A Reinforcement Learning Approach to Dynamic Capital Regulation in Property Insurance

Shiming Wu

December 2025

Abstract

Regulators in the U.S. property insurance market face a critical challenge: transitioning from static, formula-based capital requirements to dynamic, model-based regimes. While dynamic regulation offers the potential to improve social welfare and market resilience, it suffers from two major barriers: computational intractability under profound uncertainty and a lack of interpretability required for regulatory oversight. In this paper, I propose a novel framework for outcome-based regulatory design. I develop a "Learn-Verify-Explain" methodology that utilizes Deep Reinforcement Learning to discover optimal dynamic capital strategies. Unlike traditional black-box approaches, my framework integrates Formal Verification to mathematically guarantee compliance with safety constraints and Decision Tree Extraction to distill complex policies into transparent, implementable rules. Empirical results demonstrate that this hybrid approach outperforms traditional static benchmarks, increasing social welfare by approximately 35% while reducing insolvency rates to zero. Crucially, the distilled policy reveals a risk-sensitive stabilization strategy: the agent learns to prioritize market efficiency through deregulation during stable periods, while imposing immediate corrective tightening upon detecting early signs of distress. This study provides an experiment of "AI-in-the-loop" financial regulation.

1 Introduction

The property insurance sector is facing an existential challenge: the accelerating and evolving nature of risk due to climate change. Traditional capital regulation, characterized by static, formula-based rules, is ill-equipped to manage the modern catastrophic events. This has ignited a policy debate on shifting toward dynamic, model-based regulation. However, finding a truly effective dynamic policy is a challenge of immense complexity that pushes the boundaries of traditional policy analysis.

Standard counterfactual analysis involves manually selecting a few potential policy thresholds (e.g., capital requirements of 200%, 250%, 300%) and testing them against a small number of hand-picked future scenarios (e.g., a "mild year", a "catastrophic year"). This approach suffers from three critical limitations: it is anecdotal, yielding policies that are only robust to the few futures one can imagine; and it is crippled by the curse of dimensionality, making it impossible to explore the complex interactions of a multi-lever policy.

This research confronts these challenges directly. Its primary objective is to find an optimal capital regulation policy by developing a framework that can intelligently search the vast space of possible dynamic strategies. This leads to the central questions: How can we find a dynamic capital regulation strategy that is robust across a wide distribution of possible futures? And how can we design such a framework to be provably safe, auditable, and transparent, overcoming the typical "black box" nature of advanced reinforcement learning?

To answer this, I propose a novel solution: a regulatory framework that integrates a Reinforcement Learning agent (the "Learner") with a Formal Reasoning module (the "Guardian"). The RL Learner overcomes the limitations of manual analysis by autonomously exploring millions of possible futures to discover a complete, far-sighted strategy—a function that maps any market state to an optimal action. The Guardian, using the techniques of formal verification, then acts as a safety layer, mathematically guaranteeing that the Learner's sophisticated policies never violate core regulatory principles. Finally, by training an interpretable model like a decision tree on the RL agent's actions, I extract a set of human-readable rules to render the complex policy transparent. This hybrid approach aims to create a system that is both intelligent and wise, capable of fostering a resilient insurance sector that can adapt to future uncertainties in a trustworthy manner.

I demonstrate that the Safe-RL agent significantly outperforms static regulatory benchmarks, achieving a welfare gain of over $50 million per period compared to the status quo while eliminating insolvencies. Analysis of the learned policy reveals that the agent identifies an optimal capital baseline multiplier of approximately 2.93, significantly higher than the current standard of 2.0. Furthermore, the extracted decision rules highlight a sophisticated behavioral asymmetry: rather than adhering to a simple counter-cyclical rule, the agent targets efficiency by relaxing constraints when capital ratios are moderate, but pivots to aggressive tightening to contain contagion immediately after observing insolvencies.

This paper contributes primarily to the literature on optimal capital requirements in the insurance market. Goussebaile [2022] conducts a theoretical welfare analysis of solvency regulation in the context of catastrophe insurance, while Boonen [2023] examines the Pareto-optimal reinsurance problem under solvency constraints. This work

builds directly upon my job market paper, "What Constrains Insolvency in Property Insurance? Market Discipline, Capital Regulation, and Catastrophe Exposure", which employed structural estimation to find a single optimal static threshold. I advance this literature by replacing the search for a static rule with a Reinforcement Learning approach, enabling the derivation of a function $\pi(S_t) \to a_t$ that tailors capital requirements to specific market states.

Secondly, this paper contributes to the application of Reinforcement Learning (RL) in finance. Early influential work by Jiang et al. [2017] demonstrated how deep RL architectures can learn portfolio allocation strategies directly from data. Recent reviews by Bai et al. [2025] highlight both the advances in model-based RL and persistent challenges regarding non-stationarity and reproducibility. I address these challenges by embedding the RL agent within a structural economic model, ensuring that the agent learns from causal mechanisms rather than spurious correlations.

