions $s_i \in \{O, N, G\}$— opposition-biased, non-intervention, or government-biased—and the combined profile $s \in S = \{O, N, G\}^n$ determines the international environment the opposition faces.

Each potential intervener i has a utility function $u_i : S \to \mathbb{R}$ over profiles. Under expected utility, let σ_i be a probability measure over $\{O, N, G\}$, and let $U_i(\sigma) = \sum_{s' \in S} u_i(s') \prod_j \sigma_j(s'_j)$. A Nash equilibrium σ^* of the intervention subgame satisfies $U_i(\sigma^*_i, \sigma^*_{-i}) \geq U_i(\sigma'_i, \sigma^*_{-i})$ for every σ'_i and every i. The equilibrium profile feeds back to the opposition, whose expected utility of fighting is $U_O(F;\sigma^*) = \sum_{s' \in S} u_O(s') \prod_i \sigma_i^*(s'_i)$. A subgame perfect equilibrium requires that σ^* be a Nash equilibrium of the intervention subgame and that

$$u_O(SQ) \geq U_O(F;\sigma^*) \iff s_O^* = N.$$

The opposition fights if and only if the expected return to fighting, given equilibrium intervention, exceeds the status quo. The international environment enters the calculus entirely through σ^*: the anticipated behavior of outside powers can make the difference between peace and war before a single shot is fired. This is the shadow of intervention (Cetinyan, 2002).[1]

1.2 From Equilibrium to Empirics

The equilibrium condition is theoretically exact but empirically intractable as stated. With up to 193 potential interveners each choosing among three options, the expectation over U_O runs over 3^{192} profiles—larger than the number of elementary particles in the observable universe. Two assumptions reduce the problem to a tractable form without discarding its theoretical content.

Assumption 1 (Separability). The opposition's payoff from any profile is additive across interveners: $u_O(s') = \sum_i u_O^i(s'_i)$, where $u_O^i : \{O, N, G\} \to \mathbb{R}$ measures the marginal contribution of intervener i's action.

Separability captures a specific and substantively defensible theory of how opposition groups aggregate international expectations: each outside power's likely behavior enters the calculus independently, and the combined effect is their sum. A rebel coalition assessing its international environment in 1970s Angola had a view about what Cuba would likely do, a separate view about South Africa, and a separate view about the United States; the hypothesis is that these entered additively. The assumption rules out complementarities—the value of Soviet logistics may depend on whether Cuban troops are also

[1] Whether σ^* approximates the expectations that rebels actually form is not merely assumed. Under rational expectations, agents exploit available information at least as well as the researcher's model (Muth, 1961). The Stage 1 classifier described in Section 2 recovers a substantial share of the predictive content in realized intervention outcomes using only publicly observable state characteristics; rebels who can also draw on private intelligence, direct patron relationships, and local knowledge face a forecasting problem that is easier, not harder, than the one the classifier solves.

present—but coordinated multilateral operations constitute a small minority of actual interventions.[2]

Separability also enables the extension to the full system of potential interveners. Because each state's contribution enters Equation (1) additively, computing the opposition's expected utility requires only the marginal intervention probabilities σ_i—not the joint distribution over all n players' actions simultaneously. Specifying and estimating the full joint game—the approach of Gibilisco and Monteiro (2022) for the five permanent Security Council members—becomes computationally intractable beyond a small player set. Separability sidesteps that constraint while preserving the game's central implication: the equilibrium strategies of outside powers must be mutually consistent. That consistency requirement is enforced directly via the Nash fixed-point iteration described in Section 2.

Assumption 2 (Non-intervention as reference). For all i, $u^i_O(N) = 0$. This normalization makes the intervention-free onset model a limiting case: if no state intervenes, the expected utility of fighting reduces to purely domestic considerations, and the standard country-year onset model follows directly.

Under these two assumptions, the expected utility of fighting simplifies to

$$U_O(F;\sigma) = \sum_{i=1}^{n} \left[\sigma_i(O)\, u^i_O(O) + \sigma_i(G)\, u^i_O(G)\right]. \tag{1}$$

Equation (1) is the primary theoretical object, and the two-stage empirical design follows directly from its structure. The equation contains two types of unknowns: the intervention probabilities $\sigma_i(O)$ and $\sigma_i(G)$, and the marginal utilities $u^i_O(O)$ and $u^i_O(G)$ that scale each. Stage 1 (Section 2) estimates the probabilities for every directed dyad using a machine-learning classifier trained on the Regan intervention record. Stage 2 (Section 3) recovers the utility parameters from their aggregate effect on civil war onset. The

[2] Aydin and Regan (2012) find that the *alignment* among interveners matters for civil war duration: same-side interventions shorten wars only when the intervening states share similar preferences. This is evidence of complementarities in the conduct phase. Separability here pertains to the onset calculus—the rebel's assessment of the intervention environment at the moment it decides whether to fight—not to the downstream interactions among interveners once fighting begins. The spatial lags in the Stage 1 classifier (Section 2) capture interdependence among interveners' *choices* without requiring that their *effects* on the rebel enter non-additively.

theory does not merely motivate the empirical strategy; it specifies exactly what each stage must estimate and how the estimates connect.

The model assigns the strategic decision to the opposition, not because governments are passive but because this is where the onset decision sits—following the literature's longstanding focus on the conditions that facilitate or inhibit insurgency (Fearon and Laitin, 2003). Governmental behavior enters implicitly through the opposition's status quo utility and the host-country features in the Stage 1 classifier; endogenizing the government's response would require a bargaining game that severs the tight deductive connection between Equation (1) and its empirical counterpart.

2 A Learned Proxy for the Shadow of Intervention

Operationalizing the shadow requires a measure of the expected intervention environment at the moment a potential rebel group weighs the costs and benefits of open conflict. The construction proceeds in two stages: a directed-dyad classifier estimates intervention probabilities for every potential intervener, and a country-year aggregation sums them under a Nash fixed-point condition ensuring self-consistency.

2.1 Outcome Variable

The Stage 1 training labels draw primarily from Regan (2000), whose dataset records third-party military, economic, and diplomatic interventions in civil conflicts from 1944 to 1999. I extend Regan's coding through 2014 by hand-coding unambiguous military interventions in post-1999 COW onset country-years. The coding threshold requires deployment of military personnel—troops, combat advisors, or directed proxy forces attributable to a specific state—on behalf of one side; arms transfers, economic aid, and sanctuary alone are insufficient. UN peacekeeping missions are excluded. The post-1999 extension adds 31 directed dyad-year records across 14 host countries, yielding a combined intervention table of 455 records (1944–2014); the full coding with source notes appears in Online Appendix A.

The civil war universe is drawn from the Correlates of War Intra-State War Dataset (v5.1) (Dixon and Sarkees, 2016), restricting to wars fought for central control (type 4) or local/secessionist issues (type 5). The COW threshold (≥1,000 battle deaths) is theoretically appropriate: the mechanism requires rebels organized enough to form rational expectations and conflicts large enough that outside military involvement is a realistic prospect.[3] This yields 153 distinct onset country-years between 1946 and 1999 and 191 through 2014. I restrict attention to *military* interventions, coded as *government-biased* (1) or *opposition-biased* (2); corrections and deduplication are documented in Online Appendix A.

For each directed dyad ($B \rightarrow A$) active in a civil war onset year, the training label is: 0 (no military intervention), 1 (government-biased), or 2 (opposition-biased). Non-onset dyad-years are excluded from the training set. This produces 254 intervention-coded onset observations—150 government-biased and 104 opposition-biased—across 51 host countries and 58 unique intervening states. The five permanent Security Council members account for 110 of these events (43 percent); the remaining states—Cold War proxies, neighbors, regional patrons, and coethnics—account for the majority (Table 1).

Training exclusively on onset country-years is not a convenience—it is the only option. Intervention in a civil war is undefined absent a civil war; in non-onset years the outcome is not "zero intervention" but rather unobserved. The consequence is that the training sample is selected on precisely the outcome that the Stage 2 analysis tries to explain: if the shadow deters onset, then high-shadow country-years are systematically underrepresented among onsets, and the classifier learns the feature–intervention mapping from an endogenously filtered sample. There is no design that avoids this. The defense rests on what the classifier is asked to learn. The training target is the conditional relationship between *dyad-level* features—alliance portfolios, capability ratios, colonial ties, geographic proximity, regime similarity—and the direction of military intervention, given that a war has started. These bilateral characteristics are plausible determinants of how state B behaves toward state A in a conflict regardless of what brought that conflict about. The fixed-point iteration then handles the extrapolation: it asks what the equilibrium intervention environment *would* look like in any country-year, including those where onset was never observed, without requiring that the training sample be representative of the full population.

[3] At lower thresholds—such as the UCDP/PRIO ≥25-death criterion—third-party military intervention becomes vanishingly rare, swamping any predictive signal.

State	Gov.	Opp.	Total
United States†	31	15	46
USSR / Russia†	16	7	23
United Kingdom†	9	8	17
France†	11	6	17
Cuba	7	1	8
China†	2	5	7
Libya	1	6	7
Iran	2	5	7
Somalia	0	5	5
Syria	3	2	5
India	4	1	5
Uganda	3	2	5

Table 1: Most frequent military interveners (1946–2014); states with fewer than five interventions omitted. † Permanent member of the UN Security Council (P5). Government-biased (Gov.) and opposition-biased (Opp.) interventions are coded from the Regan `target` variable and post-1999 extension.

2.2 Feature Set

The classifier draws on 105 directed-dyad-year features organized around four substantive dimensions (full list in Online Appendix C). Many variables enter twice—once for the potential intervener and once for the host—so that the classifier can learn asymmetric effects. Missing data are multiply imputed using miceforest (five country-year × five undirected-dyad imputations, yielding 25 complete datasets; Online Appendix B).

- **State characteristics of both the potential intervener and the host.** Material capabilities (CINC scores and components, GDP, population) from COW v6 (Singer, 1987) and V-Dem v15 (V-Dem Institute, 2025); regime type (Polity II and lags, liberal democracy index, instability); power status (COW major power, P5 membership); and host-side conflict history, ethnic fractionalization (Fearon and Laitin, 2003), ethnic exclusion (Cederman et al., 2010), oil wealth, and terrain.

- **Bilateral ties.** Alliance commitments (Gibler, 2009), bilateral trade, geographic proximity (log capital distance, contiguity Stinnett et al., 2002), colonial history, territorial disputes (Huth and Allee, 2003), expected dispute outcomes (Carroll and Kenkel, 2019), and peace quality (Diehl et al., 2021).

- **Foreign policy alignment.** UN General Assembly ideal points (Bailey et al., 2017) for each state and the absolute distance between them; shared IGO membership from COW Dyadic IGO v3 (Pevehouse et al., 2020).
- **Spatial context.** Ten spatial lag variables constructed from a row-normalized polity-similarity weight matrix capture the weighted behavior of other potential interveners in the same host-year: the fraction coded government- or opposition-biased, US and Soviet intervention posture, and four superpower interaction terms (Gent, 2007; Salehyan, 2007) (Online Appendix D).

Before training, the 115 features (95 dyadic covariates + 20 spatial lags) are standardized and reduced via principal components analysis, retaining components to 90% of cumulative variance (59–61 components across all 25 imputation datasets).

2.3 Classifier Design

The multinomial outcome (0 = no intervention, 1 = government-biased, 2 = opposition-biased) is modeled using a super-learner ensemble (van der Laan et al., 2007).

