

EXPLAINABLE ROBUSTNESS FOR ADAPTIVE DECISION SUPPORT
UNDER DEEP UNCERTAINTY

<https://doi.org/10.59982/18294359-26.1-as-33>

Arman Grigoryan

Phd in Technical Sciences, Associate Professor

EUA, Chair of Information Technologies and Applied Mathematics

a.grigoryan@eua.am

Abstract

Decision-making in complex, non-stationary environments requires models that are not only mathematically robust but also transparent to human stakeholders. While Robust Optimization (RO) and Distributionally Robust Optimization (DRO) provide essential safeguards against worst-case scenarios and model misspecification, they often function as "black boxes". This opacity creates a significant implementation paradox, i.e., the more sophisticated the robust model, the harder it is for human operators to justify its conservative choices, leading to automation bias or the "override trap".

This paper proposes a comprehensive framework for Explainable Robustness (XR) to bridge the gap between algorithmic resilience and human trust. We introduce Vector-Valued Robust Markov Decision Processes (VV-RMDPs), which decompose scalar value functions into explicit safety and efficiency margins. This decomposition supports Contrastive Explanations that clearly identify the specific environmental risks targeted by a policy. To ensure these explanations remain reliable under perturbation, we incorporate "Formal Verification" via Lipschitz stability bounds. Furthermore, we define a Trust-Aware Index that quantifies the quality of explanations by integrating mathematical fidelity, interpretability, and ethical compliance. By transitioning robust decision-making from opaque protection to interpretable insight, this framework provides a path toward sustainable human-AI teaming in high-stakes domains such as healthcare, autonomous infrastructure, and public policy [Grigoryan].

Keywords: Explainable Robustness, Adaptive Decision Support, Deep Uncertainty, Trust-Aware Explainable AI (TAXAI), Robust Markov Decision Processes, Distributionally Robust Optimization, Vector-Valued Robust MDPs, Human-AI Collaboration.

Introduction

Modern organizations operate in environments characterized by deep uncertainty, i.e., a state where not only probabilities are unknown, but the underlying models relating actions to consequences

are often contested or unreliable. In such situations, traditional "expected value" optimization becomes fragile. To counter this, the field has moved toward Robust and Adaptive Decision Models, such as Robust Markov Decision Processes (RMDPs) and Distributionally Robust Optimization (DRO), which provide mathematical guarantees against worst-case outcomes and parameter shifts. This shift is not merely technical. It represents a philosophical move from "predicting the future" to "preparing for any plausible future".

Despite their mathematical rigor, robust models face a significant barrier to adoption: the implementation paradox. Research consistently shows that high algorithmic sophistication often correlates with persistent failure rates in real-world organizational ecosystems [Grigoryan]. When a robust system recommends a highly conservative action (e.g., aborting a flight or rejecting a high-value trade), human operators often lack the insight to understand the specific ambiguity or threat the model is mitigating. Without explainable uncertainty, operators may perceive the model as overly cautious, leading them to override essential safeguards. This lack of transparency undermines the very protection the robust model was designed to provide [Grigoryan].

To address this, we propose the concept of Explainable Robustness (XR). Unlike standard Explainable AI (XAI), which focuses on explaining what a model decided, XR focuses on explaining why a specific level of robustness is necessary. It identifies the specific states of the world within the ambiguity set that could cause the policy to fail and explains how the current action protects against those states.

This requires a shift from scalar optimization to Vector-Valued decision logic, where the trade-offs between nominal performance and worst-case protection are made explicit. By formalizing these explanations and verifying their stability against noise, we can move from black-box resilience to insightful robustness [Grigoryan].