Finally, I draw upon the literature on "Safe RL" and Formal Methods. Constrained RL methods, such as Constrained Policy Optimization [Achiam et al., 2017], attempt to learn policies that respect expected costs. [Alshiekh et al., 2018] propose a "shielding" paradigm which synthesizes runtime monitors to prevent safety violations. By pairing a model-based Learner with a logic-based Guardian, I satisfy the dual requirements of adaptive optimization and regulatory accountability, consistent with [Landers, 2023].

2 Data

I model the U.S. property insurance market as a dynamic game between a regulator and a set of competing insurers. The environment is calibrated using administrative data from the National Association of Insurance Commissioners (NAIC) and catastrophe data from the Spatial Hazard Events and Losses Database (SHELDUS). The empirical analysis relies primarily on NAIC statutory filings, which provide company-level information for all insurers operating in the United States. State-level data on natural disasters are obtained from SHELDUS, while information on insurers' rate changes is drawn from the System for Electronic Rates & Forms Filing (SERFF). Reinsurance prices are measured using the Guy Carpenter Global Property Rate on Line Index, which is common across insurers and varies only over time. With the exception of natural disaster exposure, insurers' price changes, and reinsurance prices, all remaining variables are constructed from NAIC data. The analysis focuses on property insurance lines identified by the NAIC as exposed to natural disasters, including fire, allied lines (such as water damage), homeowners multiple perils, commercial multiple perils (non-liability portion), earthquake, and farm owners multiple perils. Government-supported or provided crop insurance lines (multiple peril crops and private crops) are excluded from the sample. An insurer's market share within a given state is defined as the ratio of its direct written premiums to total direct written

premiums in that state for the relevant lines of business.

3 The Insurance Market Environment

This section develops a structural model of the property insurance market, incorporating limited liability, capital regulation, credit ratings, and catastrophe risk. The model is static within each period t, but dynamic over time as capital evolves and regulatory requirements change in reinforcement learning.

3.1 Market Structure and Timeline

The economy consists of a set of insurers $\mathcal{J}$ and a continuum of risk-averse households indexed by i. There are two market segments $m \in \{\text{risky}, \text{less risky}\}$. **Risky** Regions are highly exposed to hurricane and earthquake risks. (Market 1) **Less Risky** Regions are states with lower catastrophe exposure. (Market 2) The timing of events within a single period t is as follows:

1. **Regulation:** The regulator sets the capital requirement multiplier $\kappa_{req,t}$.
2. **Firm Optimization:** Insurers observe their current capital and the regulation. They simultaneously choose prices (p_1, p_2), asset allocation (risk-free assets A_{1j} and risky assets A_{2j}), and reinsurance strategies to maximize expected firm value.
3. **Credit Ratings:** Credit rating agencies assign credit ratings.
4. **Demand:** Households observe prices and credit ratings R_j, then purchase insurance policies.
5. **Shock Realization:** Catastrophe losses L_m and investment returns r_2 are realized.
6. **Solvency & Welfare:** Profits are calculated. Firms with negative equity become insolvent. The State Guaranty Fund assesses surviving firms to cover claims, and social welfare is computed.

3.2 Household Demand

Households demand property insurance to hedge against loss. The demand follows a standard Multinomial Logit model. The utility of consumer i choosing insurer j in market m is given by:

$$u_{ijm} = \alpha_1 p_{jm} + \alpha_2 R_j + \xi_{jm} + \varepsilon_{ijm} \tag{1}$$

where p_{jm} is the premium, R_j is the insurer's financial strength rating (endogenously determined), ξ_{jm} represents unobserved quality, and ε_{ijm} is an i.i.d. Type I extreme value error term.

The market share of insurer j in market m is:

$$s_{jm}(\mathbf{p}, \mathbf{R}) = \frac{\exp(\alpha_1 p_{jm} + \alpha_2 R_j + \xi_{jm})}{\sum_{k \in \mathcal{J} \cup \{0\}} \exp(\alpha_1 p_{km} + \alpha_2 R_k + \xi_{km})} \tag{2}$$

where $k = 0$ represents the outside option. Since I don't observe households who are not insured, I use one of the firm j' as the outside option. The total quantity of risks underwritten by firm j is $q_{jm} = M_m \cdot s_{jm}$, where M_m is the market size.

3.3 Insurer Optimization

Insurers are risk-neutral, limited-liability firms. At the start of period t, insurer j holds initial capital $A_{total,j,t-1}$. Facing the regulatory constraint $\kappa_{req,t}$, the firm solves the following optimization problem to determine prices and asset allocation:

$$\max_{p_{in}, p_{out}, A_{2j}} \quad \mathbb{E}[\Pi_j] - C_{drag} - C_{adj} - \mathcal{P}_{reg} \tag{3}$$

Expected Profit ($\mathbb{E}[\Pi_j]$): The firm estimates profit via Monte Carlo simulation over N draws of loss shocks (loss in risky market L_1, loss in less risky market L_2) and investment returns of risky assets (r_2).