Nine component classifiers spanning three model families are trained within a ten-fold cross-validation scheme: two tree ensembles (a random forest Breiman, 2001 and two histogram-based gradient boosting specifications at different learning rates), four logistic regression variants (ridge, elastic-net, lasso, and unpenalized multinomial), and two multi-layer perceptrons (shallow and deeper architectures). The unpenalized multinomial logistic regression—a softmax best-response function—directly nests the functional-form assumption of structural game-theoretic models of intervention (Gibilisco and Monteiro, 2022); its stacking weight is determined by out-of-fold predictive performance, so whether that assumption fits the data is answered empirically rather than imposed by construction. Each classifier produces a vector of class probabilities for every training observation. The super learner combines these using non-negative least squares (NNLS) weights chosen to minimize held-out multinomial log-loss, so that the ensemble adaptively upweights component classifiers that predict well on out-of-fold observations.

All reported performance statistics are based on out-of-fold predictions to guard against overfitting—following Wang's (2019) demonstration that evaluation of class-imbalanced political-event models on training data produces severely inflated performance estimates (Muchlinski et al., 2016). The primary performance metric is the proportional reduction in log-loss relative to the class-frequency null model (PRL); the area under the ROC curve (AUC) is also reported.

Spatial lags and Nash fixed-point iteration. The broader motivation for the two-stage design is that intervention decisions are endogenous to the conflict environment: interveners select into conflicts strategically, and a single-equation model that treats realized intervention as a covariate in an onset regression conflates the causal effect with the selection process that generated it (Signorino, 2002; Gent, 2008). Qualitative evidence confirms the problem directly: rebel groups in Bosnia and Kosovo explicitly conditioned their onset decisions on anticipated outside support, so realized intervention is partly a consequence of expected intervention rather than an independent cause (Kuperman, 2008). Using predicted intervention expectations rather than realized intervention is the appropriate correction.

The spatial context variables described in Appendix D depend on the intervention choices of *other* potential interveners—quantities that are themselves to be predicted. A documented pattern of countervailing intervention makes these dependencies quantitatively important: when one state intervenes on the government side, ideologically rival states face a selective incentive to support the opposition, and vice versa (Salehyan et al., 2011; Findley and Teo, 2006). Aydin and Regan (2012) confirm that these network effects are consequential: opposing interventions nearly double civil war duration, while same-side interventions shorten wars only when the intervening states share similar foreign policy preferences—evidence that the strategic alignment among interveners, not merely their count, carries the signal. This creates a concrete measurement problem. Without an explicit consistency requirement, spatial lags are computed from *observed* training-data intervention choices, which are not the model's own outputs. Applied out-of-sample—to country-years unlike the training-period equilibrium—the predictions embed a deterministic discrepancy between lag inputs and probability outputs whose sign and magnitude depend on how far the out-of-sample world departs from training-period behavior. This is systematic error, not sampling noise, and in a measure designed to capture counterfactual intervention environments the bias could be large.

The fix is to require that the predictions be self-consistent inputs to themselves:

$$\sigma^*(B, A, t) = \hat{M}(X(B, A, t, \sigma^*)) \quad \forall\, (B, A, t), \tag{2}$$

where $\hat{M}$ is the trained super-learner classifier. At σ^*, the spatial lags are derived from the model's own equilibrium predictions, not from historical observations. The deterministic discrepancy disappears by construction.

This self-consistency condition is also the Nash equilibrium condition for a natural class of games. The primitive object is the best-response correspondence, not the game itself: a game specifies a best-response correspondence *derivatively*, via utility maximization; $\hat{M}$ specifies one *directly*, from the data. Nash equilibrium is a fixed point of the best-response correspondence and does not require identifying the payoff functions that generate it. Games can be partitioned into equivalence classes under the relation "generates the same best-response correspondence"; σ^* is the Nash equilibrium of every game in the class whose correspondence $\hat{M}$ estimates.[4] If the training data reflect equilibrium play and the spatial lags are sufficient statistics for

[4] The set of rationalizing payoff functions for any interior σ^* is non-empty—choose utilities that make each player indifferent among actions assigned positive probability. What is identified is the correspondence; the payoffs that generate it are not. A natural parametric fiber of these classes is the linear-utility family $u_i(a, \sigma_{-i}) = \alpha \cdot f_i(a) + \beta \cdot W_i(a, \sigma_{-i})$, where f_i captures state-specific intervention incentives and W_i the spatial interaction term. The best-response from this family is a softmax—the functional form of the multinomial logit component of the super-learner—so the parametric case is explicitly represented in the ensemble while the broader non-parametric mixture accommodates departures from it. The iterative procedure that solves Equation (2) is the nonparametric analogue of the nested pseudo-likelihood (NPL) algorithm of Aguirregabiria and Mira (2007): the fitted super-learner replaces the parametric best-response function, and the fixed-point iteration updates beliefs (spatial lags) until self-consistency, as in A&M's two-step PML. In the parametric case, Crisman-Cox and Gibilisco (2021) show that this class of iterative estimators substantially outperforms direct maximum likelihood estimation of strategic games under equilibrium multiplicity. Uniqueness of the fixed point is not guaranteed in general, but convergence is rapid (two to three passes across all 25 imputation draws), consistent with a contractive mapping.

the strategically relevant information in σ_{-i}, then $\hat{M}$ is a non-parametric estimate of the true best-response correspondence BR^* and σ^* estimates its fixed point, without specifying which game in the equivalence class is the true data-generating process. The statistical and game-theoretic justifications therefore coincide: Equation (2) corrects a concrete measurement error and simultaneously identifies the equilibrium of the underlying intervention game, whichever member of the equivalence class that game happens to be.

Equation (2) has no closed form but is solved by fixed-point iteration. The iteration proceeds in two stages. In the *training stage*, the classifier is first fit using observed spatial lags; out-of-fold predictions then replace the observed intervention indicators to recompute spatial lags; and the classifier is refit on the updated lags. This loop repeats until convergence, typically within two to three passes. In the *prediction stage*, the trained classifier is held fixed and only the spatial lags are updated: the model predicts for every directed dyad-year (onset and non-onset alike), spatial lags are recomputed from those predictions, and the model re-predicts until convergence. Applying the fixed-point universally—not just to onset rows—ensures that the shadow reflects the counterfactual equilibrium in any country-year, including those where no conflict was observed.

2.4 Predictive Performance

Table 2 reports out-of-fold performance for the super-learner and two selected components: the lasso logistic regression (the best single component by standalone PRL) and the random forest (which despite lower standalone PRL receives the second-highest NNLS weight at 38.6%, reflecting complementary predictive information). Full results for all nine component learners appear in Appendix E, Table 10. All statistics average across the 25 complete datasets, with standard deviations across imputation draws in parentheses.

The ensemble achieves a PRL of 47.3% and a macro-average AUC of 0.969. A PRL near 47% indicates that the classifier recovers nearly half the information in intervention outcomes that the class-frequency null discards—a substantial figure for a three-class rare-event problem in which the modal outcome (no intervention) accounts for over 99% of observations. Interventions account for barely one percent of directed-dyad-year observations, so the predictive task is a needle-in-a-haystack problem. A natural question is whether the classifier earns its performance by finding the needles—correctly identifying the rare dyad-years where intervention occurs—or by calibrat-

Model	PRL	AUC (macro)	Log-loss
Super-learner	47.3% (1.0%)	0.969 (0.003)	0.033 (0.001)
MLP (100, 50)	33.4% (4.4%)	0.940 (0.014)	0.041 (0.003)
Lasso logit	35.1% (1.0%)	0.957 (0.003)	0.040 (0.001)

Table 2: Selected Stage 1 classifier performance (out-of-fold, 25 imputation draws). The MLP is the highest-weighted component (44.5% of the ensemble); lasso logit is the best parametric model. Standard deviations across draws in parentheses. The class-frequency null has PRL = 0% by construction, AUC = 0.500, and log-loss = 0.062. Full results for all nine component learners appear in Table 10.

ing the hay, sharpening predictions for the overwhelming non-intervention majority. The distinction is consequential: a model that only improved its calibration of non-events would replicate what any onset regression without shadow variables already does implicitly, sending all intervention probabilities to their unconditional near-zero rate. What the shadow measure requires is discrimination among rare events. Decomposing the log-loss reduction by true class confirms that the classifier delivers exactly this: 94 percent of the PRL originates from the 1.0 percent of observations where intervention actually occurred, with only 6 percent from improved calibration of non-events.

The result also speaks directly to the rational expectations assumption embedded in the theoretical model. The classifier is trained entirely on publicly observable state characteristics—material capabilities, alliance portfolios, regime type, colonial history, foreign policy alignment. If an estimator operating on this strict subset of information recovers 47% of the predictive content in realized intervention outcomes, then intervention is, to a substantial degree, foreseeable from information available to any careful observer. Under rational expectations, agents exploit available information at least as well as the analyst's model (Muth, 1961); rebel leaderships can additionally draw on private intelligence, direct communications with potential patron states, diaspora networks, and on-the-ground assessments of great power intentions that no dataset records. The 47% figure is accordingly a lower bound on how knowable the intervention environment is from the rebel's perspective, not a ceiling.

The performance gap between the ensemble and the best single logistic specification—47.3% vs 35.2% PRL on an identical feature set—reveals nonlinear and interactive structure that parametric models cannot fully recover. Morucci and Spirling (2024) show that in typical political science applications, complex models rarely outperform logistic regression because curated feature sets lack exploitable nonlinear structure; the intervention prediction task is an exception, precisely because the broad feature space is designed for measurement rather than theory testing. The NNLS stacking procedure quantifies this directly: it assigns near-zero weight to all four logistic specifications (ridge, lasso, elastic-net, unpenalized multinomial), concentrating ensemble mass on the deep multi-layer perceptron (44.5%) and the random forest (38.6%), with a gradient-boosted ensemble taking 11.1% (Appendix E, Table 10). Notably, the logistic family achieves the highest standalone PRL among individual learners yet receives negligible ensemble weight, indicating that the nonlinear learners subsume their predictive content while adding complementary information the logistic models miss. Because the logistic family is explicitly included in the candidate library and explicitly downweighted by the data, the ensemble's superiority is not an artifact of architecture choice.

Because Stage 1 probabilities are subsequently summed across approximately 190 potential interveners, systematic miscalibration compounds; reliability diagrams confirming calibration for each outcome class appear in Appendix E. These diagnostics follow Knox et al.'s (2022) recommendation to "always assess and report measurement performance" (their §5.4): the PRL and AUC quantify predictive accuracy, and the reliability diagrams verify the calibration assumptions on which the aggregation to country-year expectations depends.

2.5 Aggregating to Country-Year Intervention Expectations

The Stage 1 classifier produces, for each directed dyad ($B \to A$) in a civil war onset country-year, probability estimates $\hat{\sigma}_B(1)$ and $\hat{\sigma}_B(2)$ for government- and opposition-biased intervention, respectively. To reduce noise from potential interveners to whom the classifier assigns negligible probability, I apply a cutpoint filter: intervener B contributes to the aggregate only if $\hat{\sigma}_B(k) \geq \tau$ for $k \in \{1, 2\}$. The threshold τ is treated as a tuning parameter and selected by out-of-fold performance in the Stage 2 onset models; in practice, $\tau = 0.001$ performs best, excluding only the very least likely interveners. The threshold is permissive by design: the PRL decomposition above shows that the classifier's information content concentrates overwhelmingly in the rare dyads it flags as plausible interveners, so the cutpoint discards noise from the long tail of near-zero predictions while preserving essentially all of the signal.