This paper makes three primary contributions:

1. Vector-Valued Robust Markov Decision Processes (VV-RMDPs): A mathematical framework for decomposing robust value functions into interpretable components (e.g., safety vs. efficiency) to generate contrastive explanations [Cialenco and Kováčová].
2. Verified Explanation Stability (VES): The application of Lipschitz continuity and formal verification to ensure that explanations remain faithful even as environmental conditions shift [Zhu et al.].
3. The Trust-Aware Explainable Artificial Intelligence (TAXAI) Framework: A quantitative method for evaluating the "trustworthiness" of robust explanations, integrating mathematical fidelity with human-centric interpretability [Salem and Khalil].

Robustness Under Deep Uncertainty

Markov Decision Processes and the Challenge of Uncertainty. A standard Markov Decision Process (MDP) is defined by the tuple $\langle S, A, P, r, \gamma \rangle$, where S is the state space, A is the action space, $P: S \times A \rightarrow \Delta(S)$ is the transition probability kernel, $r: S \times A \rightarrow \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor [Sutton and Barto]. The goal is to find a policy $\pi: S \times A$ that maximizes the expected cumulative reward. However, in deep uncertainty, the transition kernel P is unknown or unreliable.

Robust MDPs (RMDPs) and the Rectangularity Assumption. An RMDP extends the standard MDP by introducing an ambiguity set $P(s, a)$ of possible transition probability distributions for each state-action pair [Nilim and El Ghaoui]. The goal is to find a policy π that maximizes the worst-case expected return. The Robust Bellman Operator defines the robust value function $V(s)$ for a state s :

$$V(s) = \max_{a \in A} \left\{ r(s, a) + \gamma \inf_{p \in P(s, a)} \sum_{s'} p(s'|s, a) V(s') \right\}. \quad (1)$$

Under the assumption of rectangularity, where the ambiguity set $P(s, a)$ is independent across different state-action pairs, robust value iteration converges to a stationary optimal policy π [Iyengar]. This assumption is critical because it allows the "adversary" nature to choose the worst-case transition independently at each step, which simplifies the min-max optimization significantly.

Distributionally Robust Optimization (DRO) and Wasserstein Ambiguity Sets. DRO is a generalization of robust optimization that hedges against distributional uncertainty. A DRO problem typically takes the form:

$$\inf_{x \in X} \sup_{p \in P} E_p[f(x, \xi)], \quad (2)$$

where P is an ambiguity set centered at an empirical distribution $\hat{P}_n$. DRO with Wasserstein ambiguity sets, defined as $P = \{P : W_p(P, \hat{P}_n) \leq \epsilon\}$, where W_p is the Wasserstein distance of order p , is particularly powerful, and ϵ represents the radius (or size) of the Wasserstein ambiguity set P [Mohajerin Esfahani and Kuhn]. This approach provides finite-sample guarantees by bounding the distance between the true distribution P and the empirical distribution $\hat{P}_n$.

The tractability of Wasserstein DRO is rooted in its strong dual reformulation. For many convex losses, the inner supremum over the Wasserstein ball has a closed-form solution:

$$\sup_{p \in P} E_p[\ell(x, \xi)] = \frac{1}{n} \sum_{i=1}^n \sup_{\xi} \{\ell(x, \xi) - \lambda d(\xi, \xi_i)^p\} + \lambda \epsilon^p, \quad (3)$$

where $(\lambda \geq 0)$ is the dual variable corresponding to the Wasserstein radius ϵ , and $\ell(x, \xi)$ represents the loss function (or cost function) being evaluated [Gao and Kleywegt]. This reformulation allows DRO to be solved using standard convex programming tools like Gurobi or CPLEX [Nickel et al.].

The Explainable Robustness (XR) Framework

Vector-Valued Robust MDPs (VV-RMDPs). Standard RMDPs collapse all objectives into a single scalar reward. To provide insight, we define a Vector-Valued Reward $r(s, a) \in \mathbb{R}^k$, where k represents distinct objectives.

The vector-valued value function $V^\pi(s)$ is defined recursively:

$$V^\pi(s) = \min_{p \in P(s, \pi(s))} E_p \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) \right]. \quad (4)$$

This decomposition allows us to see how different objectives contribute to the overall robustness of a policy. By analyzing the components of $V^\pi(s)$, we can identify which specific worst-case scenarios (within the ambiguity set P) are driving the decision [Cialenco and Kováčová].

