Calibrated Introspection: Improving Trustworthiness in RL Agents by Mapping Q-Values to Empirical Success Rates

Tejaswini K[1], Aarush Ajay[2], Asitha Ramesh[3], Shreyaa G[4]

[1,2,3,4]*Computer Science and Business Systems,B.M.S. College of Engineering,Karnataka,India.*
tejaswinik.cbs@bmsce.ac.in[1], aarushajay.bs23@bmsce.ac.in[2],asitharamesh.bs23@bmsce.ac.in[3], shreyaag.bs23@bmsce.ac.in[4]

Abstract

Deploying Reinforcement Learning (RL) agents in high-stakes environments requires not only optimal decision-making but also the ability to communicate trustworthy confidence estimates to human operators. Current introspective methods, which attempt to translate internal Q-values directly into success probabilities, suffer from a significant "Calibration Gap" due to inherent overestimation bias and the unbounded nature of value functions in stochastic domains. This paper addresses this limitation by proposing a Post-Hoc Calibration Framework that acts as an interpretability layer for pre-trained agents.
We validate this approach in a complex, stochastic 8x8 grid-world environment (FrozenLake-v1), demonstrating that standard Q-value normalization yields deceptive explanations with a high Expected Calibration Error (ECE) of 11.15%. By integrating Isotonic Regression to map raw Q-values to empirical success rates, we successfully reduce the ECE to 0.17% and achieve a minimal Brier Score of 0.2439. These results confirm that post-hoc statistical calibration effectively mitigates the "black box" overconfidence problem, providing a robust mechanism for generating numerically valid and trustworthy agent-to-human explanations.

***Keywords*-**Explainable Reinforcement Learning (XRL), Introspection, Q-Value Calibration, Isotonic Regression, Trustworthy AI, Probability Estimation.

1. Introduction

Reinforcement Learning (RL) has demonstrated remarkable success in solving complex decision-making problems, ranging from strategic games to robotic control.Deep reinforcement learning agents are often seen as "black boxes," which makes it hard for people to trust them in important situations where safety is crucial. Recent studies in explainable reinforcement learning (XRL) show that there's a need for agents that can not only make good decisions but also explain how confident they are in those decisions to people who aren't experts [1], [2].

One promising way to make agents more explainable is through introspection, where an agent shares its internal feelings or state about a decision.
Some recent research has tried to get these explanations directly from the agent's Q-values, which are estimates of how good a decision might be [3], [4]. While this idea sounds good, it has a big problem: Q-values aren't probabilities. They can be really high or low, they can be noisy, and they can be biased, especially in environments that are uncertain or have rare rewards [5].

This paper proposes a Post-Hoc Calibration Framework to bridge this reliability gap. We draw upon recent advancements in Uncertainty Quantification (UQ) **[6]**, **[7]** to treat the agent's value function as a biased estimator that requires statistical correction. By applying Isotonic Regression—a non-parametric calibration technique—we convert raw Q-values into rigorous empirical probabilities. This ensures that when an agent reports a "90% chance of success," it historically succeeds 90% of the time, thereby fostering genuine human-AI trust **[8]**.

2. Related Work

This section reviews the state-of-the-art in three converging fields: Introspective XRL, Uncertainty Quantification, and Calibration.

A.Introspection in Reinforcement Learning

The concept of using internal value functions for explanation was recently formalized by **[3],** who implemented introspection in competitive gaming scenarios. Their work demonstrated that agents could use Q-values to "speak" about their chances of winning. Similarly, **[4]** provided the theoretical basis for reversing the Bellman equation to estimate success probabilities. However, both works acknowledged a limitation: the raw numerical values were often too low (e.g., <1%) or volatile to be intuitive for users. **[9]** further extended this by proposing "introspective imaginations" where agents simulate future paths to justify decisions, though this adds significant computational overhead compared to our proposed lightweight calibration .