The aggregate intervention-expectation variables for host country A in year t are

$$E_A^G(t) = \sum_{B:\, \hat{\sigma}_B(1) \geq \tau} \hat{\sigma}_B(1), \quad E_A^O(t) = \sum_{B:\, \hat{\sigma}_B(2) \geq \tau} \hat{\sigma}_B(2).$$

Both sums are mechanically larger in later years as the system expands. I include a year trend in all second-stage regressions to absorb this secular growth. The sums are also right-skewed; following Burbidge et al. (1988), I apply the inverse hyperbolic sine (asinh) transformation before entering them into the onset models.

2.6 Properties of the Shadow Measure

Before turning to the onset analysis, I examine the descriptive properties of E^G and E^O to establish that they behave as the theory predicts.

The dyad-level predictions reveal a structural asymmetry between the two sides of intervention. The highest predicted probabilities of government-biased intervention involve superpowers projecting influence into client states—the United States supporting the Philippines in 1972 ($\hat{p}_{\text{gov}} = 0.81$), Laos in 1968, Cambodia in 1972. For opposition-biased intervention, the picture shifts: the maximum is Somalia opposing the Ethiopian government in 1980 ($\hat{p}_{\text{opp}} = 0.77$), followed closely by Bulgaria targeting Greece in 1947 and Libya targeting Chad in 1984, and all ten of the highest-ranked opposition predictions involve non-P5 interveners. A measurement framework restricted to P5 states would miss the majority of the opposition-biased shadow.

Table 3 makes this concrete for three canonical onsets, reporting the five states to which the classifier assigns the highest government- and opposition-biased intervention probabilities alongside an indicator of whether the state actually intervened in that direction.

Across the three cases, 10 of 12 actual interventions appear in the top five on the correct side—and in each case, restricting the measure to P5 states would discard the single most important opposition-biased intervener: Zaire

Gov-biased	$\hat{p}$		Opp-biased	$\hat{p}$	
Angola (1976) — 5 actual interventions					
USSR*	.20	✓	Zaire	.49	
Cuba	.18	✓	South Africa	.19	✓
United States*	.15		United States*	.04	✓
Zambia	.07		Portugal	.03	
Portugal	.05		USSR*	.02	
Ethiopia (1975) — 3 actual interventions					
USSR*	.14	✓	Somalia	.66	✓
South Yemen	.12		United States*	.06	
United States*	.11		North Yemen	.06	
Iran	.04		Sudan	.04	
Saudi Arabia	.03		South Yemen	.02	
Afghanistan (1978) — 4 actual interventions					
USSR*	.29	✓	Pakistan	.45	✓
United Kingdom*	.11		United States*	.08	✓
China*	.09		United Kingdom*	.06	
United States*	.07		China*	.05	
Iran	.07		Iran	.04	✓

Table 3: Top five predicted interveners for three canonical onsets. Out-of-fold probabilities, averaged across 25 imputation draws. *P5 member. ✓ indicates the state actually intervened in that direction (Regan coding). Ten of 12 actual interventions appear in the top five on the correct side.

in Angola, Somalia in Ethiopia, Pakistan in Afghanistan.[5]

Figure 2 shows the cross-sectional distribution of E^G and E^O across all country-years. Both measures are heavily right-skewed: the median country-year faces modest intervention expectations, but a long right tail captures

[5] The United States appears in the top five government-biased predictions for all three cases, yet actually intervened on the opposition side in Angola and Afghanistan. The model assigns substantial probability mass to *both* sides but defaults toward government-biased intervention because that is the modal US posture across the full sample. This directional ambiguity for superpowers is one reason the country-year aggregation is preferable to raw dyad-level predictions: summing across all potential interveners preserves the correct signal that these countries faced unusually high total intervention pressure, even where the classifier hedges on which side a specific superpower would take.

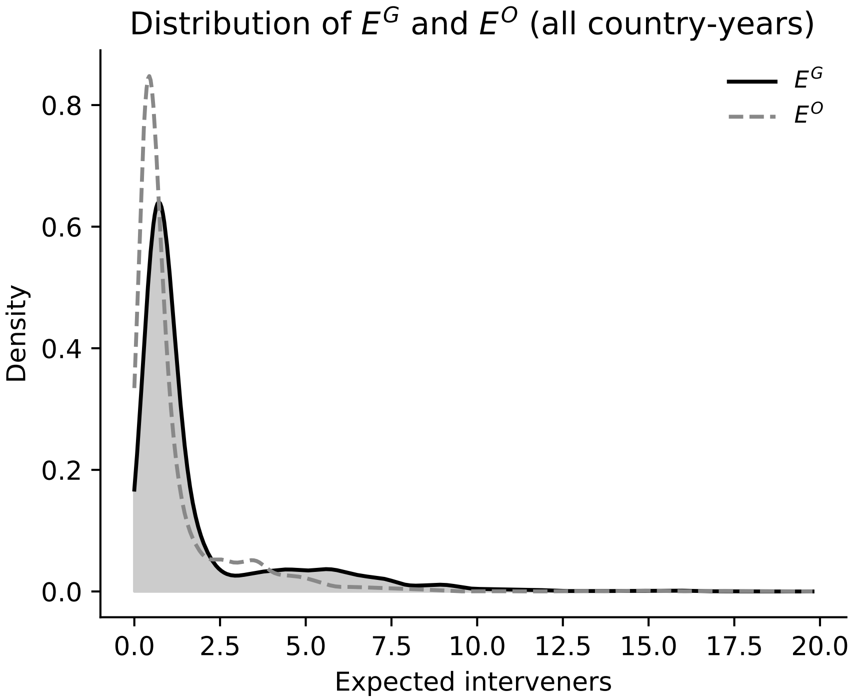

Figure 2: Kernel density estimates of E^G and E^O across all country-years (1946–2014), averaged across 25 imputation draws.

the handful of states embedded in dense alliance networks and superpower competition.

Figure 3 illustrates the temporal dynamics for three countries with well-documented intervention environments. South Vietnam shows the cleanest signal: both measures ramp from approximately 0.6 in 1960 to $E^G = 6.68$ and $E^O = 5.21$ at the peak of the war in 1972—an order of magnitude above typical countries. Afghanistan illustrates a regime-switching dynamic: the 1978 Saur Revolution raises E^G to 0.84, but by 1979 E^O overtakes E^G (1.17 vs 0.85), correctly capturing the surge in mujahideen support from Pakistan, the United States, and others. Egypt provides a critical out-of-sample check: it experiences no civil war onset in our data, so the shadow measures are pure counterfactual predictions. The classifier nonetheless tracks the geopolitical environment—the 1967 Six Day War spike ($E^G = 7.67$), the post-1972 Soviet expulsion drop, the 1991 Gulf War spike—evidence of genuine temporal tracking, not mere onset-fitting.

E^G correlates positively with both existing measures of anticipated intervention: the Lake security hierarchy used by Cunningham (2016) ($r = 0.29$, $n = 161$) and the structural P5 intervention probabilities estimated by Gibilisco and Monteiro (2022) ($r = 0.53$, $n = 145$). The stronger correlation with G&M is expected—both capture intervention propensity directly—while the moderate hierarchy correlation reflects the fact that US dominance is only one component of the broader intervention environment (Appendix F, Figure 6).

3 The Decision to Fight

Measure in hand, I now study whether intervention expectations affect the decision to initiate civil war. The *Entrants* model augments a standard onset logit with the two shadow variables constructed in the previous section:

$$U_O(F; \hat{\sigma}) = \mathbf{X}\boldsymbol{\beta} + \gamma^O \sum_i \hat{\sigma}_i(O) + \gamma^G \sum_i \hat{\sigma}_i(G) + \epsilon,$$

where $\mathbf{X}$ contains the baseline country-year predictors and the two summations are the expected total number of opposition-biased and government-biased interveners, respectively.[6] This is the simplest specification of Equation (1), in which all potential interveners carry the same utility weight ($u^i_O(O) = \gamma^O$ and $u^i_O(G) = \gamma^G$ for all i). I relax this assumption in the heterogeneous-utilities extensions below.

The two-stage design extends the logic of statistical backwards induction (Bas et al., 2008) to an n-player environment. In the two-player case, backwards induction estimates the second mover's choice probabilities and plugs them into the first mover's expected utility. Here the intervention subgame involves up to 193 simultaneous movers; separability and the Nash fixed-point reduce the problem to one in which each intervener's marginal probabilities

[6] In estimation, the summations enter through the inverse hyperbolic sine transform, $\text{asinh}(\sum_i \hat{\sigma}_i)$, for variance stabilization. This introduces a concavity—diminishing marginal returns to additional expected interveners—that is not implied by the theory. Because asinh is strictly increasing, the sign of the estimated coefficients is invariant to this choice, and the out-of-sample model comparisons test whether the shadow variables improve predictions regardless of functional form. The exact shape of the utility function over aggregated intervention expectations is not identified from these data.

suffice, and Stage 2 recovers the opposition's utility parameters from their aggregate effect on onset.

3.1 Empirical Strategy

Testing these models requires a comparison point. Following Clarke (2007), the relevant question is not whether intervention expectations matter in isolation but whether a model that includes them outperforms one that does not.

The *Baseline* model serves as the rival. It is a logistic regression for civil war onset using the covariate set from Fearon and Laitin (2003) Table 1, Model 1—the structural specification the literature has converged on as a reference: prior war, log per capita income, log population, mountainous terrain, oil exporter status, new state status, political instability, ethnic and religious fractionalization, and democracy—estimated on our sample (COW Intra-State Wars v5.1, 1946–2014). This is not a replication of F&L; it is their specification used as a comparison point on our data with updated sources. Time-varying covariates (income, population, democracy, political instability) are lagged one year to avoid post-onset contamination; income and population come from V-Dem v15 (V-Dem Institute, 2025) and democracy from the Polity project. Civil war history is drawn from COW v5.1. Slowly changing or fixed characteristics—terrain, ethnic and religious fractionalization, colonial history—are taken from the F&L replication dataset. To account for the mechanical growth of the summation-based shadow variables as the state system expands, I add a year trend to all specifications.

Cederman et al. (2010) challenge the aggregate-fractionalization baseline on the grounds that ethnopolitical exclusion from state power is the theoretically appropriate predictor. The Coethnics extended specification (Section 3) partially addresses this critique by interacting intervention expectations with ethnic matching between the intervener and the host country's population.

Model comparison draws on both in-sample fit and out-of-sample predictive performance. Within sample, I report the proportional reduction in log-loss relative to random guessing and the area under the ROC curve. Out-of-sample performance uses leave-one-war-out cross-validation: all country-years belonging to a civil war are held out together, and predictions are formed from the remaining data. This scheme respects temporal dependence and avoids the information leakage that would arise from splitting individual country-years at random.

Because the Stage 1 probabilities are estimated quantities, I retain all 25 multiply-imputed shadow measures (5 country-year imputations × 5 directed-dyad imputations) for the uncertainty analysis described below. Point estimates for model-fit metrics are computed on the averaged shadow; coefficient standard errors account for the full spread of measurement draws.

3.2 Intervention Expectations and Onset

Table 4 presents the paper's central result. The two shadow variables enter with the signs the theory requires: expected government-biased intervention is negative ($\gamma^G = -1.62$), consistent with deterrence, and expected opposition-biased intervention is positive ($\gamma^O = +0.64$), consistent with emboldening. This directional separation holds in every one of the 25 measurement draws, in all ten specifications reported in Table 5, and in the country fixed-effects estimator.

