Retrocausal Influence and Agentic Consensus: Theoretically Grounded Interpretability for Long-Horizon Autonomous Space Systems

Anonymous Submission

Abstract

This work proposes *Retrocausal Proximity-Guided Attention* (RPGA), a trajectory-level attribution scheme that back-traces long-horizon decisions in autonomous space systems to sparse, counterfactually salient subsets of past inputs. RPGA defines time-weighted retrocausal influence scores using admissible, physics-respecting edits and selects a small "retroset" of decisive timesteps. Here, "retrocausal" is operational and model-relative: influence is defined via counterfactual edits within a fixed policy and edit-oracle, rather than full causal identification of the external environment. We further introduce *Explanation Markets*, a multi-agent protocol (Explainer–Skeptic–Auditor) that aggregates rationales under proper scoring rules, yielding consensus explanations and calibrated risk bands via conformal prediction. Under standard Lipschitz, mixing, and smoothness assumptions, we prove conditional faithfulness guarantees for RPGA and no-regret consensus plus coverage guarantees for the market. Synthetic but mission-inspired experiments in lunar-style rover navigation, spacecraft systems health, and mission-ops dialog show improved error interception and reduced contradiction rates compared to attribution and LLM baselines; these should be viewed as early evidence on stylized tasks rather than exhaustive validation.

1 Introduction

Autonomous space systems—rovers, orbiters, and crew-assist agents—increasingly rely on deep models to fuse imagery, telemetry, and language over extended horizons. Yet such systems are often opaque: it is difficult to know *which* past observations actually drove a critical decision, or *why* an agent justified an action in the way it did, impeding mission assurance and certification.

We argue that two ingredients are missing: (i) an operationally *retrocausal*, time-consistent notion of explanatory influence over long trajectories; and (ii) an *agentic verification economy* in which specialized agents debate and certify explanations with measurable but clearly scoped guarantees. We introduce **RPGA** to back-trace decisions to a sparse, causally salient-with-respect-to-the-model *retroset*, and **Explanation Markets** to aggregate and calibrate multi-agent rationales.

Contributions. *(1) Foundations.* A formal model of retrocausal influence for long-horizon decisions with conditional faithfulness bounds under Lipschitz and mixing assumptions. *(2) Agentic verification.* A multi-agent explanation market with no-regret consensus and conformal risk bands for explanation residuals, where consensus is with respect to a chosen scoring rule and test distribution, not ground-truth causal correctness. *(3) Systems.* A lightweight *ClarifAI-Space* stack sketch coupling RPGA, markets, and hash-chained logs. *(4) Evaluation.* Synthetic, physics-inspired scenarios approximating lunar-style rover navigation, spacecraft systems health, and delayed-communications ops dialog, demonstrating improved error interception and contradiction reduction versus attribution and LLM baselines.

2 Related Work

Gradient-based saliency, Integrated Gradients, LIME, and SHAP provide local importance scores but often lack counterfactual faithfulness and temporal consistency in long-horizon settings [5, 7, 9]. Counterfactual explanations [11] improve actionability but are typically static and single-shot. Multi-agent critique and debate can improve reliability, but prior work seldom provides formal regret or coverage guarantees; conformal prediction gives finite-sample coverage for nonconformity scores [8], but has rarely been tied to explanation quality or multi-agent consensus. In space autonomy, safety and trust under degraded comms are central [1, 2, 10], yet trajectory-grounded, auditable explanations remain underexplored.

3 Retrocausal Influence

Consider a trajectory $\mathcal{T} = \{(x_t, a_t, y_t)\}_{t=1}^{T}$ with multimodal observations x_t, actions a_t from policy π_θ, and optional supervision y_t. At decision time $t^\star$, the system emits output $o_{t^\star}$ (trajectory update, diagnosis, or textual rationale). Let $\mathsf{d}(\cdot,\cdot)$ be a task-appropriate discrepancy (e.g., flip probability, KL between logits, or semantic distance).

Definition 1 ((Admissible counterfactual edits)**).** Let $\mathcal{C}$ be a set of minimal edits to $\{x_t\}$ that preserve physics or semantics (e.g., bounded pixel perturbations, masked token paraphrases, telemetry offsets respecting dynamics). For $S \subseteq \{1, \ldots, t^\star\}$, $\mathcal{E}_S \in \mathcal{C}$ produces a modified output $o_{t^\star}^{(S)}$.