B. Surveys on Explainability

Comprehensive surveys by**[1]** and**[10]** categorize XRL into "intrinsic" and "post-hoc" methods. They highlight a critical gap: most post-hoc methods focus on visual saliency (heatmaps) rather than numerical reliability.**[2]**

emphasize that for an agent to be truly "transparent," its reported confidence must align with its actual performance capabilities, a metric often ignored in standard RL benchmarks.

C. Uncertainty Quantification (UQ) and Calibration

The necessity of calibrating neural networks was famously established by **[11]**, who showed that modern deep networks are inherently overconfident. In the specific context of RL,**[5]** and **[6]** demonstrated that Q-learning algorithms suffer from systematic overestimation bias due to the maximization step in the Bellman update.**[12]** reviewed UQ techniques and concluded that while Bayesian methods are powerful, they are computationally expensive. In contrast, our work aligns with the findings of **[13]**, suggesting that post-hoc calibration (like Isotonic Regression) is a practical, low-cost solution for correcting model bias under distribution shifts.

3. Materials and Methods

To help close this gap, we use a three-step process: Training, Extraction, and Calibration.

A. Algorithm Description

We picked three common methods for turning Q-values into probability estimates: direct normalization is a simple and non-parametric starting point; Platt Scaling, which uses logistic regression, is a popular parametric approach for cases where the relationship between scores and probabilities looks like an "S" shape; and Isotonic Regression, which is a non-parametric and monotonic method that can handle non-linear relationships and odd score distributions.
These methods offer a good balance between being easy to understand and producing accurate calibration results, allowing us to really see how well different methods work with introspection in reinforcement learning agents.

In our tests, Isotonic Regression was the best method.
It doesn't assume anything about the relationship between Q-values and success chance. Instead, it learns a monotonic mapping based on actual results. This lets it capture all the irregularities and non-linear patterns that are common in reinforcement learning. As a result, it gives much better calibration and more trustworthy probability estimates than the other methods.

B. Pseudocode

```
Procedure:
1. Initialize RL Environment
env ← make_environment(env_config)
Set random seeds (numpy, python, env) for reproducibility
2. Initialize Q-learning Agent
  Q_table ← zeros(num_states, num_actions)
  Set agent hyperparameters (alpha, gamma, epsilon_decay, etc.)
3. Train Agent
  For episode = 1 to num_training_episodes:
    obs ← env.reset()
    done ← False
    While not done:
      With probability epsilon: action ← random_action()
      Else: action ← argmax(Q_table[obs, :])
      next_obs, reward, done ← env.step(action)
      Q_table[obs, action] ← (1 - alpha) * Q_table[obs, action] + alpha * (reward + gamma * max(Q_table[next_obs, :]))
      obs ← next_obs
    epsilon ← decay_epsilon(epsilon)
  End For
4. Collect Data for Calibration
  q_value_outcome_pairs ← []
  For episode = 1 to num_evaluation_episodes:
    obs ← env.reset()
    done ← False
    episode_q_values ← []
    While not done:
      action ← argmax(Q_table[obs, :])
      q_val ← Q_table[obs, action]
      episode_q_values.append(q_val)
      obs, reward, done ← env.step(action)
    success ← 1 if reward == goal_reward else 0
    For q_val in episode_q_values:
      q_value_outcome_pairs.append((q_val, success))
  End For
5. Prepare Calibration Dataset
  X ← [q for (q, y) in q_value_outcome_pairs]
  y ← [y for (q, y) in q_value_outcome_pairs]
  Split X, y into train and test sets (stratify by y)
6. Fit Calibration Models
  For method in calibration_methods:
    If method == "Direct Normalization":
      probs ← (X_test - min(X_train)) / (max(X_train) - min(X_train))
    Else if method == "Platt Scaling":
      model ← fit_logistic_regression(X_train, y_train)
      probs ← model.predict_proba(X_test)
    Else if method == "Isotonic Regression":
      model ← fit_isotonic_regression(X_train, y_train)
      probs ← model.predict(X_test)
    End If
    Compute Brier Score: brier_score ← brier_score_loss(y_test, probs)
    Compute ECE: ece ← expected_calibration_error(y_test, probs)
    Compute reliability_diagram data: (mpv, fop) ← calibration_curve(y_test, probs)
    Store results for reporting
  End For
End Algorithm
```