The Entrants specification improves on the Baseline both in-sample (PRL from 10.1% to 11.9%, AUC from 0.771 to 0.787) and out-of-sample (OOS PRL = +0.4%, OOS AUC from 0.650 to 0.672). For a rare-event model with a 2% base rate, positive OOS PRL is a demanding bar: the log-loss metric penalizes confident false positives severely, and the Baseline specification—which encodes decades of accumulated knowledge about structural onset risk—does not clear it.

Generated-Regressor Uncertainty

Because the shadow variables are predicted from a Stage 1 classifier rather than directly observed, standard logistic-regression errors understate uncertainty. Knox et al. (2022) show that this is the most common methodological failure in learned-proxy applications: of the 48 studies they review, nearly all ignore measurement-stage variability. I follow their recommended $T \times P$ correction: the primary analysis runs separately for each of the $T = 25$ shadow measures, each is bootstrapped $P = 200$ times (pairs-cluster, resampling countries), and the resulting 5,000 coefficient vectors are pooled.

The corrected standard errors in Table 4 are roughly 3.5× larger for $\hat{E}^G$ and 2.9× larger for $\hat{E}^O$ than naive MLE output. A variance decomposition shows that 88–92% of this total variance originates in the measurement stage—variation *between* shadow draws—with only 8–12% from primary-analysis sampling. The measurement model, not the regression, is the binding source of uncertainty.

Two features of the data make this unsurprising. First, 184 onsets across 25 measurement draws leave the Stage 2 signal thin relative to variation in the learned proxy. Second, $\hat{E}^G$ and $\hat{E}^O$ are correlated at $r = 0.94$: countries that attract government-biased intervention expectations also attract opposition-biased expectations. This collinearity means that small perturbations in the Stage 1 classifier move both shadow variables together, inflating the between-draw variance of each coefficient. Tested jointly, however, the shadow variables are unambiguous: a likelihood-ratio test rejects the Baseline in favor of Entrants at $p < 0.001$ (LR $= 32.7$, df $= 2$), and most type-disaggregated specifications reject Entrants in turn (Appendix G, Table 12). The collinearity inflates *individual* coefficient uncertainty but does not prevent the pair from carrying strong joint signal. The consistent sign separation across all 25 draws—$\gamma^G < 0$ and $\gamma^O > 0$ in every one—reinforces this: the logit reliably separates the two effects despite having only ≈6% independent variation to work with.

Neither coefficient achieves conventional significance individually: the 95% confidence interval for γ^G runs from -4.63 to $+0.69$ and for γ^O from -1.94 to $+2.59$. Under the average monotonicity condition of Knox et al. (2022) (§3.1)—more of the true intervention expectation implies more of the proxy, which the Stage 1 calibration diagrams verify empirically (Appendix E)—the signs of the coefficients are valid even if the magnitudes are attenuated. The evidence for the directional claim therefore rests not on any single p-value but on the conjunction of three facts: both coefficients carry their theoretically predicted signs across all specifications and measurement draws; the shadow variables improve out-of-sample prediction; and the pattern survives country fixed effects.

Restricting Stage 2 to the labeled subsample (country-years with observed intervention coding) yields coefficients close to the full-sample estimates, confirming that the results are not driven by out-of-sample extrapolation of the classifier (Knox et al. 2022, §5.6).

The fixed-effects specification (Column 3) absorbs all time-invariant confounders—terrain, ethnic composition, colonial history—and identifies the shadow effect from within-country temporal variation alone. Both coefficients retain their signs: $\gamma^G = -1.74$ (SE $= 0.47$), $\gamma^O = +0.87$ (SE $= 0.47$). Only 71 of 171 countries exhibit variation in onset and therefore contribute to identification. (These standard errors are from maximum likelihood and do not incorporate the $T \times P$ correction; they should be interpreted as lower bounds on uncertainty.)

3.3 Heterogeneous Utilities

The learned-proxy approach permits a question that structural designs cannot easily address: which types of interveners carry disproportionate weight in the onset calculus? The Entrants model constrains all interveners to carry the same utility weight regardless of type. The theoretical framework permits the utility parameters to vary with intervener characteristics: setting $u^i_O(O) = \gamma^O_0 + \gamma^O_P z_i$ and $u^i_O(G) = \gamma^G_0 + \gamma^G_P z_i$ for some intervener attribute z_i and aggregating produces the Entrants variables plus a pair of type-weighted interaction terms. I apply this to major-power status (*Powers*), geographic contiguity (*Neighbors*), shared dominant ethnic group (*Coethnics*), colonial relationship (*Rulers*), bilateral rivalry (*Rivals*), and military balance (*DOE*). Each type-disaggregated specification nests the Entrants variables.

Six of the nine shadow-augmented specifications outperform the Baseline out-of-sample, with Powers, Neighbors, Rivals (bin), and Rivals (cts) achieving the largest gains (OOS PRL $\approx$ 1.7–2.6%, OOS AUC $\approx$ 0.71), indicating that major powers, contiguous states, and hostile dyads carry disproportionate weight in shaping onset expectations. The Full model achieves the highest discriminative ability (OOS AUC = 0.778) and is the only specification with OOS PRL above 4%, though its 26-parameter specification risks overfitting.

The shadow measure also subsumes the two most direct existing proxies for anticipated intervention. On the common sample (6,973 country-years, 1950–2005), the Lake security hierarchy score from Cunningham (2016) adds no explanatory power beyond the shadow variables: the hierarchy coefficient is effectively zero when the Entrants variables are included (Appendix F, Figure 6). On the 7,772 country-years where the Gibilisco and Monteiro (2022) structural P5 intervention probabilities are available, the pattern is the same: the G&M aggregate score adds nothing to the Entrants specification (LR = 0.06, p = 0.80), but the shadow variables remain highly significant when added to a model that already includes the G&M score (LR = 26.4, $p < 0.001$). The learned proxy absorbs what both the hierarchy and the structural-game estimates capture, while adding variation from non-P5 interveners, bilateral relationships, and the spatial environment that neither alternative covers.

4 Conclusion

Civil war does not begin in a vacuum. Opposition groups weigh the prospect of outside military assistance against the threat of international resistance before they decide to fight. I have argued that these anticipatory calculations—the shadow of intervention—are a systematic cause of civil war onset, and I have developed a theory, a measure, and an empirical test to support this claim.

The theoretical framework models the opposition group's onset decision as a function of rational expectations over the intervention choices of up to n potential outside powers. Under two simplifying assumptions, the expected utility of fighting reduces to a weighted sum of intervener-specific probabilities and utilities—a quantity that is both theoretically interpretable and empirically tractable. The framework generalizes prior work by addressing both government- and opposition-biased intervention, by modeling the full set of potential state interveners rather than superpowers alone, and by allowing for strategic interdependence among interveners through a spatial weighting scheme.

Operationalizing this framework required constructing a measure of intervention expectations that did not yet exist. The Stage 1 classifier draws on a rich directed-dyad feature set—state capabilities, regime type, bilateral relationships, foreign policy alignment, host-country characteristics, and spatial context—to produce calibrated probability estimates for each potential intervener in each civil war onset year. Multiple imputation across 25 complete datasets propagates uncertainty from both the country-year and undirected-dyad imputation stages through to the final model estimates.

The onset analysis delivers three principal findings. First, the two shadow variables are jointly significant ($p < 0.001$) and enter with the signs the theory requires. Expected government-biased intervention enters negatively ($\gamma^G = -1.62$): the prospect of outside support for the incumbent deters onset. Expected opposition-biased intervention enters positively ($\gamma^O = +0.64$): the prospect of outside support for rebels encourages fighting. The two variables are correlated at $r = 0.94$—countries that attract one type of intervention expectation attract both—so measurement-corrected individual confidence intervals are wide, but the directional separation holds in every one of the 25 measurement draws, all ten specifications, and the country fixed-effects estimator. The shadow subsumes both the Lake security hierarchy (Cunningham, 2016) and the Gibilisco and Monteiro (2022) structural P5 estimates: neither adds explanatory power once the shadow variables are included. Second, the shadow variables carry genuine predictive content that survives out-of-sample validation. Eight of the ten specifications achieve positive out-of-sample proportional reduction in log-loss under leave-one-onset-group-out cross-validation, with the strongest gains from type-disaggregated specifications—Powers, Neighbors, Rivals (bin), and Rivals (cts) all exceed the aggregate Entrants model—and out-of-sample AUC rises from 0.650 (Baseline) to 0.712 (Rivals (cts)); the in-sample PRL rises from 10.1% to 11.9% and AUC from 0.771 to 0.787.

The directional separation is not merely an empirical regularity of this particular analysis; it is grounded in a complementary literature on civil war outcomes. Thyne (2006) provides direct evidence that both governments and opposition groups incorporate expectations of external support into their prewar decisions, with cheap interstate signals—day-to-day diplomatic events that shift perceived intervention probabilities—substantially affecting onset. Sullivan and Karreth (2014) show that pro-rebel intervention raises the probability of rebel victory by roughly 40 percentage points unconditionally, while pro-government intervention improves government prospects only when rebels are militarily strong. Dudley (2024) finds a parallel asymmetry in negotiations: rebel-biased intervention increases the probability of talks by approximately 60 percentage points, while government-biased intervention reduces it. Forward-looking rebels who anticipate these asymmetric downstream effects should respond very differently to E^G and E^O, making the directional distinction theoretically necessary rather than merely a descriptive refinement.

The results carry implications beyond the civil war literature. Many political decisions are made in the shadow of anticipated multiparty responses: party platforms are staked out in anticipation of rivals' reactions; bilateral agreements are struck with an eye to market responses; leaders calibrate repression partly on expectations about international condemnation or support. The two-stage empirical strategy introduced here—predict the probabilities of downstream actions, then use those predictions as regressors in a structural model of the initiating decision—is applicable wherever a strategic choice is made in anticipation of others' responses. Within the civil war literature, the approach addresses a persistent limitation of the macro-correlates framework: country-level structural risk factors evolve slowly and cannot explain the timing of onset within already-risky countries. Intervention expectations shift with the decisions of specific leaders and governments—year to year, conflict to conflict—providing the temporal variation that structural factors cannot.

The paper's two-stage design also speaks to a productive debate about the appropriate role of machine learning in quantitative political science. Grimmer et al. (2021) distinguish four tasks—discovery, measurement, causal inference, and prediction—each demanding different methods and evaluation criteria; the cardinal sin is conflating prediction with inference. The present design instantiates their prescription: Stage 1 is measurement (agnostic ensemble, evaluated by PRL and AUC), Stage 2 is causal inference (parametric model, identification assumptions, corrected standard errors). Morucci and Spirling (2024) sharpen the point further: in curated political science data, complex models rarely outperform logistic regression because researchers select variables guided by theory, stripping out the nonlinear structure ML exploits. ML earns its keep in *measurement* tasks where the feature space is broad and uncurated—exactly the Stage 1 setting, where the 12-percentage-point PRL gap between the ensemble and the best single logistic specification confirms exploitable nonlinear structure. Stage 2, by contrast, uses the same five to ten covariates the onset literature has relied on for decades; logistic regression suffices because the feature set was designed for it. Montgomery and Olivella (2018) advocate tree ensembles for prediction when the data-generating process is unknown; their propensity-score illustration is structurally analogous to the present two-stage design, with flexible ML first-stage weights feeding a parametric second-stage model. Knox et al. (2022) close the loop: once the proxy is learned, the remaining challenge is to propagate measurement-stage uncertainty into inference and to verify that the proxy's imperfections do not distort the directional conclusions. Under their "average monotonicity" condition (§3.1)—more of the true concept implies more of the proxy, which the Stage 1 calibration verifies empirically—the signs of $\gamma^G < 0$ and $\gamma^O > 0$ are valid even if the magnitudes are attenuated. Whether anticipated intervention shapes the decision to fight is a substantive question; that it can be measured—at scale, across time, and across the full population of potential outside powers—is the prior claim on which the answer depends.

