

Uncertainty-Aware Policy Optimization for Robust Multi-Agent Decision Making under Partial Observability

Erik Andersson¹, Johan Svensson², Lars Johansson^{3*}

Department of Information Technology, Uppsala University, 751 05 Uppsala, Sweden

Corresponding author: l.johansson@kth.se

Abstract.

Decision-making in partially observable environments introduces uncertainty that can degrade coordination performance. This study develops an uncertainty-aware reinforcement learning framework by integrating soft actor-critic (SAC) with entropy regularization and Bayesian value estimation. The approach explicitly models uncertainty in both state representation and reward estimation, enabling more stable policy updates. Evaluation is performed on 10,000 tasks with varying levels of observation noise and incomplete information, including negotiation and multi-stage planning scenarios. The proposed method improves success rate from 69.2% to 81.5% and reduces variance in cumulative reward by 27.4% compared to standard SAC. Additionally, calibration metrics indicate better alignment between predicted and actual outcomes. The results demonstrate that uncertainty-aware optimization enhances robustness in collaborative decision systems.

Keywords:

Soft actor-critic; Uncertainty modeling; Partial observability; Multi-agent reinforcement learning; Robust decision-making

INTRODUCTION

Decision-making in partially observable environments remains a central challenge for collaborative reinforcement learning because each agent must act on incomplete, local, and sometimes noisy information while still maintaining effective coordination with others. This difficulty arises in many practical settings, including negotiation, distributed planning, and autonomous multi-agent systems, where the full environment state is not directly available to any single participant. In such tasks, local observations are often insufficient to capture latent task structure, hidden dependencies, or changes in the surrounding environment. As a result, agents must make decisions under uncertainty not only at the action stage, but also during state interpretation and coordination formation. Recent studies show that multi-agent reinforcement learning (MARL) has achieved encouraging progress in cooperative control, communication learning, and distributed decision-making, yet several core issues remain unresolved, including non-stationarity, credit assignment, and weak robustness when observations are uncertain or incomplete [1]. This challenge is particularly evident in multi-step inference settings, where policy quality depends heavily on how incomplete information is structured, represented, and propagated

during sequential decision-making [2]. Existing work on partially observable environments further indicates that hidden states, missing observations, and sensory noise can significantly weaken policy performance when uncertainty is not modeled in a principled manner [3]. To improve learning stability under such conditions, a large body of research has focused on entropy-based reinforcement learning and actor-critic methods. Maximum-entropy reinforcement learning promotes broader exploration by encouraging policies to remain sufficiently stochastic during training, which helps reduce premature convergence to poor local solutions. Within this family, soft actor-critic (SAC) has become one of the most widely used approaches because of its stable updates, sample efficiency, and strong empirical performance across continuous and sequential decision tasks [4,5]. Even so, recent studies suggest that value estimation error remains a major source of performance degradation. When the critic produces biased or unreliable estimates, policy updates may become unstable, especially in environments with noisy observations, delayed rewards, or ambiguous state transitions [6,7]. This problem is more serious in partially observable settings because the critic must infer long-term value from imperfect state information, which increases the risk of error propagation during training [8,9]. To address this issue, recent work has increasingly incorporated uncertainty into value estimation and policy learning. Uncertainty-aware methods attempt to distinguish reliable learning signals from ambiguous ones so that the policy is updated more cautiously when observations or returns are less trustworthy. Some studies reduce the influence of uncertain samples through weighting strategies, whereas others introduce probabilistic or Bayesian formulations to model uncertainty directly within the value function [10,11]. These developments suggest that policy performance depends not only on effective exploration, but also on whether uncertainty is explicitly accounted for during learning. In other words, stable decision-making under partial observability requires more than a good exploration mechanism; it also requires an estimation process that can recognize when its own predictions may be unreliable [12]. In multi-agent systems, this issue becomes even more complicated because uncertainty is not confined to the environment alone. Each agent must coordinate with others under changing local observations, incomplete information, and evolving interaction patterns. Errors in state interpretation may therefore affect both individual action selection and group-level coordination. Several studies have examined robustness in such settings and reported that ignoring state uncertainty can reduce coordination quality and weaken collective performance [13,14]. Related research has explored robust and safe MARL under disturbances, environmental variability, and time-varying observations, showing that uncertainty can strongly affect coordination outcomes when agents depend on partial information [15,16]. However, much of this work emphasizes control constraints, safety guarantees, or adversarial robustness, rather than policy optimization itself under partial observability. As a result, the relationship between uncertainty modeling and cooperative policy learning is still not fully understood. Another limitation of the current literature is the restricted scope of empirical evaluation. Many uncertainty-aware methods are developed for single-agent settings, offline learning, or relatively controlled benchmarks, and are therefore not directly suited to collaborative decision-making problems involving multiple agents with coupled objectives [17]. In addition, uncertainty is often addressed only at the environment level, while its influence on state representation, value estimation, and reward interpretation during training receives less attention. This is an important gap because observation noise and missing information do not only affect what agents see; they also influence how agents encode states, estimate returns, and evaluate coordination quality. Existing evaluations are also frequently limited to small benchmarks with narrow task variation, which makes it difficult to judge whether reported improvements can generalize to more complex scenarios involving