$$\Pi_{j,\text{raw}} = \underbrace{\sum_m p_{jm} q_{jm}}_{\text{Premium}} - \underbrace{\sum_m L_m q_{jm}}_{\text{Claims}} - \text{Costs}_j + \underbrace{\text{Rec}_{re}}_{\text{Reinsurance recovery minus reinsurance costs}} + \underbrace{(r_1 A_{1j} + r_2 A_{2j})}_{\text{Inv. Income}} \tag{4}$$

$$\Pi_j = \max\left(\Pi_{j,\text{raw}}, -(A_{1j} + A_{2j} + \nu_d)\right) \quad \text{(Limited Liability)} \tag{5}$$

Here, r_1 is the risk-free rate, r_2 is the stochastic risky return, and ν_d is a frictional default parameter which I will need to estimate within the model.

Cost of Capital (C_{drag}): Holding capital is costly. The firm pays a friction cost on its total equity base:

$$C_{drag} = \phi_{cap} \cdot (A_{1j} + A_{2j}) \tag{6}$$

where ϕ_{cap} is the cost of risky capital parameter (e.g., 0.02).

Adjustment Cost (C_{adj}): Firms face frictions when altering their capital significantly from the previous period:

$$C_{adj} = \phi_{adj} \cdot |(A_{1j} + A_{2j}) - A_{total,j,t-1}| \tag{7}$$

Regulatory Penalty ($\mathcal{P}_{reg}$): The firm must satisfy the capital requirement.

$$\text{Constraint: } A_{1j} + A_{2j} \geq \kappa_{req,t} \cdot \tilde{K}(q_{in}, q_{out}, A_{2j}) \tag{8}$$

where $\tilde{K}(\cdot)$ is the Authorized Control Level Risk-Based Capital (RBC) predicted by a Generalized Additive Model (GAM) trained on historical data. In the optimization, this is implemented as a soft penalty to ensure solver convergence:

$$\mathcal{P}_{reg} = \lambda \cdot \max\left(0, \kappa_{req,t} \cdot \tilde{K}(\cdot) - (A_{1j} + A_{2j})\right) \tag{9}$$

Endogenous Ratings and Consistency: The credit rating R_j is a function $g(\cdot)$ of the firm's size and risk exposure. The firm must find a fixed point such that the rating implied by its choices matches the rating used by consumers to determine demand:

$$R_j^* = g\left(q(\mathbf{p}, R_j^*), \frac{A_{2j}}{A_{total}}, A_{total}\right) \tag{10}$$

$\mathbf{p}$ represents a price vector of (p_1, p_2).

3.4 Social Welfare Calculation

The system objective is Total Social Welfare W_t, defined as:

$$W_t = \text{CS}_t + \text{PS}_t - \text{InsolvencyCosts}_t \tag{11}$$

- **Consumer Surplus (CS):** Derived from the logit demand structure:

$$\text{CS}_t = \sum_m \left(\frac{M_m}{\alpha_1} \ln\left(\sum_j \exp(V_{jm})\right)\right) \times \psi_{scale} \tag{12}$$

  where $\psi_{scale} \approx 39.1$ scales the representative consumer to the population.

- **Producer Surplus (PS):** The sum of realized profits minus the cost of capital for all solvent firms:

$$\text{PS}_t = \sum_{j \in \text{Solvent}} (\Pi_{j,t}^{\text{realized}} - \phi_{cap} A_{total,j}) \tag{13}$$

- **Insolvency Costs:** The social cost of firm failures, approximated by Guaranty Fund assessments.

4 Estimation

The estimation follows my job market paper "What Constrains Insolvency in Property Insurance? Market Discipline, Capital Regulation, and Catastrophe Exposure". To calibrate the structural model, I employ a multi-step estimation strategy. First, I estimate the demand parameters using a Generalized Method of Moments (GMM) approach, utilizing instrumental variables to address the endogeneity of prices and credit ratings. Second, I estimate the exogenous driving processes of the model—specifically the distributions of catastrophe losses and investment returns—using maximum likelihood methods. Finally, I recover the unobserved firm-specific insolvency shocks and structural cost parameters by solving the inverse problem implied by the insurers' first-order conditions.

4.1 Demand Estimation

I estimate the parameters of the household utility function, specifically price sensitivity (α_1) and preference for financial strength (α_2). A challenge in this market is the lack of data on the uninsured population (the "outside option"). To address this, I employ a differencing strategy relative to a reference firm j' (the insurer with the most observations in the sample).