References

Aguirregabiria, V. and P. Mira (2007). Sequential estimation of dynamic discrete games. *Econometrica 75*(1), 1–53.

Aydin, A. and P. M. Regan (2012). Networks of third-party interveners and civil war duration. *European Journal of International Relations 18*(3), 573–597.

Bailey, M. A., A. Strezhnev, and E. Voeten (2017). Estimating dynamic state preferences from United Nations voting data. *Journal of Conflict Resolution 61*(2), 430–456. Data updated June 2024. https://doi.org/10.7910/DVN/LEJUQZ.

Bas, M. A., C. S. Signorino, and R. W. Walker (2008). Statistical backwards induction: A simple method for estimating recursive strategic models. *Political Analysis 16*(1), 21–40.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Burbidge, J. B., L. Magee, and A. L. Robb (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association 83*(401), 123–127.

Carroll, R. J. and B. Kenkel (2019). Prediction, proxies, and power. *American Journal of Political Science 63*(3), 565–581.

Cederman, L.-E., A. Wimmer, and B. Min (2010). Why do ethnic groups rebel? New data and analysis. *World Politics 62*(1), 87–119. EPR Core Dataset 2021 used in this paper.

Cetinyan, R. (2002). Ethnic bargaining in the shadow of third-party intervention. *International Organization 56*(3), 645–677.

Clarke, K. A. (2007). The necessity of being comparative: Theory confirmation in quantitative political science. *Comparative Political Studies 40*(7), 886–908.

Crisman-Cox, C. and M. Gibilisco (2021). Estimating signaling games in international relations: Problems and solutions. *Political Science Research and Methods 9*(3), 565–582.

Cunningham, D. E. (2016). Preventing civil war: How the potential for international intervention can deter conflict onset. *World Politics 68*(2), 307–340.

Diehl, P. F., G. Goertz, and Y. Gallegos (2021). Peace data: Concept, measurement, patterns, and research agenda. *Conflict Management and Peace Science 38*(5), 605–624. Peace Data v2.01.

Dixon, J. S. and M. R. Sarkees (2016). *A Guide to Intra-State Wars: An Examination of Civil, Regional, and Intercommunal Wars, 1816–2014*. SAGE Publications.

Dudley, R. (2024). Turning the tables: Military intervention and the onset of negotiations in civil war. *Journal of Conflict Resolution 68*(6), 1139–1167.

Fearon, J. D. and D. D. Laitin (2003). Ethnicity, insurgency, and civil war. *American Political Science Review 97*(1), 75–90.

Findley, M. G. and T. K. Teo (2006). Rethinking third-party interventions into civil wars: An actor-centric approach. *Journal of Politics 68*(4), 828–837.

Gent, S. E. (2007). Strange bedfellows: The strategic dynamics of major power military interventions. *Journal of Politics 69*(4), 1089–1102.

Gent, S. E. (2008). Going in when it counts: Military intervention and the outcome of civil conflicts. *International Studies Quarterly 52*(4), 713–735.

Gibilisco, M. and S. Monteiro (2022). Do major-power interventions encourage the onset of civil conflict? A structural analysis. *Journal of Politics 84*(2), 940–956.

Gibler, D. M. (2009). *International Military Alliances, 1648–2008*. CQ Press. COW Formal Alliances v4.1.

Grimmer, J., M. E. Roberts, and B. M. Stewart (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science 24*, 395–419.

Huth, P. K. and T. L. Allee (2003). Domestic political accountability and the escalation and settlement of international disputes. *Journal of Conflict Resolution 47*(3), 320–352. ICOW Territorial Claims Project v10.1.

Knox, D., C. Lucas, and W. K. T. Cho (2022). Testing causal theories with learned proxies. *Annual Review of Political Science 25*, 419–441.

Kuperman, A. J. (2008). The moral hazard of humanitarian intervention: Lessons from the balkans. *International Studies Quarterly 52*(1), 49–80.

Kydd, A. H. and S. Straus (2013). The road to hell? Third-Party intervention to prevent atrocities. *American Journal of Political Science 57*(3), 673–684.

Langø, H.-I. (2023). Intervention, war expansion, and the international sources of civil war. *Conflict Management and Peace Science 40*(3), 304–324.

Montgomery, J. M. and S. Olivella (2018). Tree-based models for political science data. *American Journal of Political Science 62*(3), 729–744.

Morucci, M. and A. Spirling (2024). Model complexity for supervised learning: Why simple models almost always work best. Working paper.

Muchlinski, D., D. Siroky, J. He, and M. Kocher (2016). Comparing random forests with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis 24*(1), 1–17.

Muth, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica 29*(3), 315–335.

Pevehouse, J. C. W., T. Nordstrom, R. W. McManus, and A. S. Jamison (2020). Tracking organizations in the world: The Correlates of War IGO version 3.0 datasets. *Journal of Peace Research 57*(3), 492–503.

Regan, P. M. (2000). *Civil Wars and Foreign Powers: Outside Intervention in Intrastate Conflict.* University of Michigan Press.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Salehyan, I. (2007). Transnational rebels: Neighboring states as sanctuary for rebel groups. *World Politics 59*(2), 217–242.

Salehyan, I., K. S. Gleditsch, and D. E. Cunningham (2011). Explaining external support for insurgent groups. *International Organization 65*(4), 709–744.

Sambanis, N., S. Skaperdas, and W. Wohlforth (2020). External intervention, identity, and civil war. *Comparative Political Studies 53*(14), 2155–2182.

Signorino, C. S. (2002). Strategy and selection in international relations. *International Interactions 28*(1), 93–115.

Singer, J. D. (1987). Reconstructing the Correlates of War dataset on material capabilities of states, 1816–1985. *International Interactions 14*, 115–132. NMC v6 (2021 update) used in this paper.

Stinnett, D. M., J. Tir, P. Schafer, P. F. Diehl, and C. Gochman (2002). The Correlates of War project direct contiguity data, version 3. *Conflict Management and Peace Science 19*(2), 58–66.

Sullivan, P. L. and J. Karreth (2014). The conditional impact of military intervention on internal armed conflict outcomes. *Conflict Management and Peace Science 31*(4), 363–384.

Thyne, C. L. (2006). Cheap signals with costly consequences: The effect of interstate relations on civil war. *Journal of Conflict Resolution 50*(6), 937–961.

V-Dem Institute (2025). V-Dem [country–year/country–date] dataset v15. Varieties of Democracy (V-Dem) Project. https://www.v-dem.net.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology 6*(1), Article 25.

Wang, Y. (2019). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data: A comment. *Political Analysis 27*(1), 107–110.

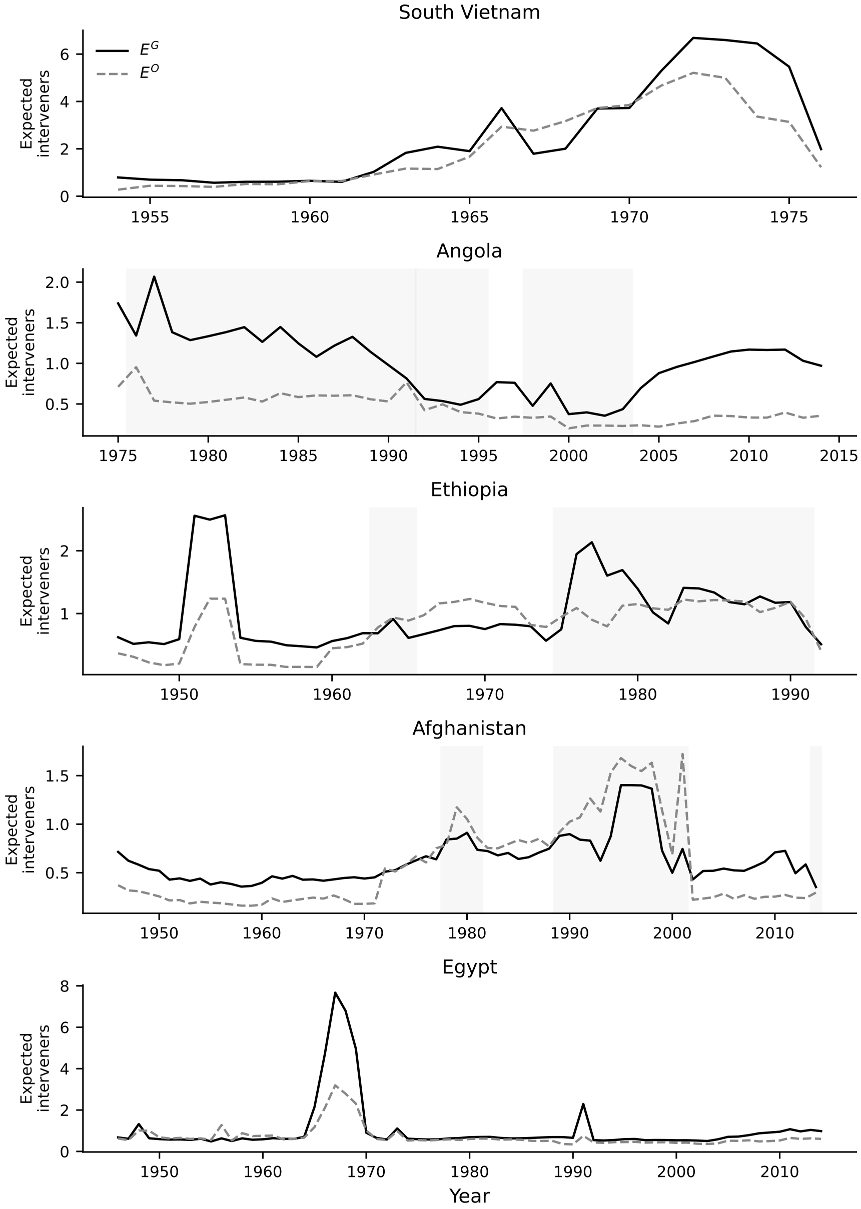

Figure 3: Shadow measure time series (E^G and E^O) for five countries with well-documented intervention environments. Probabilities averaged across 25 imputation draws.