different noise levels, hidden information patterns, and multi-stage interaction structures [18]. For collaborative decision systems, strong average performance alone is not sufficient. A reliable method should also produce predictions and value estimates that remain consistent with actual outcomes when the observation process becomes less certain. These observations indicate the need for a more comprehensive learning framework in which uncertainty is treated as part of the cooperative decision process rather than as an external disturbance handled only at the environment level. For partially observable multi-agent problems, a useful framework should improve exploration, reduce critic bias, stabilize policy updates, and strengthen the consistency between predicted and realized returns. It should also remain effective across tasks with different uncertainty patterns, since real collaborative systems rarely operate under a single fixed level of noise or information completeness. From this perspective, integrating entropy-regularized policy learning with explicit uncertainty modeling offers a promising direction for improving robustness in collaborative reinforcement learning. This study develops an uncertainty-aware reinforcement learning framework for multi-agent decision-making under partial observability. The proposed method combines soft actor-critic with entropy regularization and Bayesian value estimation, and models uncertainty in both state representation and reward estimation. By incorporating uncertainty into these two components, the framework is designed to improve the stability of policy updates, reduce variance in learning outcomes, and provide more reliable value estimates during cooperative training. The method is evaluated on 10,000 tasks with different levels of observation noise, including negotiation and multi-stage planning scenarios, so that its behavior can be examined across a broader range of collaborative settings than is common in existing work. The purpose of this study is to determine whether explicit uncertainty modeling can strengthen cooperative policy learning in partially observable environments and thereby improve the robustness of collaborative decision systems. Its significance lies in moving uncertainty treatment beyond a narrow environmental perspective and embedding it directly into representation learning and value estimation, two components that strongly influence training stability and policy quality. Through large-scale evaluation across tasks with varying noise intensity and missing information, this work aims to provide empirical evidence on how uncertainty-aware learning affects success rate, reward variance, and agreement between predicted and actual outcomes. In this sense, the study contributes not only a practical reinforcement learning framework for collaborative systems with incomplete information, but also a clearer understanding of how uncertainty should be incorporated into multi-agent learning when reliable coordination under partial observability is the central goal.

2. Materials and Methods

2.1. Sample and Environment Description

Experiments were conducted in a simulated multi-agent environment with partial observability. A total of 10,000 tasks were generated. These tasks differed in observation noise and information completeness. Each task included 3–5 agents. Agents made decisions based on local observations and did not have access to the full state. The environment followed stochastic transitions to reflect uncertainty in state changes and rewards. Tasks were grouped by noise level and information loss. The dataset was divided into training, validation, and test sets with a ratio of 70%, 15%, and 15%. Task settings varied across episodes to avoid learning fixed patterns.

2.2. Experimental Design and Control Setup

Two methods were tested: standard soft actor-critic (SAC) and an uncertainty-aware version of SAC. Both methods used the same network structure and training settings. This ensured that differences in results came from the learning method rather than other factors. The uncertainty-

aware method extended SAC by including uncertainty in state features and value estimates. Training conditions were kept the same for both methods, including training steps and batch size. Each experiment was repeated five times with different random seeds. Results were averaged across runs. The comparison focused on success rate, reward variance, and calibration.

2.3. Measurement Methods and Quality Control

Three metrics were used: task success rate, variance of cumulative reward, and calibration error. Task success rate was defined as the ratio of tasks completed within a fixed number of steps. Reward variance measured the stability of returns across episodes. Calibration error was obtained by comparing predicted values with actual returns. All results were averaged over repeated runs, and standard deviations were reported. The same hardware and software setup was used for all experiments. Hyperparameters were fixed across methods. Early stopping based on validation results was applied to limit overfitting.

2.4. Data Processing and Model Formulation

Training data consisted of state–action–reward trajectories collected during interaction. Observations were normalized before being used as input. Rewards were scaled to a fixed range to avoid unstable updates. Policy learning followed the SAC framework. The objective function is given by

$$J(\pi) = \mathbb{E}_{(s,a) \sim D} [Q(s,a) - \alpha \log \pi(a|s)]$$

where $Q(s,a)$ is the value estimate and α controls the entropy term.