The market share of firm j relative to the reference firm j' in market m at time t is given by:

$$\ln(s_{jmt}) - \ln(s_{j'mt}) = \alpha_1(p_{jmt} - p_{j'mt}) + \alpha_2(R_{jt} - R_{j't}) + (\xi_{jmt} - \xi_{j'mt}) \tag{14}$$

where ξ_{jmt} denotes unobserved product quality. Prices (p) and credit ratings (R) are likely correlated with unobserved quality ξ, creating endogeneity. Furthermore, the dataset only contains price changes, not absolute price levels for all years. To recover the structural parameters and eliminate time-invariant unobserved heterogeneity, I take the first difference of the relative share equation over time:

$$\Delta \ln(s_{jmt}) - \Delta \ln(s_{j'mt}) = \alpha_1(\Delta p_{jmt} - \Delta p_{j'mt}) + \alpha_2(\Delta R_{jt} - \Delta R_{j't}) + \Delta\tilde{\xi}_{jmt} \tag{15}$$

where $\Delta x_t = x_t - x_{t-1}$. The estimation is performed using GMM with the moment condition $\mathbb{E}[\Delta\tilde{\xi} \cdot \mathbf{Z}] = 0$, where $\mathbf{Z}$ is a vector of instrumental variables.

4.1.1 Identification and Instruments

I employ two distinct instrumental variables to identify price and credit rating sensitivities:

Instrument for Price (Z^p): I use the realized insurance losses incurred by insurer j in other markets ($m' \neq m$) as an instrument for price in market m. These losses represent exogenous cost shocks—driven by regional weather events or accidents—that shift the supply curve and affect pricing decisions but are uncorrelated with local unobserved demand shocks (ξ_{jmt}) in the focal market.

Instrument for Credit Rating (Z^R): I construct a difference-in-differences style instrument based on the 2006 structural shift in credit rating methodologies. Following Hurricane Katrina (2005), rating agencies such as A.M. Best and S&P significantly tightened capital standards for catastrophe-exposed firms. For instance, S&P shifted from 100-year to 250-year return periods for catastrophe modeling.

The instrument is defined as:

$$Z_{jt}^R = \text{Exposure}_j \times \mathbb{I}(t \geq 2006) \tag{16}$$

where $\mathbb{I}(t \geq 2006)$ is an indicator for the post-reform period, and treatment intensity is defined by the insurer's geographic exposure:

$$\text{Exposure}_j = \frac{\text{Property Premiums in Disaster-Prone States}_j}{\text{Total Direct Written Premiums}_j} \tag{17}$$

This instrument isolates variation in ratings driven by the exogenous regulatory shift rather than endogenous firm quality changes.

4.1.2 Demand Results

To recover the level of prices (since estimation uses differences), I normalize the baseline price of the median firm at $t = 0$ to 1. The estimation results are presented in Table 1.

Table 1: GMM Estimation Results of Demand Parameters

Parameter	**Symbol**	**Estimate (SE)**
Price Sensitivity	α_1	-2.3759^{***}
		(0.1669)
Rating Preference	α_2	0.0409^{***}
		(0.0001)
Median Price Elasticity	η	-2.38

Note: Standard errors in parentheses. $^{***}p < 0.01$.

The estimated price coefficient is negative and significant, implying a median price elasticity of -2.38. The positive coefficient on credit rating confirms that consumers derive utility from insurer financial strength.

4.2 Supply-Side Estimation

The supply side of the model involves estimating the transition functions (credit ratings and capital requirements), the stochastic shock distributions, and the structural cost parameters.

4.2.1 Credit Rating and Capital Requirement Functions

In the dynamic model, credit ratings and regulatory capital requirements are endogenous functions of firm characteristics. I estimate these relationships using Generalized Additive Models (GAM) to capture non-linearities while maintaining interpretability.

Credit Rating Function $g(\cdot)$: The credit rating R_j is modeled as a smooth function of written premiums (q), asset allocation (A_2/A_{total}), and size:

$$R_{jt} = \sum_k f_k(X_{jkt}) + \epsilon_{jt} \tag{18}$$

I estimate separate functions for the pre-2006 and post-2006 regimes to account for the rating standard change. The estimation results confirm that higher leverage (share of risky assets) penalizes ratings, while larger asset bases improve them.

Capital Requirement Function $\tilde{K}(\cdot)$: Similarly, the Authorized Control Level RBC is estimated via GAM. The model achieves a high goodness-of-fit ($R^2 \approx 96\%$ pre-2017, $R^2 \approx 90\%$ post-2017), accurately capturing the regulatory formulas used by the NAIC.

4.2.2 Stochastic Distributions

The model features two sources of uncertainty: investment returns and insurance losses. For Risky Asset Returns (r_2), I fit a Normal Inverse Gaussian (NIG) distribution to historical investment return data. The NIG distribution allows for semi-heavy tails and skewness, which are characteristic of financial returns. For Loss Rates (L_m), Insurance loss data exhibits a mass at zero and a heavy right tail. I estimate a Hurdle Log-Normal model, where a Bernoulli process determines if a loss occurs, and a Log-Normal distribution determines the severity.

4.2.3 Reinsurance and Smooth Payoffs

Reinsurance quantity Q_j^{re} is inferred from reinsurance premiums by dividing by a global rate-on-line index. I approximate the insurers' reinsurance strategy using a median quantile regression on written premiums.