	Baseline	Entrants	FE Entrants
Polity2$_{\text{lag}}$	-0.026 (0.014)	-0.032^{*} (0.014)	-0.043^{*} (0.019)
GDP/cap (log)$_{\text{lag}}$	-0.670^{***} (0.112)	-0.500^{***} (0.122)	-0.232 (0.270)
Pop (log)$_{\text{lag}}$	$+0.110$ (0.061)	$+0.095$ (0.066)	$+0.411$ (0.697)
Mountainous	$+0.223^{***}$ (0.063)	$+0.126$ (0.067)	
Noncontiguous	$+0.532^{*}$ (0.225)	$+0.538^{*}$ (0.225)	
Oil	$+0.450^{*}$ (0.203)	$+0.328$ (0.208)	
New state	$+1.003^{**}$ (0.342)	$+1.168^{**}$ (0.348)	$+1.218^{**}$ (0.385)
Instability$_{\text{lag}}$	$+0.567^{**}$ (0.180)	$+0.513^{**}$ (0.179)	$+0.617^{**}$ (0.196)
Prior war	$+0.529^{**}$ (0.195)	$+0.450^{*}$ (0.200)	-0.313 (0.203)
Ethnic frac.	$+0.409$ (0.293)	$+0.562$ (0.299)	
Religious frac.	$+0.673$ (0.392)	$+1.225^{**}$ (0.411)	
Year	$+0.011^{*}$ (0.005)	$+0.006$ (0.005)	-0.003 (0.019)
$\hat{E}^{G}$ (asinh)		-1.62 (1.37)	-1.74^{***} (0.47)
$\hat{E}^{O}$ (asinh)		$+0.64$ (1.10)	$+0.87$ (0.47)
Country FEs	No	No	Yes
N	8,792	8,792	3,750
Onsets	184	184	184
PRL	10.1%	11.9%	
AUC	0.771	0.787	

Table 4: Logistic regression estimates for civil war onset. Columns 1–2 are pooled logit; Column 3 is conditional (fixed-effects) logit. Point estimates and standard errors for the shadow variables ($\hat{E}^{G}$, $\hat{E}^{O}$) in Column 2 are from a $T \times P$ bootstrap following Knox et al. (2022): $T = 25$ measurement-model draws $\times$ $P = 200$ pairs-cluster bootstrap replications, pooling all 5,000 coefficient vectors. This accounts for uncertainty in both the Stage 1 measurement model and the Stage 2 primary analysis. All other standard errors are from maximum likelihood. Column 3 standard errors do not incorporate the $T \times P$ correction. Time-invariant covariates are absorbed by country fixed effects in Column 3. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Model	IS PRL	IS AUC	OOS PRL	OOS AUC
Baseline	10.1%	0.771	−1.0%	0.650
Entrants	11.9%	0.787	+0.4%	0.672
Powers	14.3%	0.809	+1.7%	0.707
Neighbors	14.5%	0.814	+2.4%	0.710
Coethnics	12.8%	0.798	+0.4%	0.680
Rulers	11.9%	0.787	−0.8%	0.668
Rivals (bin)	14.4%	0.814	+1.7%	0.706
Rivals (cts)	14.7%	0.820	+2.6%	0.712
DOE	11.9%	0.788	+0.2%	0.671
Full	22.3%	0.866	+4.2%	0.778

Table 5: In-sample and out-of-sample fit across model specifications. Each extended specification nests the aggregate Entrants variables. IS = in-sample; OOS = leave-one-onset-group-out cross-validation. PRL = proportional reduction in log-loss relative to the class-frequency null. AUC = area under the ROC curve.

Conflict	**State**	**Correction**	**Rationale**
Congo-Brazzaville (1997)	France (220)	target: neutral → 2 (opp)	France supplied arms to Cobra militia, siding with Angola/opposition
Indonesia (1958)	USA (002)	target: gov → 2 (opp)	CIA Operation Archipelago backed anti-communist rebels
Cameroon (conflict 17)	W. Germany	ccode: 255 → 260	Miscoded; correct COW code for West Germany
Congo Crisis (1960s)	All	host ccode: 484 → 490	DRC better represents this conflict than Republic of Congo
Congo Crisis (1964)	Belgium (211)	target flipped gov↔opp	Sides effectively reversed in 1964 phase
Ethiopia (1980s)	Somalia (520)	target: gov → 2 (opp)	Somalia generally supported Ogaden opposition
Cambodia	Vietnam (816)	target: opp → 1 (gov)	Vietnam supported incumbent government throughout
Mozambique	Zimbabwe (552)	target: → 1 (gov)	Zimbabwe consistently backed FRELIMO government
Mozambique	Malawi (553)	target: → 3 (neutral)	Malawi supported both sides; excluded as neutral
Zimbabwe/South Africa	South Africa	target: → 3 (neutral)	Supported both sides; excluded
Sri Lanka	UK (200)	target: → 2 (opp)	Miscoded; correct side assignment
Sri Lanka	India (750)	target: → 1 (gov)	India supported government
Iran	Iraq (645)	target: gov → 1	Iraq favored government; miscoded in original

Table 6: Regan coding corrections applied before analysis.

Online Appendix

A Regan Intervention Coding Corrections

The raw Regan (2000) data contain ten per-conflict coding decisions that are corrected before analysis. These are listed in Table 6. In all cases the corrections are applied to the raw Stata file *before* country-code standardization, matching the original R pipeline's order of operations.

After corrections, two directed dyads with conflicting target assignments across multiple records are resolved by fixing the target directly: Georgia–Russia (opposition-biased) and Liberia–Sierra Leone (government-biased). Final deduplication retains the first record per directed dyad-year after sorting by the `ddyear` identifier.

A.1 Post-1999 Intervention Coding

The Regan (2000) data end in 1999. To extend the Stage 1 training set through 2014, I code military interventions in all 38 post-1999 COW onset country-years using the same threshold: deployment of military personnel (troops, combat advisors, or directed proxy forces) attributable to a specific state, on behalf of one side. Arms transfers, economic aid, sanctuary, and strategic direction alone are insufficient. UN peacekeeping missions are excluded. Table 7 lists the 32 coded interventions; the remaining 17 onset country-years are coded as having no intervention meeting the threshold. Full source notes for each coding decision are available in the replication archive.

Onset country-years coded as having no military intervention meeting the personnel-deployment threshold: Guinea (2000), Philippines (2000, 2003, 2005), Burundi (2001), Rwanda (2001), Nepal (2001, 2003), Sudan (2003, 2011), Indonesia (2003), Yemen (2004, 2007), Sri Lanka (2006), Syria (2011), Myanmar (2011), and Central African Republic (2012).

B Multiple Imputation Procedure

Missing values in continuous predictors are handled through a two-stage multiple imputation strategy using miceforest (Multiple Imputation by Chained Equations with LightGBM). The strategy generates $5 \times 5 = 25$ complete datasets; estimates are combined across datasets following Rubin (1987).

The staged structure reflects the data's own hierarchy and is a design choice, not a computational convenience. Country-year quantities (GDP per capita, regime scores, CINC components) must be imputed at the country-year level: imputing them at the directed-dyad level would treat state A's GDP as a separate quantity in each of A's dyads, producing inconsistent imputed values across rows that share the same country and year. Dyadic quantities (bilateral trade, peace scores, capital distances) must be imputed before directed expansion so that the symmetry constraint $x_{AB} = x_{BA}$ holds within

Host	**Year**	**Intervener**	**Side**
Sierra Leone	2000	United Kingdom	Gov.
Ivory Coast	2002	France	Gov.
Liberia	2002	Guinea	Opp.
Pakistan	2004	United States	Gov.
Chad	2005	France	Gov.
Chad	2005	Sudan	Opp.
Pakistan	2005	United States	Gov.
Somalia	2006	Ethiopia	Gov.
Chad	2007	France	Gov.
Chad	2007	Sudan	Opp.
Pakistan	2007	United States	Gov.
Somalia	2009	Ethiopia	Gov.
Iraq	2010	United States	Gov.
Ivory Coast	2011	France	Opp.
Libya	2011	United States	Opp.
Libya	2011	United Kingdom	Opp.
Libya	2011	France	Opp.
Libya	2011	Qatar	Opp.
Libya	2011	UAE	Opp.
Mali	2012	France	Gov.
Mali	2012	Chad	Gov.
Nigeria	2013	Chad	Gov.
South Sudan	2013	Uganda	Gov.
Ukraine	2014	Russia	Opp.
Somalia	2014	Ethiopia	Gov.
Somalia	2014	Uganda	Gov.
Libya	2014	Egypt	Gov.
Libya	2014	UAE	Gov.
Yemen	2014	Saudi Arabia	Gov.
Yemen	2014	UAE	Gov.
Afghanistan	2014	United States	Gov.
Afghanistan	2014	United Kingdom	Gov.

Table 7: Post-1999 military interventions in COW onset country-years. Gov. = government-biased; Opp. = opposition-biased.

each imputed dataset—a constraint that would be broken if each directed-dyad row were imputed independently. Imputing in sequence, country-year first and undirected dyad second, preserves both requirements simultaneously.

Stage 1—Country-year imputation. The analysis sample (1946–2014, $n = 8{,}897$ country-years) is imputed five times. Variables imputed include regime scores, GDP per capita, population, oil wealth, mountainous terrain, ethnic fractionalization, ethnic exclusion, all six NMC capability components, and UN ideal points. Six right-skewed NMC variables are asinh-transformed before imputation (with individually optimised scale parameters minimising the Kolmogorov–Smirnov statistic against the standard normal) and back-transformed afterwards. Five iterations of chained equations are run. Post-imputation clipping enforces substantive constraints ($-10 \leq$ `polity2` ≤ 10, probabilities $\in [0, 1]$, capabilities ≥ 0).

Stage 2—Undirected-dyad imputation. For each of the five country-year datasets, the undirected dyad frame ($\sim$599,000 rows) is imputed five times before expansion to directed dyads. Imputation at the undirected-dyad stage ensures that symmetric quantities (trade flows, peace scores, capital distance) satisfy $x_{AB} = x_{BA}$ by construction within each imputed dataset. Variables imputed include bilateral trade flows, peace quality scores, capital distances, and dispute outcome expectations. Two iterations of chained equations are run.

The 25 resulting complete datasets are processed identically through the spatial-weights and intervention-coding stages before combination.[7]

C Variable List

Table 8 lists all country-year variables included in the analysis. Table 9 lists the dyadic variables added at the directed-dyad stage.

[7] The 25 branches are mutually independent at every stage upstream of the final combination: each (cy, ud) pair flows from imputation through Stage 1 training, Stage 1 prediction, and Stage 2 estimation without reference to any other branch. Rubin's Rules averaging is the only step that requires all 25 to be complete. This means the pipeline is embarrassingly parallel—all 25 classifiers could in principle be trained simultaneously—and that updating the sample in a future study would not require retraining all 25 models from scratch. A single additional imputed dataset, processed through the same pipeline, contributes marginally to the final combination.

Variable	Description	Source	Coverage
onset	Civil war onset (COW types 4–5, $\geq$ 1,000 battle deaths)	COW ISW v5.1	1946–2014
polity2	Polity II score (-10 to $+10$)	V-Dem v15 (e_polity2)	1946–2014
v2x_polyarchy	Electoral democracy index (0–1)	V-Dem v15	1946–2014
instab	Political instability (large Polity swing or coup)	V-Dem v15	1946–2014
lgdp	Log GDP per capita (2011 PPP int'l $)	V-Dem v15 (e_gdppc)	1946–2014
lpop	Log population (thousands)	V-Dem v15 / NMC v6 backup	1946–2014
oil	Oil/gas income per capita > 0 (binary)	V-Dem v15	1946–2014
cinc	Composite Index of National Capability	COW NMC v6	1946–2014
milex	Military expenditure (thousands USD)	COW NMC v6	1946–2014
milper	Military personnel (thousands)	COW NMC v6	1946–2014
major_power	COW major power status (binary)	COW Major Powers 2024	1946–2014
is_P5	UN Security Council P5 member (binary)	Coded from COW ccodes	1946–2014
ideal_point	UNGA ideal point (liberal–conservative)	Bailey et al. (2017)	1946–2023
recent_int	Military interventions by this state in prior 5 years	Regan (2000)	1946–1999
lmtnest	Log(% mountainous terrain + 1)	Fearon and Laitin (2003)	time-invariant
ncontig	Non-contiguous state (islands, exclaves)	Fearon and Laitin (2003)	time-invariant
colbrit	Former British colony	Fearon and Laitin (2003)	time-invariant
colfra	Former French colony	Fearon and Laitin (2003)	time-invariant
ethfrac	Ethnic fractionalization index (0–1)	Fearon and Laitin (2003)	time-invariant
eth_excl_frac	Population share of excluded ethnic groups	Cederman et al. (2010)	1946–2021
nwstate	New state (independent $\leq$ 2 years)	COW States 2011	1946–2014
prior_war	Ongoing civil war in previous year	COW ISW v5.1	1946–2014

Table 8: Country-year variables.