Uncertainty in value estimation was modeled using a simple stochastic form:

$$Q(s,a) = \mu(s,a) + \sigma(s,a) \cdot \epsilon$$

where $\mu(s,a)$ is the mean estimate, $\sigma(s,a)$ represents uncertainty, and ϵ is sampled from a standard normal distribution. This form allows the model to reflect uncertainty during training.

2.5. Training Procedure and Implementation Details

Training was performed for 1.2 million interaction steps. A replay buffer was used to store trajectories. Mini-batches were sampled uniformly for updates. The learning rate was set to 3×10^{-4} , and the discount factor was 0.99. Target networks were updated using soft updates. The entropy coefficient was adjusted during training to balance exploration and exploitation. The model was implemented in PyTorch and trained on a GPU. After training, the policy was tested on unseen tasks with different noise levels to evaluate robustness and generalization.

3. Results and Discussion

3.1. Overall Performance under Observation Uncertainty

The uncertainty-aware method achieved better results than standard SAC across all task settings. The difference became clearer as observation noise increased. Standard SAC showed a drop in performance under incomplete information, while the proposed method maintained a higher success rate. As shown in Fig. 1, performance improved more steadily across different noise levels. The success rate increased from 69.2% to 81.5%. This suggests that explicit treatment of uncertainty helps maintain decision quality when observations are limited. Similar patterns have been reported in recent studies, where policy performance depends on how uncertainty is handled during learning [19,20].

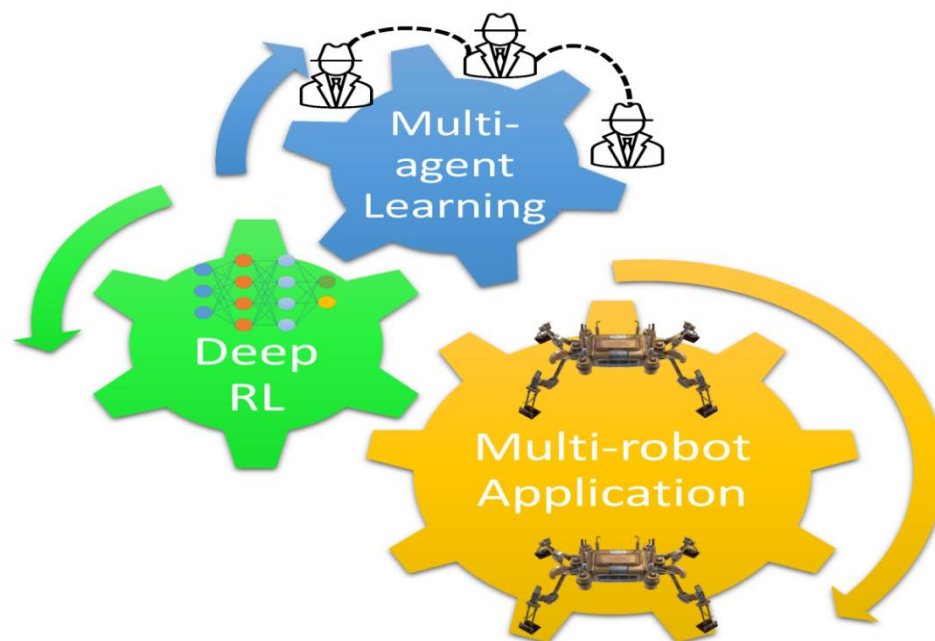


Figure 1. Task success rates for SAC and the uncertainty-aware method across noise levels.

3.2. Reward Stability and Policy Robustness

Differences were also observed in reward stability. Standard SAC produced larger fluctuations in cumulative reward, especially under high noise. The uncertainty-aware method reduced reward variance by 27.4%. This result suggests that policy updates became more stable when uncertainty was included in value estimation. In partially observable settings, errors in state perception can affect value learning and then influence action selection. The present results show that modeling uncertainty can reduce this effect and lead to smoother learning. This observation is consistent with earlier work that identifies critic reliability as a key factor in stable actor–critic training [21,22].

3.3. Calibration Performance under Partial Observability

The proposed method also improved calibration between predicted values and observed returns. This aspect is important in decision-making under uncertainty. A reliable model should not only perform well but also produce confidence estimates that match actual outcomes. As shown in Fig. 2, the uncertainty-aware method produced a closer match between predicted and realized returns. Standard SAC showed larger gaps under noisy conditions. This result indicates that Bayesian value estimation improves the quality of value prediction under incomplete information [23,24]. Better calibration can support safer decisions, especially in multi-agent settings where errors may propagate across agents.