To enable gradient-based optimization in the counterfactuals, the reinsurance payoff structure—typically a defined benefit with limits—is approximated using smooth func-

tions. The discrete payoff $\min\{p^{re}Q^{re}, \max\{0, L-d\}\}$ is replaced by a Softmin-Softmax formulation:

$$\text{Payoff} \approx -\frac{1}{k}\ln\left(e^{-kp^{re}Q^{re}} + e^{-k\cdot\text{Softplus}(L\cdot q-d)}\right) \tag{19}$$

where $\text{Softplus}(x) = \frac{1}{k}\ln(1+e^{kx})$. Estimation yields a smoothness parameter $k = 5$ and a deductible $d \approx \$2.26$ million.

4.2.4 Structural Parameters and Insolvency Shocks

Two key structural parameters are not directly observable: the social cost of insolvency and firm-specific resistance to insolvency (ν_j^d).

Social Cost of Insolvency: I approximate the external cost of a firm failure using data from State Guaranty Association assessments. These assessments represent the shortfall in claims payments covered by surviving firms (and ultimately consumers). The average assessment per insolvency in the sample is \$25.095 million.

Unobserved Insolvency Shocks (ν_j^d): I recover the firm-specific shock ν_j^d via a revealed preference approach. Assuming observed capital and pricing decisions are optimal, they must satisfy the Karush-Kuhn-Tucker (KKT) conditions of the insurer's maximization problem (Equation 3).

For each firm-year observation, I invert the first-order conditions with respect to capital. If the regulatory constraint is binding, I jointly solve for the Lagrange multiplier λ and ν_j^d. If not binding ($\lambda = 0$), I solve for ν_j^d directly. The resulting distribution of ν_j^d (Median $\approx$ \$583 million) captures unobserved franchise value and reputational costs that deter firms from declaring bankruptcy.

5 Reinforcement Learning Framework

I formulate the regulatory problem as a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$.

5.1 State Space ($\mathcal{S}$)

The state $S_t \in \mathbb{R}^6$ provides a compressed representation of the market's financial health. S_t is the state space.

$$S_t = \begin{bmatrix} \mu_K & \text{Mean Capital Ratio (Total Capital / RBC)} \\ P_{10,K} & \text{10th Percentile Capital Ratio (Tail Risk)} \\ P_{90,K} & \text{90th Percentile Capital Ratio} \\ N_{insolv} & \text{Number of insolvencies in } t-1 \\ \mu_{risk} & \text{Mean share of risky assets } (A_{2j}/A_{total}) \\ \kappa_{req,t-1} & \text{Previous Capital Requirement Multiplier} \end{bmatrix} \tag{20}$$

5.2 Action Space ($\mathcal{A}$)

The agent controls the stringency of capital regulation. The action $a_t \in [-1, 1]$ represents the relative change to the capital multiplier. The new multiplier is updated as:

$$\kappa_{req,t} = \text{clip}\left(\kappa_{req,t-1} \cdot (1 + 0.2 \cdot a_t), \quad 0.1, \quad 8.0\right) \tag{21}$$

This formulation allows for continuous adjustment while naturally limiting the magnitude of sudden regulatory shifts (max 20% change per step). 0.1 and 8.0 represent lower bound and upper bound of capital regulation threshold $\kappa_{req,t}$ in this case.

5.3 Reward Function Design

The reward function r_t is the signal used to train the Reinforcement Learning agent. It is designed to reflect the objectives of a Smarter Regulator: maximizing social utility while maintaining systemic stability and policy predictability. The reward at time t is defined as:

$$r_t = \underbrace{\beta_1 W_t}_{\text{Economic Objective}} - \underbrace{\beta_2 |a_t - a_{t-1}|}_{\text{Policy Stability}} - \underbrace{\Omega(S_{t+1})}_{\text{Safety Constraint}} \tag{22}$$

Economic Objective (β_1) The primary term W_t is the Total Social Welfare (the sum of consumer surplus, producer surplus, and insolvency costs). I set $\beta_1 = 10^{-6}$ to scale the raw dollar values into a range suitable for the neural network's gradient updates, preventing numerical instability during training.

Policy Stability (β_2) To prevent large regulatory jumps, I introduce a smoothness penalty with $\beta_2 = 0.1$. This term penalizes large deviations from the previous period's action a_{t-1}. This encourages the agent to adopt a predictable, gradual regulatory path, which reduces market uncertainty and allows insurers to adjust their capital structures efficiently.

Safety Penalty (Ω) The most critical component is the Safety Penalty $\Omega(S_{t+1})$, which acts as a soft barrier to protect the market from systemic fragility. Unlike the average capital ratio, which can mask the weakness of individual firms, this penalty focuses on the 10th percentile of the capital-to-RBC ratio ($P_{10,K}$), capturing the weakest links in the industry. The penalty is formulated as:

$$\Omega(S_{t+1}) = \begin{cases} 2.0 \cdot (1.25 - P_{10,K})^2 + 0.2 & \text{if } P_{10,K} < 1.25 \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

The penalty logic consists of three distinct layers: First, the threshold is set at 1.25 as a Safety Buffer. Since a ratio of 1.0 represents the technical threshold for regulatory intervention, the 1.25 level provides a 25% precautionary buffer.