Contrastive Explanations and the Robust Advantage Vector. An XR-enabled system provides a Contrastive Explanation by comparing the chosen action $\mathbf{a}$ against a counterfactual action $\mathbf{b}$. The explanation is driven by the Robust Advantage Vector:

$$\Delta V(s, a, b) = V^{\pi_a}(s) - V^{\pi_b}(s). \quad (5)$$

The system identifies the component $j^* = \underset{j}{\operatorname{argmax}} \Delta V_j$ to explain the choice. For example, “Action A” is chosen because it maintains a 90% safety margin, even if visibility drops by 50%, whereas “Action B's” safety margin collapses to 20% under the same conditions [Salem and Khalil]. This provides the human operator with a clear, actionable understanding of the model's “fear.”

The Role of Uncertainty Decomposition in Explanation Selection. Recent research suggests that the type of explanation provided should depend on the nature of uncertainty. Epistemic uncertainty, i.e., the lack of knowledge, should trigger counterfactual explanations (what would need to change?), while aleatoric uncertainty, i.e., inherent randomness, should trigger feature-importance explanations (what are the key drivers?) [Zhu et al.]. The XR framework integrates this by monitoring the “uncertainty budget” of the VV-RMDP to dynamically select the most appropriate explanation modality.

Formal Verification of Explanation Stability

For an explanation to be useful in real-time adaptive support, it must be stable. If a tiny change in sensor data causes the explanation to flip, the human operator will lose trust in the system.

ՏԵՂԵԿԱՏՎԱԿԱՆ ՏԵԽՆՈԼՈԳԻԱՆԵՐ

Lipschitz Continuity and Explanation Astuteness. We define the stability of an explanation function $\mathcal{E}: S \times A \rightarrow \mathbb{R}^m$ using the Lipschitz Constant $L_{\mathcal{E}}$:

$$d(\mathcal{E}(s, a), \mathcal{E}(s', a)) \leq L_{\mathcal{E}} \|s - s'\|_2, \quad \forall s, s' \in S \quad (6)$$

A smaller $L_{\mathcal{E}}$ indicates a more stable explanation. We use Probabilistic Lipschitzness to bound the probability that an explanation remains faithful within a local neighborhood [Khan et al.].

Verified Robust Explanations. Borrowing from formal methods, a Verified Robust Explanation for a feature subset A is one where the explanation's "importance ranking" is guaranteed not to change for any perturbation within an ϵ – ball. In this context, ϵ represents the perturbation radius defining the local neighborhood $B(s, \epsilon)$ around a state or feature vector s . Formally, for all $s' \in B(s, \epsilon)$:

$$\text{sign}(\mathcal{E}(s', a)_j) = \text{sign}(\mathcal{E}(s, a)_j). \quad (7)$$

This ensures that the direction of the justification is mathematically sound [Wu et al.].

Quantifying Trust: The Trust-Aware Explainable Artificial Intelligence Framework

The effectiveness of XR is ultimately measured by human trust. We adopt the Trust-Aware Explainable AI (TAXAI) model to evaluate the quality of robust explanations.

The Trust Index (Fidelity, Interpretability, Compliance). The Trust Index $T(s, a)$ is a normalized score $[0, 1]$ calculated as:

$$T(s, a) = \omega_1 \cdot F(s, a) + \omega_2 \cdot I(s, a) + \omega_3 \cdot C(s, a), \quad (8)$$

where:

Fidelity F : The mathematical accuracy of the explanation relative to the model's internal logic (verified via Lipschitz bounds).

Interpretability I : The cognitive ease of understanding contrastive logic (e.g., number of features highlighted).

Compliance C : The degree to which the robust action satisfies safety thresholds and ethical constraints [Salem and Khalil].