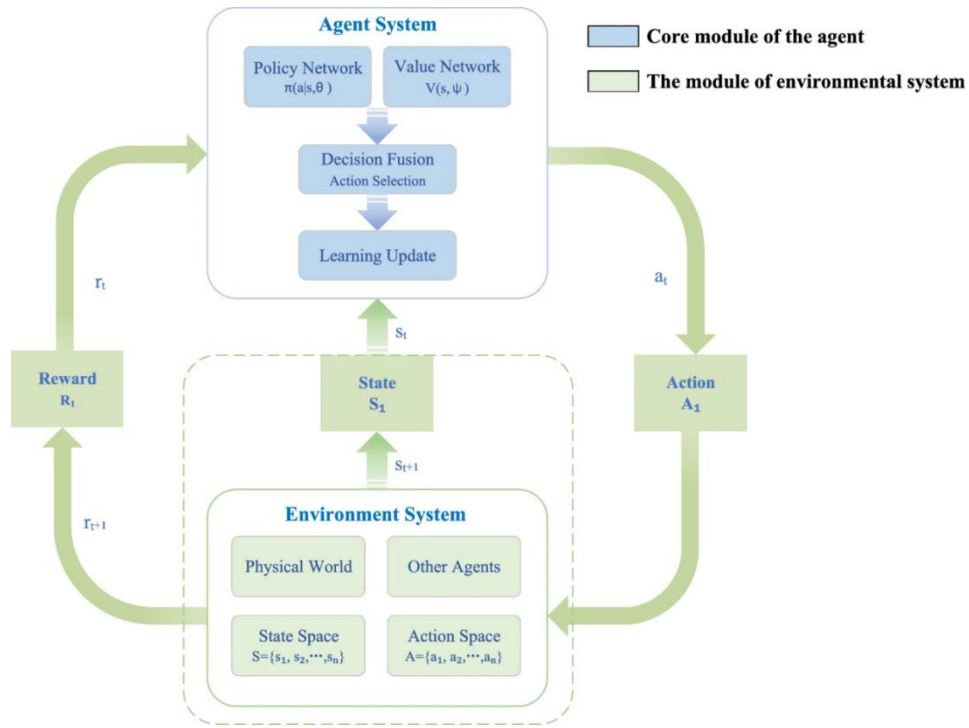


Figure 2. Predicted values versus observed returns under partial observability.

3.4. Comparison with Existing Studies and Practical Implications

Compared with existing studies, this work focuses on uncertainty in both state representation and value estimation during learning. Many earlier methods rely on entropy regularization or task-specific designs, while fewer studies consider uncertainty in both perception and evaluation at the same time. The results suggest that this combined approach improves policy robustness in partially observable tasks. However, some limitations remain. The experiments were conducted in a simulated environment, and the number of agents was limited. The task settings, although varied, still follow controlled conditions. Future work may include larger systems, richer communication structures, and real-world data. Despite these limits, the findings support uncertainty-aware learning as a practical approach for improving robustness in collaborative decision systems.

4. Conclusion

This study examined an uncertainty-aware learning scheme for multi-agent decision tasks under partial observability. Uncertainty was included in both state features and value estimation. This setting led to more stable learning and reduced sensitivity to noisy observations. Compared with standard SAC, the method achieved a higher success rate and lower variance in cumulative reward across task settings. It also produced value estimates that were closer to actual returns, which improves reliability under incomplete information. These results show that handling uncertainty during learning is important for policy quality and stability. The approach can be used in scenarios such as negotiation, distributed control, and multi-stage planning, where agents act with limited information. However, several limitations remain. The experiments were conducted in a simulated environment, and the number of agents was limited. The uncertainty model also relied on predefined forms. Future work may consider larger systems, more diverse tasks, and adaptive uncertainty models to improve robustness and scalability.