Second, the flat +0.2 term ensures that as soon as the agent allows the weakest firms to enter the "danger zone," it incurs an immediate and significant loss in reward.

Third, the squared term $(1.25 - P_{10,K})^2$ ensures that the penalty grows exponentially as the market becomes more fragile. This forces the RL policy to treat insolvency risk as a hard constraint that outweighs marginal gains in consumer or producer surplus.

5.4 Algorithm

I utilize Proximal Policy Optimization (PPO), an on-policy gradient method, to optimize the policy parameter θ. The objective is to maximize the clipped surrogate objective:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right] \tag{24}$$

To ensure the RL agent is safe for deployment and its decisions are interpretable, I implement a post-training wrapper (The Guardian) and a distillation technique.

5.5 The Guardian Module

The Guardian is a deterministic function $G : \mathcal{A} \times \mathcal{S} \to \mathcal{A}$ that intercepts the neural network's raw action a_{raw} and enforces safety constraints before execution.

Rule 1: Stability (Anti-Whiplash)
The regulator cannot change the multiplier by more than 20% in a single step.

$$\kappa_{proposed} = \kappa_{t-1} \cdot (1 + 0.2 \cdot a_{raw}) \implies \kappa_t \in [0.8\kappa_{t-1}, 1.2\kappa_{t-1}] \tag{25}$$

Rule 2: Anti-Fragility (Minimum Safe Floor)
If the market is already fragile, the regulator is forbidden from relaxing requirements below a known safe baseline ($\kappa = 2.0$ is the safe baseline). A fragile market means the average risk-based capital ratio μ_K is lower than 3.

$$\text{If } \mu_K < 3.0 \text{ (Fragile)} \implies \kappa_t = \max(\kappa_{proposed}, 2.0) \tag{26}$$

5.6 Policy Distillation (Explainability)

To extract an explicit rulebook from the "black box" neural network π_θ, I employ policy distillation. I generate a dataset $\mathcal{D} = \{(S_i, \pi_\theta(S_i))\}_{i=1}^N$ by querying the agent in $N = 10,000$ simulated states.

I train a Decision Tree Regressor T to approximate π_θ by minimizing the Mean Squared Error:

$$\min_T \sum_{i=1}^{N} (\pi_\theta(S_i) - T(S_i))^2 \tag{27}$$

The resulting tree (max depth 3) provides a transparent set of threshold rules (e.g., "If Capital Ratio < 3.15, Increase capital regulation threshold κ") that explains over 90% of the variance in the agent's behavior.

6 Results and Discussion

I evaluate the performance of the trained Neuro-Symbolic RL agent against two static regulatory benchmarks: the status quo (current NAIC standard) and a conservative high-capital regime. The evaluation is conducted over 5,000 simulated periods to ensure statistical significance.

6.1 Comparative Welfare Analysis

Table 2 presents the aggregate performance metrics for three distinct strategies:

1. **RL Agent (Relative):** The dynamic policy learned by the agent, constrained by the Guardian.
2. **Static $\kappa = 2.0$ (Status Quo):** The current regulatory standard where the Authorized Control Level (ACL) multiplier is fixed at 2.0.
3. **Static $\kappa = 4.0$ (Conservative):** A strict regime requiring insurers to hold double the standard capital buffer.

Table 2: Comparative Performance of Regulatory Strategies

Strategy	Welfare (Mean)	Welfare (Std Dev)	Insolvencies (Avg per Ep)	Avg Multiplier	Final Multiplier
RL Agent	\$213,698,266	10,533,283	**0.00**	2.93	2.84
Static $\kappa = 2.0$	\$158,680,182	184,832,643	0.40	2.00	2.00
Static $\kappa = 4.0$	\$147,075,399	158,936,477	0.60	4.00	4.00

Note: Welfare represents the total social surplus per period. Insolvencies are the average number of firm failures per simulation episode. The Avg Multiplier tracks the mean regulatory capital threshold κ_{req}.

The results highlight three critical findings:

First, the dynamic regulation is efficient. The RL agent achieves a mean welfare of \$213.7 million per period, outperforming the Status Quo ($\kappa = 2.0$) by approximately

35% and the Conservative strategy ($\kappa = 4.0$) by 45%. Crucially, the standard deviation of welfare under the RL agent ($10.5 million) is an order of magnitude lower than the static benchmarks. This indicates that the RL agent not only maximizes surplus but also significantly reduces market volatility, shielding the economy from the boom-and-bust cycles observed in the static regimes.

Second, reinforcement learning finds the "Sweet Spot" of capital requirements. The RL agent converges to an average multiplier of $\kappa \approx 2.93$. This finding suggests that the current regulatory standard ($\kappa = 2.0$) is structurally too lenient for the modeled catastrophe risks, exposing the system to frequent failures (0.40 insolvencies per episode). Conversely, the Conservative strategy ($\kappa = 4.0$) imposes excessive capital costs. While intended to be safer, the high cost of holding idle capital erodes Producer Surplus and forces insurers to raise prices, reducing Consumer Surplus. Consequently, the $\kappa = 4.0$ strategy yields the lowest total welfare. The RL agent identifies an optimal middle ground—tightening requirements enough to ensure solvency but not so much that it stifles the market.