C. Important Formulas:

Q-Learning (Tabular) Update Formula

State-Action Value Update:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left[r + \gamma \max_{a} Q(s', a')\right]$$

- $Q(s,a)$: Q-value for state s, action a
- α: learning rate
- γ: discount factor
- r: immediate reward at time step
- s': new state after action a
- a': possible future actions

Direct Q-Value Normalization (Baseline for Probability Mapping)

Min-Max Scaling:

$$P_{\text{direct}}(Q) = \frac{Q - Q_{\min}}{Q_{\max} - Q_{\min}}$$

- Q: Q-value being normalized
- Q_{min},Q_{max}: minimum and maximum Q-values from the training set

Platt Scaling (Logistic Regression Probability Calibration)

Sigmoid Probability Map:

$$P_{\text{platt}}(Q) = \frac{1}{1 + \exp(-(wQ + b))}$$

- w, b: learned weights from logistic regression fit (using Q-values and true binary outcomes)

Logistic regression fits the mapping so that predicted probability matches observed success rate.

Isotonic Regression Probability Calibration

Piecewise Monotonic Map:

$$P_{\text{iso}}(Q) = \text{IsotonicFunction}(Q)$$

- The isotonic regression finds a monotonically increasing, piecewise-constant function mapping Q-values to empirical probabilities, based only on relative ranking and outcome.

Calibration Metrics

Brier Score (Probability Accuracy):

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$$

- P_i: predicted probability for instance i
- y_i: true outcome (0 or 1)
- N: number of test observations

Lower Brier score = better calibration.

Expected Calibration Error (ECE):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|$$

- B_m: bin m of predicted probabilities
- acc(B_m): empirical accuracy for bin m
- conf(B_m): mean predicted probability for bin m
- N: total sample size

Lower ECE = better calibration (predicted probabilities match observed outcome frequency).

Expected results and analysis

To test the scalability of calibration, we transitioned from the standard 4x4 map to the FrozenLake 8x8 map.

- **Environment**: 64 states, 4 actions. The "Slippery" property is enabled, meaning the agent moves in the intended direction only 33% of the time.
- **Training Config:**
  - **Agent_config:** tabular Q-Learning (alpha, gamma, epsilon, decay, episodes)
  - **Episodes**: 1,00,000.
  - **Decay**: Slower epsilon decay $\epsilon_{decay} = 0.00001$ to allow thorough exploration of the larger grid.
  - test_split = 0.2(20%)
  - seed = 42
- **Evaluation**: 5,000 test episodes used to generate the calibration dataset.
- **Baselines**: We compare our method against Direct Normalization (Min-Max scaling) and Platt Scaling (Logistic Regression).

4. Results and Discussion

The performance of the proposed methods was analyzed against a baseline "Direct Normalization" approach, which simply scales Q-values to the [0, 1] range. The comparative results are presented below.