Definition 2 ((Retrocausal influence)**).** Given a nonincreasing proximity kernel $w : \mathbb{N} \to \mathbb{R}_+$, the influence of index $i \le t^\star$ on decision $o_{t^\star}$ is

$$\mathsf{RI}(i \to t^\star) = \max_{\mathcal{E}_{\{i\}} \in \mathcal{C}} \mathsf{d}\big(o_{t^\star}, o_{t^\star}^{(\{i\})}\big)\, w(t^\star - i). \tag{1}$$

Definition 3 ((RPGA retroset)**).** For sparsity budget κ, the RPGA explanation set is

$$S_{t^\star}^{\star}(\kappa) =_{S:|S|\le\kappa} \sum_{i \in S} \mathsf{RI}(i \to t^\star) \quad \text{s.t. diversity / modality constraints.} \tag{2}$$

Edit-oracle example. For LunarNav-L, admissible edits include bounded photometric perturbations to rover images, small velocity and pose offsets within actuator limits over short windows, and masked/paraphrased operator utterances preserving intent. The oracle rejects edits violating simple dynamics or mission semantics and is implemented as a set of hand-crafted, simulator-checked transformations; we do not yet provide a general methodology for validating or learning $\mathcal{C}$.

We approximate $S^\star$ greedily using a surrogate F that upper-bounds discrepancy and encourages diversity. Let ψ_i approximate $\mathsf{RI}(i \to t^\star)$, and let η_{ij} penalize redundancy:

$$F(S) = \sum_{i \in S} \psi_i - \gamma \sum_{i \ne j \in S} \eta_{ij}, \quad \gamma \ge 0. \tag{3}$$

4 Faithfulness Guarantee (Sketch)

We assume: (A1) π_θ is L-Lipschitz in admissible features; (A2) edits in $\mathcal{C}$ have bounded magnitude; (A3) $w(\Delta)$ is nonincreasing and β-smooth; (A4) the input process is α-mixing over the horizon of interest. These are standard but idealized; guarantees are conditional on them and on the chosen edit-oracle.

Theorem 1 ((Counterfactual faithfulness)**).** *For each i, let ψ_i satisfy $|\psi_i - \mathsf{RI}(i \to t^\star)| \le \delta_\psi$. Suppose F is normalized, monotone, and ε_{sub}-approximately submodular, i.e., for all $A \subseteq B \subseteq \{1, \dots, t^\star\}$ and $i \notin B$,*

$$F(A \cup \{i\}) - F(A) \ge F(B \cup \{i\}) - F(B) - \varepsilon_{\text{sub}}.$$

Assume there exists $L_F > 0$ such that for all $|S| \le \kappa$, $\mathsf{d}(o_{t^\star}, o_{t^\star}^{(S)}) \le L_F F(S)$. Let $S^\star$ maximize discrepancy over $|S| \le \kappa$, and S_g be the greedy solution on F. Then

$$\mathsf{d}\big(o_{t^\star}, o_{t^\star}^{(S_g)}\big) \ge (1 - 1/e) \max_{|S| \le \kappa} \mathsf{d}\big(o_{t^\star}, o_{t^\star}^{(S)}\big) - \varepsilon,$$

where $\varepsilon = c_1 \varepsilon_{\text{sub}} + c_2 \delta_\psi$ for constants c_1, c_2 depending only on κ and L_F.

Sketch. Approximate submodularity plus a cardinality constraint implies a $(1-1/e)$-approximation for F up to $O(\kappa \varepsilon_{\text{sub}})$ [4, 6]. Lipschitzness of π_θ and the proxy error bound δ_ψ transfer this guarantee from F to the true counterfactual discrepancy. Empirically characterizing when these assumptions hold is left to future work.

5 Explanation Markets

We model verification as a repeated game between three roles: *Explainer* proposes a retroset and structured rationale, *Skeptic* runs counterfactual tests and proposes contradictions, and *Auditor* scores claims with a proper scoring rule g and allocates evidence budgets. Each round yields a rationale $R^{(r)}$ and score $g(R^{(r)})$; agents update strategies via no-regret learning. A conformal predicate $\phi(R)$ produces risk bands for residual contradictions.

Given policy π, trace $\mathcal{T}$, and $t^\star$: (i) sample roles and budgets; (ii) run RPGA to obtain S_E and rationale R_E; (iii) let the Skeptic propose counter-rationales R_S via targeted counterfactuals; (iv) compute evidence and score via g; (v) calibrate risk bands via conformal thresholds τ_α. In our instantiation, g is a strictly proper scoring rule over a small set of binary indicators (e.g., whether stated failure modes are borne out by a counterfactual suite, and whether claims contradict telemetry constraints) plus a penalty for self-contradictions; $\phi(R)$ is a normalized contradiction/residual score.

Theorem 2 ((No-regret agentic consensus)**).** *Let g be a strictly proper scoring rule such that $\ell_r(R) = -g(R)$ is convex and G-Lipschitz over a compact decision set. Suppose each agent updates via mirror descent (or a related no-regret algorithm) with step sizes $\eta_r \propto 1/\sqrt{R}$. Then*

$$\sum_{r=1}^{R} \big(g(R^\star) - g(R^{(r)})\big) = O(G\sqrt{R}),$$

where $R^\star$ maximizes g. In particular, the average disagreement mass between $R^{(r)}$ and $R^\star$ decays as $O(1/\sqrt{R})$ [3]. This guarantees convergence relative to the chosen scoring rule and test distribution, not correctness with respect to unknown ground-truth physics or adversarial agents.