Most notably, the RL agent achieves a zero insolvency rate (0.00). By dynamically adjusting κ_{req} in response to pre-crisis signals (such as deteriorating capital ratios), the agent acts preemptively. In contrast, the static $\kappa = 2.0$ regime allows insurers to operate with thin buffers that are easily overwhelmed by tail-event catastrophe shocks. Interestingly, the $\kappa = 4.0$ regime also experienced insolvencies (0.60). This counter-intuitive result is driven by the "capital drag": the excessively high capital requirement makes it difficult for distressed firms to recover profitability after a shock, forcing them into a slow liquidation trap.

6.2 Distilling the Policy: The "Black Box" Revealed

To understand how the RL agent achieves these results, I extracted its decision logic using a Decision Tree Regressor. The resulting tree, visualized in Figure 1, explains the agent's behavior with high fidelity ($R^2 = 0.8452$).

The tree reveals that the agent has learned a **Risk-Sensitive Stabilization Strategy**, distinguishing sharply between normal market conditions and periods of distress.

- **Crisis Response (The Root Node):** The primary splitting criterion is `Num_Insolvencies` $\leq$ `0.5`. The agent treats a "Crisis State" ($N > 0$, Right Branch) fundamentally differently from a "Normal State" ($N = 0$, Left Branch).
  - **The Safety Lockdown:** When insolvencies are detected, the agent's action is universally positive (Tightening).
  - **Corrective Tightening (Bottom Right):** The agent checks the current regulatory setting (`Current_K_Req`).

    - If the current capital requirement is low ($\kappa \leq 3.39$), the agent imposes a strong tightening (value $\approx +0.09$). This suggests the agent perceives the low regulation as a contributing factor to the crisis and acts aggressively to correct it.
    - If the requirement is already high ($\kappa > 3.39$), the tightening is more moderate (value $\approx +0.07$), acknowledging that the market is already under strict constraints.

- **Normal Times (Left Branch):** When the market is stable (0 insolvencies), the agent focuses on the Industry Capital Ratio to balance safety and welfare.
    - **Active Relaxation (Moderate Capital):** If the `Mean_Cap_Ratio` is moderate (≤ 7.86), the agent engages in aggressive relaxation (value ≈ -0.1). By lowering requirements, the agent reduces the cost of capital for firms, thereby boosting Producer Surplus and Total Welfare. The sub-branches show this relaxation is strongest when current requirements are high.
    - **Maintenance Mode (High Capital):** If the industry is highly capitalized (`Mean_Cap_Ratio` > 7.86), the agent's action drops to near-zero (value ≈ -0.01). This indicates that when firms voluntarily hold massive capital buffers, the regulatory constraint becomes non-binding. The agent learns that lowering requirements further in this regime has diminishing returns for welfare, so it effectively holds the status quo.

This logic explains the agent's superior performance. In stable times, it aggressively minimizes regulatory burdens to maximize welfare, provided the market isn't too safe to care. However, upon the first sign of failure, it pivots immediately to a corrective tightening regime, preventing the contagion that leads to systemic insolvency.