References

- Oroojlooy, A., & Hajinezhad, D. (2023). A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11), 13677-13722.
- Xu, D., Liu, H., Qiu, D., & Ma, Q. (2026). Structured Modeling and Representation Methods for Post-Retrieval Inference Processes in Large Video Language Models.
- Lambrechts, G., Bolland, A., & Ernst, D. (2022). Recurrent networks, hidden states and beliefs in partially observable environments. *arXiv preprint arXiv:2208.03520*.
- Liu, H., Xu, D., Ma, Q., Xu, S., & Qiu, D. (2026). Memory Poisoning Propagation and Repair Mechanism in Multi-Agent Collaborative Environments.
- Messaoud, S., Mokeddem, B., Xue, Z., Pang, L., An, B., Chen, H., & Chawla, S. (2024). S² AC: Energy-Based Reinforcement Learning with Stein Soft Actor Critic. *arXiv preprint arXiv:2405.00987*.
- Zhang, Y., Gu, W., & Wang, J. (2025). Research on First Article Inspection (FAI)-Driven Quality Assurance Methods for Wind Turbine Installation and Operation & Maintenance and Their Effect on Reliability Improvement. Available at SSRN 6094206.
- Palminteri, S. (2025). Human reinforcement learning processes and biases: computational characterization and possible applications to behavioral public policy. *Mind & Society*, 1-25.
- Xu, D., Gui, H., & Chen, H. (2026). Research on Layered Control and Fault Recovery Mechanisms for Fast Charging Safety Diagnosis of High Voltage Battery Systems Under Charging Network Interoperability Conditions.
- Pozzi, A., Incremona, A., & Toti, D. (2025). Imitation learning-driven approximation of stochastic control models. *Applied Intelligence*, 55(12), 838.
- Wang, Y., Chen, J., Wang, Y., & Yin, X. (2026). Application of Obtainable Biological Agent Characteristics in Efficacy Stratification of Oral Anti-Obesity Drugs.
- Campagner, A., Biganzoli, E. M., Balsano, C., Cereda, C., & Cabitza, F. (2025). Modeling unknowns: A vision for uncertainty-aware machine learning in healthcare. *International Journal of Medical Informatics*, 203, 106014.
- Zhang, Y., & Wang, J. (2026). Design and Implementation of a Computer-Aided Full Lifecycle Quality Management System for Wind Farms in Upgrades, Renovations, and Subcontractor Supervision.
- Kucuksayacigil, F., Zhang, Z., & Davidson, M. R. (2025). Coordinating power sector climate transitions under policy uncertainty. *Nature Communications*, 16(1), 3786.
- Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). The Impact of Visual Language Strategies in Public Art Creation on Community Spatial Perception and Public Behavior.
- Taghavi, M., & Vahidi, J. (2025). MARL-CC: A Mathematical Framework for Multi-Agent Reinforcement Learning in Connected Autonomous Vehicles: Addressing Nonlinearity, Partial Observability, and Credit Assignment for Optimal Control. *arXiv preprint arXiv:2511.17653*.
- Gao, G., Ma, X., Lu, C., & Gao, R. (2026). Reliability Analysis and Application Research of SMS Communication Systems in Medical Notification Scenarios.
- Shah, M. I. A., Barrett, E., & Mason, K. (2025). Uncertainty-aware knowledge transformers for peer-to-peer energy trading with multi-agent reinforcement learning. *arXiv preprint arXiv:2507.16796*.
- Xu, D., Chen, H., & Gui, H. (2026). Unified Online Estimation Method for SOC, SOH, and Power Capacity Considering Safety Boundary Consistency in Battery Management Systems.

- Anderson, D., DeLeo, R. A., Albright, E. A., Taylor, K., Birkland, T., Zhang, M., ... & Cage, C. (2023). Policy learning and change during crisis: COVID-19 policy responses across six states. *Review of Policy Research*, 40(1), 10-35.
- Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). Quantitative Study on the Construction and Application Effectiveness of Graffiti Wall Painting Teaching Models in Public Space Contexts.
- Springenberg, J. T., Abdolmaleki, A., Zhang, J., Groth, O., Bloesch, M., Lampe, T., ... & Riedmiller, M. (2024). Offline actor-critic reinforcement learning scales to large models. *arXiv preprint arXiv:2402.05546*.
- Wang, Y., Yin, X., Chen, J., & Wang, Y. (2026). Evidence-Based Study on Low-Burden Digital Phenotyping for Precision Screening of Oral Anti-Obesity Drug Efficacy.
- Schetinin, V., & Jakaite, L. (2025). Bayesian Learning Strategies for Reducing Uncertainty of Decision-Making in Case of Missing Values. *Machine Learning and Knowledge Extraction*, 7(3), 106.
- Yue, L., Xu, D., Qiu, D., Shi, Y., Xu, S., & Shah, M. (2025, December). Sequential Cooperative Multi-Agent Online Learning and Adaptive Coordination Control in Dynamic and Uncertain Environments. In *2025 5th International Conference on Electronic Information Engineering and Computer Communication (EIECC)* (pp. 692-697). IEEE.