Calibrated Trust and Adaptive Governance. A simple formalization of calibrated trust can be expressed as [Grigoryan]:

$$T_{\text{cal}} = \frac{1}{1 + e^{-\alpha(p-\theta)}}, \quad (9)$$

where p is the AI's predicted probability of a correct outcome, θ is a decision-maker-specific trust threshold, and α controls the steepness of the trust response. The XR system employs a Dynamic Trust Threshold τ . If $T(s, a) < \tau$, the system triggers a Reflective Mode, where the AI provides more detailed (but higher latency) explanations or requests a human override [Salem and Khalil].

Conclusion

This paper introduced Explainable Robustness, a framework that transforms "black-box" robust models into insightful decision support tools. By utilizing Vector-Valued RMDPs to provide contrastive logic and Lipschitz-based verification to ensure stability, we provide a path toward high-trust human-AI teaming under deep uncertainty.

In the next stage, we will focus on Integrated Explainable Robustness, extending the framework to multi-agent systems where agents must share "explanation fragments" without revealing private data. We also plan to integrate Causal Robustness, using causal inference to ensure explanations are not just statistically robust but also interventional. Finally, longitudinal trust studies will be conducted to empirically validate how Explainable Robustness affects human override rates over months of operational use.

Explainable Robustness represents a significant leap forward in the design of adaptive decision support systems. By bridging the gap between the mathematical rigor of robust optimization and the cognitive needs of human operators, we ensure that "safety" is not just a calculated margin but a transparent, verifiable, and trustworthy insight. This framework is essential for the next generation of autonomous systems, where the cost of a misunderstood "safe" action is as high as the cost of an unsafe one.

References

1. Cialenco, I., & Kováčová, G. (2026). Vector-Valued Robust Stochastic Control. *SIAM Journal on Financial Mathematics*, 17(3), pp. 770-800,
2. Gao, R. & Kleywegt, A. (2023). Distributionally Robust Stochastic Optimization with Wasserstein Distance, *Mathematics of Operations Research*, INFORMS, 48(2), pp. 603-655,
3. Grigoryan, A. (2025). A Systematic Review of Robust and Adaptive Decision Models in Complex and Uncertain Environments, *Proceedings of NPUA: Information technologies, Electronics, Radio Engineering*, 2025(2), pp. 27-37,
4. Iyengar, G. N. (2005). Robust Dynamic Programming. *Mathematics of Operations Research*, 30(2), pp. 283-304,
5. Khan, A., et al. (2024). Analyzing Explainer Robustness via Probabilistic Lipschitzness of Prediction Functions. *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, PMLR 238: pp. 1378-1386,

6. Mohajerin Esfahani, P., & Kuhn, D. (2017). Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations. *Mathematical Programming*, 171(1-2), pp. 115-166,
7. Nickel, S., Steinhardt, C., Schlenker, H., & Burkart, W. R. (2022). *Decision Optimization with IBM ILOG CPLEX Optimization Studio*. Springer. <https://doi.org/10.1007/978-3-662-65481-1>,
8. Nilim, A., & El Ghaoui, L. (2005). Robust Markov Decision Processes with Uncertain Transition *Matrices*. *Operations Research*, 53(5), pp. 780-798,
9. Salem, M. A., & Khalil, Z. A. (2026). Scalable Bayesian–XAI Framework for Multi-Objective Decision-Making in Uncertain Dynamic Systems. *Algorithms*, 19(5), 340 p.,
10. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Second Edition, MIT Press, 552 p.,
11. Wu, M., Wu, H., & Barrett, C. (2023). VERIX: Towards Verified Explanation of Deep Neural Networks. *Advances in Neural Information Processing Systems*, 36, 22247-22268,
12. Zhu, C., Bounia, L., Nguyen, V., Destercke, S. & Hoarau, A. (2025). Robust Explanations Through Uncertainty Decomposition: A Path to Trustworthier AI. *CoRR*, abs/2507.12913.

Submitted on 10.04.2026.
Sent for review on 07.05.2026.