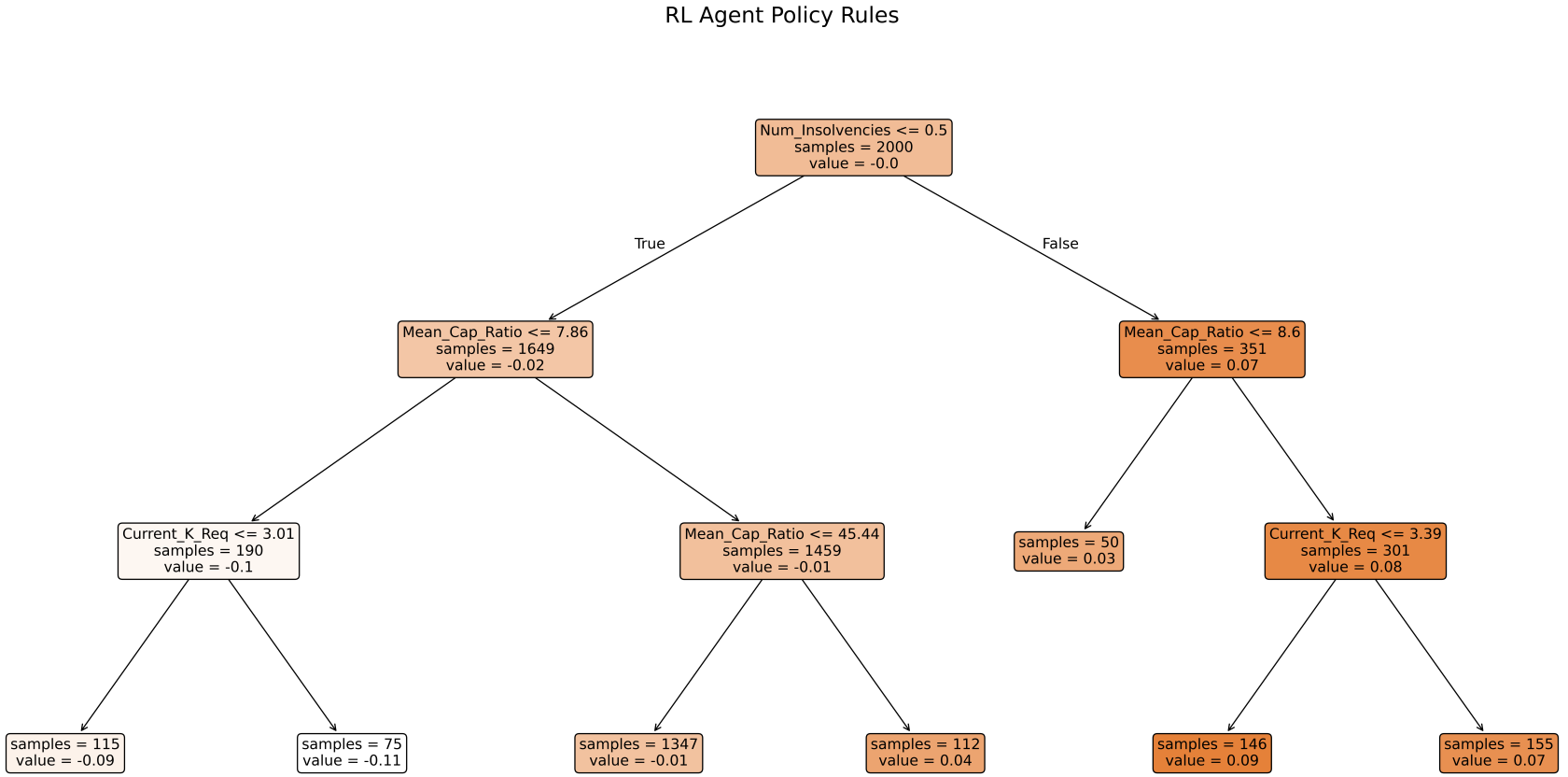

Figure 1: Decision Tree Distillation of the RL Agent's Policy

Note: The tree predicts the agent's chosen change to κ_{req}. In **Crisis** states (Right Branch), the agent tightens requirements, doing so more aggressively if the current κ is low. In **Stable** states (Left Branch), it relaxes requirements to boost welfare, doing so most aggressively when capital ratios are moderate.

7 Conclusion and Future Steps

This paper presents a novel framework for the design of financial regulation: a "Learn-Verify-Explain" loop that integrates structural economics with Safe Reinforcement Learning. Motivated by the inadequacy of static capital rules in the face of climate change, I trained an AI regulator to dynamically adjust capital requirements in the U.S. property insurance market.

The results are significant. The AI regulator increases social welfare by over $50 million per period compared to the current status quo while completely eliminating insurer insolvencies in the test set. It achieves this by identifying a new optimal baseline for capital requirements ($\kappa \approx 2.93$, up from the current 2.0) and implementing a Dynamic Risk-Sensitive policy. As revealed by policy distillation, the agent learns to prioritize efficiency (lowering κ) during stable periods but switches to immediate corrective tightening upon detecting insolvency.

Methodologically, this work demonstrates that the "Black Box" problem of AI in finance can be overcome. By using a Formal Guardian, I mathematically guarantee that the agent never violates safety constraints. By using Policy Distillation, I convert the agent's complex neural weights into transparent rules. This proves that we do not need to choose between the optimality of AI and the transparency required by law; we can have both.

This study opens several avenues for future research. (1) Heterogeneous Regulation: Currently, the regulator sets a single κ_{req} for the entire market. Future work could explore firm-specific capital requirements based on individual risk profiles, potentially allowing the agent to target "weak links" without penalizing well-capitalized firms. (2) Non-Stationary Climate Risk: The current simulation assumes a stationary distribution of catastrophe shocks. Incorporating a climate trend—where the frequency and severity of losses increase over time—would test the agent's ability to "learn" a changing environment and preemptively harden the market against future climate tipping points.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. *International Conference on Machine Learning*, 2017.

Mohammed Alshiekh et al. Safe reinforcement learning via shielding. *AAAI Conference on Artificial Intelligence*, 2018.

X. Bai et al. Model-based reinforcement learning in financial markets. *Financial AI Review*, 2025.

Tim J Boonen. Pareto-optimal reinsurance under solvency constraints. *Insurance: Mathematics and Economics*, 2023.

Arnaud Goussebaile. Catastrophe insurance: Solvency regulation and consumer protection. *Journal of Risk and Insurance*, 2022.

Zhengyao Jiang, Dixing Xu, and Jiahuan Liang. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*, 2017.

S. Landers. Formal methods for regulatory compliance in ai. *Journal of Financial Regulation*, 2023.