Variable	Description	Source	Coverage
ud_defense	Defense alliance (binary)	COW Alliances v4.1 (Gibler, 2009)	1946–2012
ud_entente	Entente alliance (binary)	COW Alliances v4.1	1946–2012
ud_A_biImports	Host's imports from intervener (millions USD)	COW Trade v4.0	1946–2014
ud_log_capdist	Log capital distance (km)	COW capdist	static
ud_conttype	Contiguity type (1=land, . . . , 6=none)	COW Contiguity v3.2 (Stinnett et al., 2002)	1946–2016
ud_peace	Peace quality score (0–1)	Diehl et al. (2021)	1946–2015
ud_icow_nclaims	Number of active territorial claims	ICOW v10.1 (Huth and Allee, 2003)	1946–2001
igo_shared	Shared full IGO memberships (count)	COW IGO v3 (Pevehouse et al., 2020)	1946–2014
ideal_point_distance	Absolute UNGA ideal-point difference	Derived from ideal_point	1946–2023
doe_pr_win_A	P(A wins bilateral dispute)	Carroll and Kenkel (2019)	1946–2012
A_wasColOf_B	A was a colony of B (binary)	COW coldata	static
B_wasColOf_A	B was a colony of A (binary)	COW coldata	static
ud_sharedColonizer	Shared former colonizer (binary)	COW coldata	static
ud_sameFirstEth	Same dominant ethnic group (binary)	Ellingsen (2000)	1945–2002
log_capratio	Log(cinc$_B$ / cinc$_A$) capability ratio	Derived from NMC v6	1946–2014
spat_gov	Polity-weighted fraction of other interveners: gov-biased	Derived (see App. D)	1946–1999
spat_opp	Polity-weighted fraction: opp-biased	Derived (see App. D)	1946–1999
spat_US_G	USA is gov-biased in this conflict (binary)	Derived	1946–1999
spat_USSR_G	USSR/Russia is gov-biased (binary)	Derived	1946–1999
spat_US_O	USA is opp-biased (binary)	Derived	1946–1999
spat_USSR_O	USSR/Russia is opp-biased (binary)	Derived	1946–1999
spat_US_USRG	B=USA and USSR gov-biased (binary)	Derived	1946–1999
spat_US_USRO	B=USA and USSR opp-biased (binary)	Derived	1946–1999
spat_USR_USG	B=USSR and USA gov-biased (binary)	Derived	1946–1999
spat_USR_USO	B=USSR and USA opp-biased (binary)	Derived	1946–1999

Table 9: Directed-dyad-year variables (B = potential intervener, A = host).

D Spatial Weights Derivation

For each year t in the analysis window, I construct a row-normalized polity-similarity weight matrix W_t over the set of potential intervener states. Let p_i denote the Polity II score of state i in year t. The raw weight is:

$$w_{ij} = \begin{cases} 1/|p_i - p_j| & \text{if } p_i \neq p_j \text{ and } i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

Row-normalization gives $W_{ij} = w_{ij}/(\sum_k w_{ik})$, so that each row sums to 1 (rows summing to zero remain zero). States with missing Polity scores are excluded from W_t in that year.

The polity-similarity weighting encodes the assumption that states pay more attention to the intervention choices of ideologically similar peers, and less to the choices of ideological adversaries. This departs from the standard inverse-distance spatial weight used in most spatial-lag models, which would give *higher* weight to more *different* states; the similarity-based weight is theoretically more appropriate for the ideological-diffusion channel.

Yemen (COW code 678) is excluded from the 1990 weight matrix to avoid anomalous results during the unification transition year.

For each onset event (host country A, year t), the spatial lag for potential intervener B is the weighted sum of other potential interveners' predicted intervention probabilities, with A removed from the weight matrix before computation:

$$\text{spat\_gov}_B = \sum_{\substack{j \neq A \\ j \neq B}} W_{Bj} \cdot \hat{\sigma}_j^*(1),$$

where $\hat{\sigma}_j^*(1)$ is the Nash equilibrium probability that state j intervenes on behalf of the government, as produced by the fixed-point iteration described in Section 2. An analogous expression gives `spat_opp` using $\hat{\sigma}_j^*(2)$.

In the first training iteration (initialisation), $\hat{\sigma}_j^*(k)$ is replaced by the observed indicator $\nVdash[y_j = k]$, where $y_j \in \{0, 1, 2\}$ is the Regan-coded intervention outcome. Subsequent iterations substitute the model's own out-of-fold predictions until convergence, at which point the lags are self-consistent with the predictions they condition on.

	PRL (%)			NNLS weight (%)		
Learner	X	W	XW	X	W	XW
MLP (100, 50)	33.4	14.5	30.9	24.9	0.0	19.5
Random forest	18.9	−29.9	16.5	22.6	0.0	16.0
HGB (lr=0.05)	27.3	16.4	25.6	7.9	0.0	3.2
HGB (lr=0.10)	−44.4	−44.6	−76.3	2.4	0.0	1.3
MLP (25)	1.3	−93.4	1.3	0.8	0.0	0.8
Multinomial	32.7	16.2	33.8	0.0	0.0	0.4
Ridge	33.1	16.3	34.0	0.0	0.0	0.1
Elastic net	33.2	16.3	34.5	0.0	0.0	0.0
Lasso	34.1	16.4	35.1	0.0	0.0	0.0
Super-learner		**47.3**			**100**	

Table 10: Out-of-fold PRL and NNLS weight for all 27 candidate models (9 learners × 3 feature sets) and the super-learner ensemble. X = dyad features only; W = spatial lags only; XW = both. All statistics average across 25 imputation draws. Learners are sorted by total weight ($X + W + XW$).

E Stage 1 Classifier Diagnostics

This appendix provides the extended diagnostic output for the Stage 1 super-learner that is summarised in the main text. All statistics are out-of-fold unless stated otherwise; all figures and tables average across the 25 complete imputation datasets (5 country-year draws × 5 dyad draws) with standard deviations in parentheses or as error bars.

E.1 Component Learner Performance and Weights

Each of the nine component classifiers is trained on three feature sets: X (dyad characteristics only, no spatial lags), W (spatial lags only), and XW (both), yielding 27 candidate models. The NNLS stacking layer assigns weights to all 27 jointly. Table 10 reports out-of-fold PRL and NNLS weights for every learner broken out by feature set. Total weight by feature set: X = 58.6% (SD 8.1%), XW = 41.3% (SD 8.1%), $W \approx 0\%$.

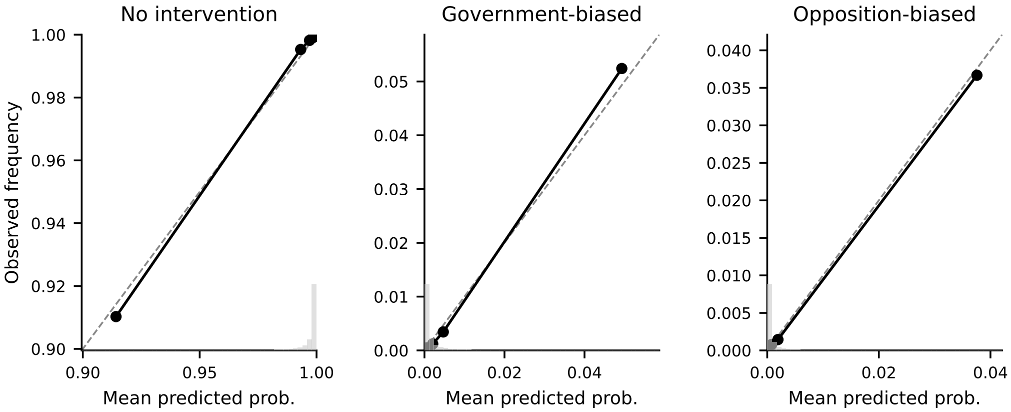

Figure 4: Reliability diagrams for each outcome class (no intervention, government-biased, opposition-biased). Each panel plots observed frequency against mean predicted probability in 10 quantile bins, pooled across all 25 imputation draws; the dashed diagonal represents perfect calibration. Axes are zoomed to the range where data reside. Shaded histograms show the distribution of predicted probabilities.

E.2 Calibration

Because Stage 1 probabilities are summed across approximately 190 dyads to form the country-year shadow measure, a calibration error of δ per dyad compounds to roughly 190δ in the aggregate—so even small miscalibration matters. Figure 4 presents reliability diagrams for each outcome class, pooled across all 25 imputation draws. In each panel, observations are sorted by predicted probability and divided into 10 quantile bins; the observed frequency within each bin is plotted against the mean prediction. The super-learner is well calibrated across all three classes: points track the diagonal closely, with no systematic over- or under-prediction.

E.3 Fixed-Point Convergence

E.3.1 The Consistency Requirement

The spatial lags `spat_gov` and `spat_opp` enter the Stage 1 classifier as predictors: they summarize how other potential interveners in the same conflict are behaving, weighted by ideological similarity. But the classifier itself produces predicted intervention probabilities, and it is those probabilities—rather than the realized 0/1/2 outcomes—that should, in principle, enter the spatial weights. The reason is that observed outcomes are single draws from mixed-strategy equilibrium probabilities; a government-biased intervention (outcome 1) carries no information about whether the intervening state was playing a pure or mixed strategy.

Formally, the requirement is *self-consistency*: the spatial lags fed into $\hat{M}$ should be the weighted sums of the probabilities that $\hat{M}$ itself assigns to each dyad. Let $\hat{\mathbf{p}} = \hat{M}(\mathbf{X}(\mathbf{s}))$ be the vector of predicted probabilities, where $\mathbf{X}(\mathbf{s})$ denotes the feature matrix with spatial lags computed from $\mathbf{s}$. Self-consistency requires $\mathbf{s} = W\hat{\mathbf{p}}$, a fixed-point condition.

E.3.2 Diagnosis of Non-Convergence

The main training loop (Section 2) runs five rounds of retrain-then-update, initializing $\mathbf{s}^{(0)}$ from the realized binary outcomes. To assess whether five rounds are sufficient, I compute the mean absolute change in spatial lags between the final iteration and the initialization:

$$\Delta = \frac{1}{n}\sum_i |s_i^{(5)} - s_i^{(0)}|.$$

Averaging across all onset rows and both lag channels, $\Delta \approx 0.007$ after five iterations. The theoretical tolerance threshold for self-consistency is 10^{-4}; five iterations leave the spatial lags roughly 70 times above threshold.

E.3.3 Frozen-Model Burnout

Retraining the full super-learner is computationally expensive (∼3 hours per draw). An alternative is *frozen-model burnout*: after training is complete, hold the fitted $\hat{M}$ fixed and iterate spatial lags using its predictions until convergence. Each burnout pass costs only a prediction sweep (milliseconds versus hours).

The burnout algorithm is:

1. Initialize from the five-iteration approximation $\mathbf{s}^{(5)}$.
2. Apply $\hat{M}$ to features $\mathbf{X}(\mathbf{s}^{(k)})$ to obtain $\hat{\mathbf{p}}^{(k)}$.
3. Update $\mathbf{s}^{(k+1)} = W\hat{\mathbf{p}}^{(k)}$.