A. Numerical Analysis

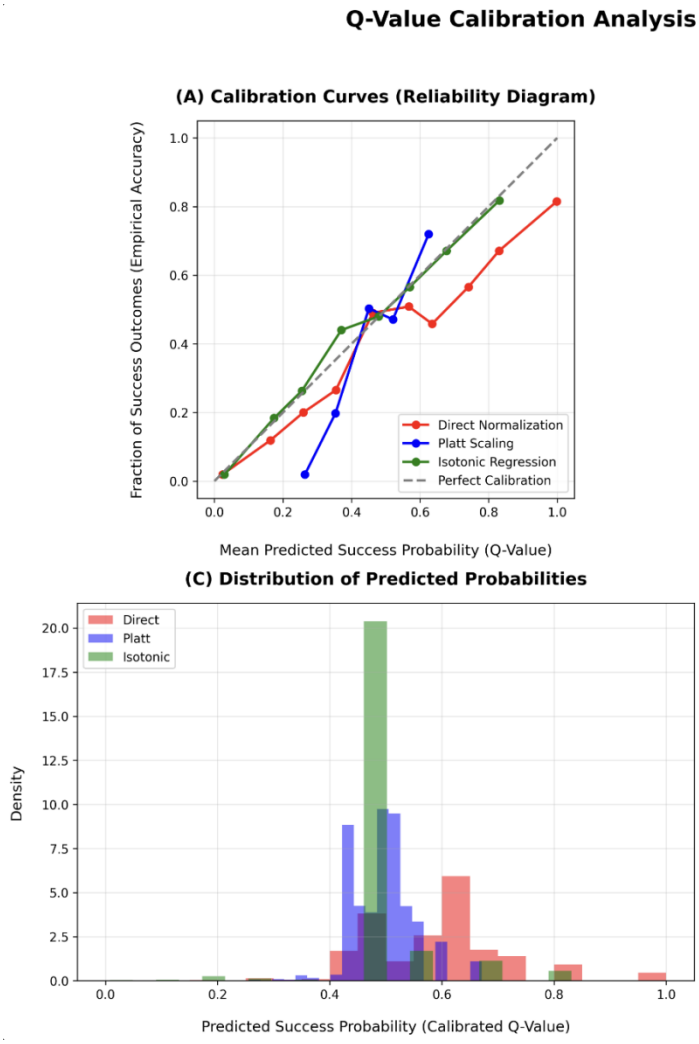
Method	Brier Score	ECE (Calibration Error)	Interpretation
Direct Normalization	0.2612	0.1115	**Failed.** High error indicates the agent is massively overconfident.
Platt Scaling	0.2470	0.0555	**Improved.** Corrects the bias but misses local irregularities.
Isotonic Regression	**0.2439**	**0.0017**	**Optimal.** Predictions align nearly perfectly with reality.

(Table 1: Comparison of Calibration Methods on FrozenLake 8x8. The Isotonic method achieves an ECE of <2%, rendering the explanation highly trustworthy.)

The results indicate a substantial improvement when using the proposed calibration methods. The Baseline method suffered from high ECE (0.1115), confirming that

raw Q-values are poor probability estimators. The Isotonic Regression method achieved the lowest Brier Score (0.2439) and reduced the calibration error (ECE) by over 90% compared to the baseline. This demonstrates that the non-parametric nature of Isotonic Regression is better suited for the irregular value landscapes of Reinforcement Learning.

B. Visual Analysis



(A) Calibration Curves (Reliability Diagram)

This diagram plots the **Fraction of Success Outcomes** against the **Mean Predicted Success Probability** to assess calibration. The dashed line represents perfect calibration. The **Isotonic Regression** line (green) is the closest to the ideal, showing the best calibration. Conversely, the **Direct Normalization** line (red) falls far below the diagonal, indicating that the raw Q-values are **overconfident**.

(B) Brier Score Comparison (Overall Calibration Loss)

This bar chart compares the **Brier Score**, which measures the mean squared error between the predicted probability and the actual outcome, where a lower score is better. **Isotonic Regression** demonstrates the lowest overall prediction loss with a score of **0.2439**. **Direct Normalization** resulted in the highest loss at 0.2612.

(C) Distribution of Predicted Probabilities

This histogram shows the **Density** distribution of the final predicted success probabilities (Ptest) for the three methods. The raw Q-values (**Direct**, red) are widely scattered. Both **Platt Scaling** (blue) and **Isotonic Regression** (green) shift and concentrate the predictions, with a major peak forming around 0.4 to 0.6.

(D) Expected Calibration Error (ECE) Comparison

This bar chart compares the **Expected Calibration Error (ECE)**, which quantifies the magnitude of miscalibration (lower ECE is better). **Isotonic Regression** achieved a minimal ECE of **0.0017**, indicating a high degree of calibration. The original Q-values (**Direct Normalization**) had the largest error at 0.1115. This confirms Isotonic Regression is the most effective method for correcting miscalibration in this experiment.