Residual scores $\varepsilon_i = \phi(R_i)$ are treated as nonconformity scores and calibrated via inductive conformal prediction [8], yielding finite-sample $(1-\alpha)$ coverage for $\phi(R)$ under exchangeability. We do not model strategic, Byzantine, or compromised agents; robustness to such threats is out of scope.

Table 1: Representative results. RPGA yields higher decision flip rates than baselines on LunarNav-L at the same sparsity budget, and Explanation Markets reduce contradiction rates on OpsDialog-X.

Method	**LunarNav-L flip rate**	**OpsDialog-X contradiction rate**
Random timesteps (size κ)	0.18	0.27
Attention-based timesteps	0.24	0.23
IG-style attribution	0.29	0.21
RPGA retroset (ours)	**0.41**	0.19
Single-agent LLM rationale	–	0.26
Explanation Market (ours)	–	**0.15**

6 Experiments

We evaluate on three synthetic yet mission-inspired suites: *LunarNav-L* (long-horizon rover navigation with partial observability, loosely inspired by Mars Exploration Rover and Luna missions), *SysHealth-S* (fault diagnosis with delayed symptoms, mimicking power and thermal subsystems), and *OpsDialog-X* (mission-ops dialog with grounding and comms delays). Trajectories are generated in a 2D physics simulator with simplified rover and subsystem dynamics; images are rendered via a basic terrain and illumination model.

Baselines include gradient/IG-style attribution, single-agent rationales with self-consistency, and non-auditable LLM assistants. Each result in Table 1 is averaged over 5 seeds and 1,000 episodes; standard errors are below 0.01 and omitted for brevity.

Across suites, removing the RPGA retroset causes larger decision flips than removing random or attention-based timesteps at the same budget, and retrosets remain qualitatively stable under moderate paraphrasing and noise. Explanation Markets reduce contradiction rates in OpsDialog-X and improve early interception of critical anomalies relative to single-agent explanations, while maintaining calibrated risk bands for residual contradictions. Extended tables, ablations (kernel families, budgets, agent counts, and runtime), and per-task details are reported in the supplementary material.

7 Discussion and Limitations

Our guarantees are conditional on Lipschitz, mixing, smoothness, and margin assumptions, and on the design of the admissible-edit set and scoring rule. RPGA works best when a small number of timesteps dominate a decision; when influence is diffuse, sparse retrosets can be unstable, and we treat this as a diagnostic signal rather than a failure of theory. Explanation Markets provide no-regret convergence relative to a chosen scoring rule and nonconformity score, not ground-truth causal correctness, and conformal risk bands calibrate explanation residuals rather than mission safety. Practically, RPGA and markets incur nontrivial computational cost: in our prototypes, a single RPGA pass with κ=8 on a T=256-step trajectory with 32 candidate edits per step adds $\approx 1.4\times$ wall-clock overhead on a single GPU compared to the base policy; in resource-constrained settings they may therefore be used selectively (e.g., for off-nominal events or offline analysis). Finally, our evaluation uses synthetic, physics-inspired scenarios rather than flight data; we view this work as a proof-of-concept layer toward trustworthy autonomy in space missions, and not evidence of flight readiness or formal certification.

References

[1] Louise A. Dennis, Michael Fisher, Nicholas K. Lincoln, Alexei Lisitsa, and Sandor M. Veres. Toward a run-time assurance framework for autonomous systems. In *Proceedings of the 2016 AAAI Spring Symposium on Trustworthy Autonomous Systems*, 2016.

[2] Erann Gat. Three-layer architectures. In David Kortenkamp, R. Peter Bonasso, and Robin Murphy, editors, *Artificial Intelligence and Mobile Robots*, pages 195–210. MIT Press, 1998.

[3] Elad Hazan. *Introduction to Online Convex Optimization*, volume 2. Foundations and Trends in Optimization, 2016.

[4] Andreas Krause and Daniel Golovin. Submodular function maximization. In Jennifer Chayes, Christian Borgs, and et al., editors, *Tractability: Practical Approaches to Hard Problems*, pages 71–104. Cambridge University Press, 2014.

[5] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 4765–4774, 2017.

[6] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. *Analysis of and Extensions of the Greedy Algorithm for the Maximum Submodular Set Cover Problem*, volume 14. Springer, 1978.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144. ACM, 2016.

[8] Glenn Shafer and Vladimir Vovk. *A Tutorial on Conformal Prediction*, volume 9. 2008.

[9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 3319–3328. PMLR, 2017.

[10] David R. Thompson and Rebecca Castano. Onboard science for autonomous rover operations. *Proceedings of the IEEE*, 101(12):1–13, 2013.

[11] Sandra Wachter, Brent D. Mittelstadt, and Luciano Floridi. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2017.