4. If $\max_i |s_i^{(k+1)} - s_i^{(k)}| < \tau$, stop; otherwise go to step 2.

The tolerance is set at $\tau = 5 \times 10^{-4}$. This is above the theoretical 10^{-4} threshold because random-forest predictions have intrinsic stochasticity (different prediction calls can return slightly different probabilities for the same input due to tree-averaging over bootstrap samples). Empirical experiments confirm that Δ plateaus around 5×10^{-4} and cannot be reduced further without retraining.

Across all 25 imputation draws, burnout converges in 2–5 iterations (mean 3.3), with final Δ ranging from 0.9×10^{-4} to 4.7×10^{-4}—confirming that the plateau reflects irreducible prediction noise rather than failure to converge. Figure 5 shows the convergence trajectories: all draws exhibit rapid geometric decay (roughly halving each iteration) before plateauing at the tolerance threshold.

E.3.4 Interpretation: Approximate Equilibrium Selection

The burnout iteration does not find an equilibrium of the underlying intervention game; it finds a fixed point of the *fitted classifier* $\hat{M}$. The procedure is a nonparametric instance of the nested pseudo-likelihood (NPL) algorithm of Aguirregabiria and Mira (2007): where A&M iterate between estimating structural parameters and updating choice probabilities, the present application holds the fitted $\hat{M}$ fixed and iterates only the spatial lags (choice probabilities), corresponding to A&M's two-step PML with a nonparametric first stage. These coincide with the true equilibrium only if $\hat{M}$ exactly recovers the true equilibrium mapping, which no finite-sample estimator can guarantee.

The weaker, defensible claim is as follows. Realized outcomes are draws from *some* equilibrium of the underlying game—they satisfy Nash consistency by assumption, since they were generated by the players. But the equilibrium correspondence for a game of this scale (up to 190 potential interveners, each choosing a mixed strategy) generically admits multiple equilibria, and different conflict cases may be near different connected components. The realized binary outcomes provide only coarse information about which equilibrium each case is near: a 0/1/2 outcome is consistent with any mixed strategy that places positive probability on that action.

The burnout iteration uses the model to extract additional information. Given the realized action as a starting condition, iterating the model refines

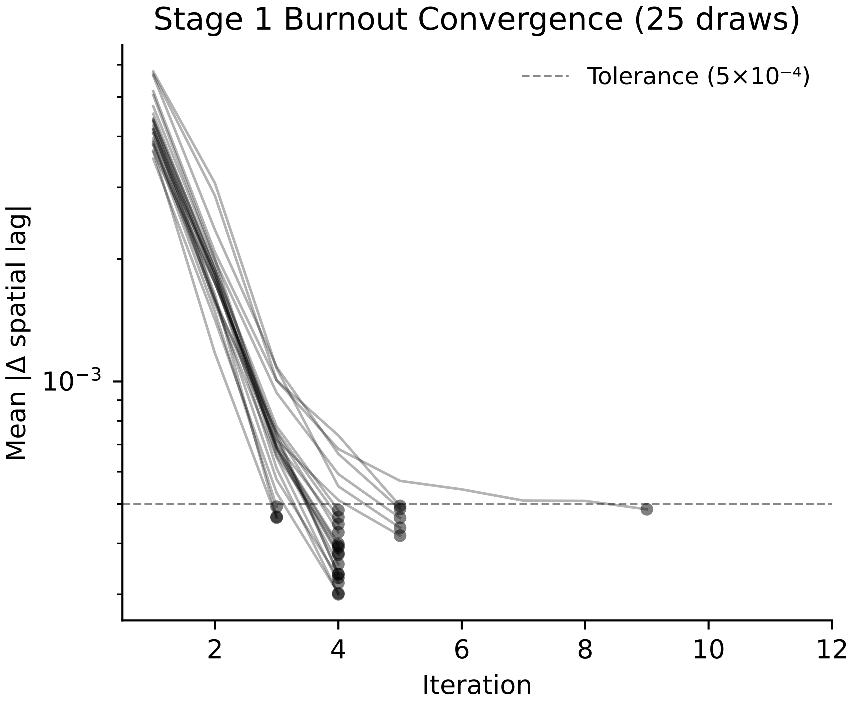

Figure 5: Burnout convergence across 25 imputation draws. Lines show the per-iteration trajectory of mean |Δspatial lag| for the 13 draws with full iteration histories; dots mark endpoints for the remaining 12 draws. Dashed red line: convergence tolerance (5×10^{-4}).

the estimated spatial lags toward values that are approximately consistent with the smooth probability distribution implied by $\hat{M}$. In this sense, the iteration performs *approximate equilibrium selection*: it identifies, for each conflict case, which region of the equilibrium correspondence the realized outcome is most consistent with, according to the model's learned mapping.

Three caveats are important. First, the model is trained on spatial lags derived from realized outcomes (the five-iteration warm start), not from the burnout-converged values. Retraining on the converged values would raise a circularity concern: the model's weights would depend on predictions the model itself generated. By freezing the model before burnout, the learned coefficients remain anchored to the training-time feature distribution, and the converged lags represent a small perturbation ($\Delta \approx 0.003$ relative to the warm start) rather than a wholesale shift to a foreign equilibrium. Second, the equilibrium selected for each training case by the burnout need not be the same equilibrium that governs prediction-time cases. The identifying assumption is that the equilibrium selection mechanism is stable across the sample—the same forces that push cases toward particular equilibria in training also govern unobserved cases. Third, the burnout-converged lags are used only to compute the shadow measure (the country-year expectations E_{gov} and E_{opp} in Stage 2) and not to re-estimate Stage 1. The shadow measure therefore reflects the approximately-equilibrium spatial environment, while the classifier's weights remain from training on the realized-outcome initialization.

F Comparison with the Cunningham Hierarchy

Cunningham (2016) measures anticipated government-biased intervention using the Lake security hierarchy, a cross-sectional index of U.S. patron-client relationships. Table 11 merges the hierarchy variable with the shadow measure on the common sample (6,973 country-years, 1950–2005, restricted to non-missing hierarchy scores) and compares four logistic onset models.

Neither proxy adds to the Baseline model (PRL and AUC are unchanged in both panels), and conditional on the shadow variables both coefficients are effectively zero. The shadow measure completely subsumes whatever onset-relevant information these proxies contain. This is consistent with Figure 6: the classifier recovers the patron-client signal embedded in the Lake hierarchy and the P5 strategic-game estimates while adding substantial variation from non-superpower interveners, bilateral relationships, and the spatial environment.

G Likelihood-Ratio Tests

Table 12 reports likelihood-ratio tests for the nested model hierarchy. Two sets of tests are presented: each specification against the Baseline (do the shadow variables as a group improve fit?) and each type-disaggregated specification against Entrants (does type-specific decomposition add beyond the aggregate shadow?).

Model	PRL	AUC	$\hat{\beta}_{\text{proxy}}$	p
Panel A: Lake hierarchy (6,973 CY, 1950–2005)				
Baseline	11.1%	0.782	—	—
Base + Hierarchy	11.2%	0.783	−0.37 (0.47)	0.43
Entrants	12.9%	0.796	—	—
Entrants + Hierarchy	12.9%	0.796	+0.05 (0.46)	0.92
Panel B: G&M structural P5 (7,772 CY, 1946–2014)				
Baseline	10.5%	0.772	—	—
Base + G&M	10.5%	0.773	+0.05 (0.62)	0.94
Entrants	12.1%	0.787	—	—
Entrants + G&M	12.1%	0.787	+0.16 (0.63)	0.80

Table 11: Subsumption tests against existing intervention proxies. Panel A adds the Lake security hierarchy from Cunningham (2016); Panel B adds the aggregate P5 intervention probability from Gibilisco and Monteiro (2022). Neither proxy improves PRL or AUC when added to any specification, and both proxy coefficients are effectively zero conditional on the shadow variables. Samples differ because the Lake hierarchy covers 1950–2005 and G&M covers 150 countries over 1950–1999.

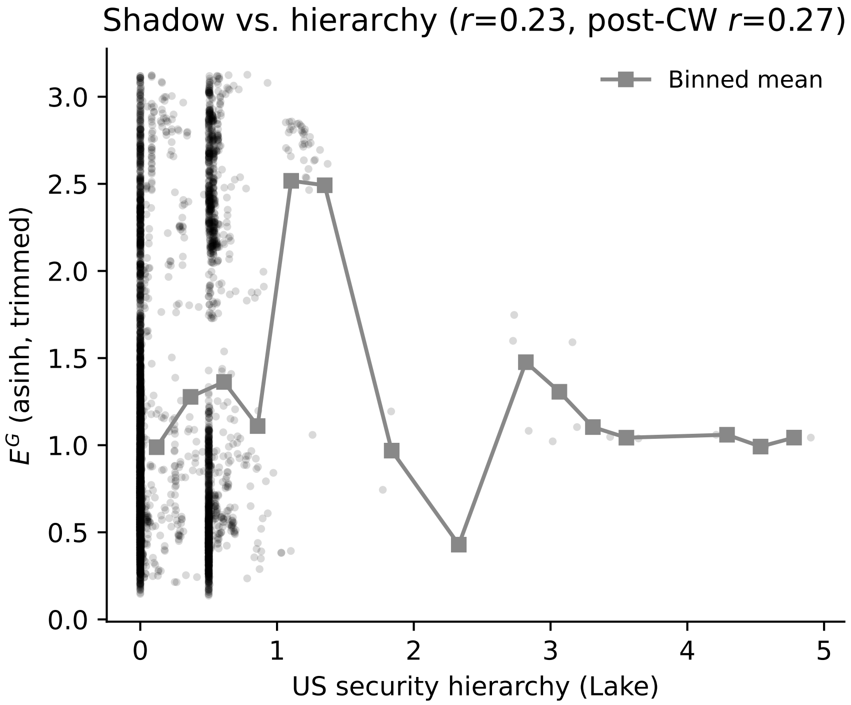

Figure 6: E^G (asinh, trimmed) versus US security hierarchy score from the Lake hierarchy data used in Cunningham (2016). Each point is a country-year (1950–2005); binned means show the conditional trend. The correlation is positive (r = 0.23) and stronger post-Cold War (r = 0.35), confirming that the classifier recovers the patron-client signal while adding substantial variation from non-superpower interveners.

			vs. Baseline		vs. Entrants	
Model	LL	k	LR	p	LR	p
Baseline	−799.5	13	—	—	—	—
Entrants	−783.2	15	32.7	<0.001	—	—
Powers	−761.6	17	75.8	<0.001	43.1	<0.001
Neighbors	−760.0	17	79.1	<0.001	46.4	<0.001
Coethnics	−775.1	17	48.8	<0.001	16.2	<0.001
Rulers	−783.2	17	32.7	<0.001	0.03	0.98
Rivals (bin)	−759.9	17	79.2	<0.001	46.5	<0.001
Rivals (cts)	−757.5	17	84.0	<0.001	51.3	<0.001
DOE	−782.3	17	34.4	<0.001	1.8	0.41
Full	−690.5	27	217.9	<0.001	185.3	<0.001

Table 12: Likelihood-ratio tests for nested onset specifications. LL = maximized log-likelihood; k = number of parameters (including constant). The "vs. Baseline" column tests whether adding shadow variables improves fit (df $= k - 13$); the "vs. Entrants" column tests whether type-disaggregated interactions improve on the aggregate shadow (df $= k - 15$). The shadow variables are jointly significant at $p < 0.001$ in every specification. Colonial ties (Rulers) and military balance (DOE) add nothing beyond the aggregate shadow.