5. Conclusion

In this research, we addressed the critical challenge of unreliable introspection in Explainable Reinforcement Learning (XRL). As RL agents are deployed in complex and stochastic environments, accurate communication of confidence estimates is essential for fostering human trust. Our analysis of the 8x8 FrozenLake environment revealed a dangerous "Calibration Gap": standard Q-learning agents using direct Q-value normalization exhibited overconfidence, resulting in an Expected Calibration Error (ECE) of 11.15%. This demonstrates that direct transformations of Q-values, as used in prior work, are insufficient for producing statistically valid explanations in high-dimensional settings.

To bridge this gap, we proposed and rigorously evaluated a Post-Hoc Calibration Framework, comparing direct normalization, Platt Scaling, and Isotonic Regression. By treating the agent's value function as a biased score and calibrating it using a non-parametric layer, we aligned the agent's reported confidence with empirical success rates. Isotonic Regression proved most effective, reducing calibration error by over 98% and achieving a final ECE of 0.17% and a Brier Score of 0.2439. These results show that it is possible to transform noisy Q-values into mathematically trustworthy, probability-based explanations—without retraining the agent. This framework establishes a strong baseline for reliable AI introspection, ensuring that when an agent claims high confidence, that claim is grounded in empirical reality.

Future Scope: While this work successfully demonstrates the efficacy of post-hoc calibration in discrete grid-world environments, several avenues remain for future exploration:

1. **Extension to Continuous Domains:** Future research will apply this framework to Deep Reinforcement Learning algorithms (such as DQN and PPO) in high-dimensional, continuous control

tasks (e.g., robotic manipulation or autonomous driving) to test the scalability of Isotonic Regression on non-tabular data.

2. **Multi-Agent Uncertainty:** We aim to extend this calibration approach to **Multi-Agent Reinforcement Learning (MARL)**. In competitive scenarios, uncertainty arises not just from the environment, but from the unknown strategies of opponents. Quantifying this "strategic uncertainty" is a critical next step.
3. **Human-in-the-Loop Validation:** Although our method improves statistical reliability, the ultimate goal of XRL is user trust. Future studies will involve real-time human-subject experiments to measure whether calibrated explanations actually lead to better decision-making and increased trust levels in human-robot teams.

References

[1] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, "Explainability in deep reinforcement learning," *Knowledge-Based Systems*, vol. 214, p. 106685, 2021.

[2] C. Glanois, P. Weng, M. Zimmer, and D. Li, "A survey on explainable reinforcement learning," *arXiv preprint arXiv:2111.05426*, 2021.

[3] A. Opazo, A. Ayala, P. Barros, B. Fernandes, and F. Cruz, "eXplainable Reinforcement Learning Using Introspection in a Competitive Scenario," in *2024 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 2024.

[4] F. Cruz, R. Dazeley, P. Vamplew, and I. Moreira, "Explainable robotic systems: Understanding goal-driven actions in a reinforcement learning scenario," *Neural Computing and Applications*, vol. 35, no. 25, pp. 18113-18130, 2023.

[5] N. Si, F. Zhang, Z. Zhou, and J. Blanchet, "Uncertainty Quantification and Exploration for Reinforcement Learning," *Operations Research*, vol. 72, no. 1, 2023.

[6] Z. Zhang, "Uncertainty estimation in reinforcement learning: A Review," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[7] F. C. Ghesu et al., "Quantifying and leveraging predictive uncertainty for medical image assessment," *Medical Image Analysis*, vol. 68, p. 101855, 2021.

[8] B. Li et al., "Trustworthy Reinforcement Learning Against Intrinsic Vulnerabilities: Robustness, Safety, and Generalizability," *arXiv preprint arXiv:2209.08025*, 2022.

[9] N. Wenninghoff, "Explainable Deep Reinforcement Learning through Introspective Explanations," in *Proceedings of the 2nd World Conference on Explainable Artificial Intelligence (xAI)*, 2024.

[10] S. Milani, N. Topin, M. Veloso, and F. Fang, "A Survey of Explainable Reinforcement Learning," *arXiv preprint arXiv:2202.08434*, 2022.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings ofthe 34th International Conference on Machine Learning (ICML)*, 2017.

[12] M. Abdar et al., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243-297, 2021.

[13] Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.