EXACTLY FAITHFUL, BARELY INFORMATIVE: AN ATTRIBUTION AUDIT OF INTERPRETABLE INVENTORY REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpretable reinforcement learning is often motivated by an implicit syllogism: transparent policies are inspectable, inspectable policies are trustworthy, therefore transparent policies are safer to deploy. We stress-test the hidden steps of this syllogism in dual-source inventory control, where one product is replenished through fast, high-emission air freight or slower, lower-emission rail, under a documented *partial* Scope 3 ledger. We train PPO with dense MLP, scalar Neural Additive Model (NAM), and interaction-NAM actors whose action means decompose *exactly* into named feature and pairwise contributions, and audit them against a validation-tuned dual-index heuristic, MPC, and downside-penalized PPO on a six-condition held-out grid. We report four dissociations, each with hierarchical intervals. *Faithfulness does not warn of fragility*: the interaction-NAM falls 0.30–0.36 service short of the structured heuristic on early-long spikes and low initial stock, with intervals excluding zero in ten of twelve condition cells. The dense MLP fails too, and the heuristic also beats every learned policy on the policies' *own* training objective, so this is under-optimization rather than a price paid for transparency. Nothing in the exact decomposition tells the reader that. *Faithfulness does not imply an identifiable explanation*: one of the six domain-specified interaction terms is identically zero not because the policy ignores it but because its two features have *disjoint support*, so the zero-anchored term cannot be non-zero on any reachable state. That structural tie inflates the headline seed-stability statistic (0.79–0.89 Spearman), and removing it alone costs 0.09–0.15. When we restrict to the terms that carry magnitude, we find 0.44–0.64, above chance (permutation $p < 0.04$) but far below what the headline conveys. Agreement on the single quantity a practitioner reads first, which interaction is largest, does not significantly exceed chance in four of five cells. *Faithfulness does not pin down the explanation*: reported importances average $|g|$ over a state distribution that the architecture does not fix and exactness does not constrain. Recomputing the same decomposition on the states visited by each policy, rather than on the fixed reference trajectory our own pipeline used, inverts which interaction is largest in 20 of 20 cells, moves one term by $435\times$, and shifts the interaction share by up to $2.6\times$, while the rank correlation between the two explanations is a reassuring 0.83. *Attributed magnitude is not evidence about operational effect*: attributed share (0.9–2.9%) shows no positive association with the measured effect of including those terms, which this study cannot resolve at all ($+0.069$ service, interval $[-0.074, +0.191]$). Exact faithfulness is a property of how an output is written down; it certifies neither the robustness, the identifiability, the well-posedness, nor the operational relevance of what is written.

1 INTRODUCTION

Expedited supply protects customer service when demand or supply is disrupted, but it usually costs more and emits more than planned replenishment. Such choices suit reinforcement learning (RL), since the action must adapt to inventory, backlog, pipeline arrivals, and demand history. A high-reward policy can be hard to audit precisely where oversight matters most, so a growing line of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

work replaces opaque actors with transparent ones. Neural Additive Models (NAMs) are a prominent instance. Their outputs can be written as named feature-wise contributions (Agarwal et al., 2021; Siems et al., 2023), and this exact decomposition is often read as evidence that the resulting policy is safer to deploy.

Two steps in that inference are rarely tested together. The first holds that a *faithful* decomposition, one reconstructing the action exactly from named terms, tells the reader something reliable about robustness. The second holds that the *content* of the decomposition, meaning which features and interactions matter and by how much, is itself identifiable, stable, and operationally meaningful. Faithfulness in the sense of exact reconstruction is an architectural property of how an output is written down (Jacovi & Goldberg, 2020). It carries no guarantee about control performance, and the attribution literature has shown elsewhere that an explanation can be exact yet fragile or uninformative (Adebayo et al., 2018; Rudin, 2019). We ask whether both steps hold in a concrete, auditable operations setting. Each fails in this audited setting, in a different way.

Our simulator has three inventory stages and two raw-material modes, slow rail and fast air. It logs operating profit, on-time service, and a documented *partial* Scope 3 ledger for purchased material and freight (Figure 1). We train dense, scalar-NAM, and interaction-NAM PPO policies and evaluate them under six held-out disruptions against a tuned dual-index heuristic and MPC baselines. We then audit the explanations the transparent policies produce, as a practitioner would, with the same statistical care we apply to the control results.

The paper makes five scoped contributions.

1. **An auditable testbed.** A dual-source inventory environment coupling operational disruptions with a partial, source-documented emissions ledger. The contribution is the evaluation setting and its accounting boundary, not a claim that carbon-aware dual sourcing is new.
2. **A robustness dissociation, correctly attributed.** Exactly-faithful transparent policies *and* the dense MLP fail to match a tuned structured heuristic under disruption; the logged training reward shows the cause is under-optimization rather than additive structure, so the finding is that faithfulness gives no warning of fragility, not that transparency costs robustness (Section 6.1).
3. **The problem of non-identifiability, and a test for it.** One domain-specified interaction cannot be non-zero, because its features never co-occur off-anchor. We prove this and provide empirical checks consistent with it, trace how it inflates the reported stability statistic, and give the support check that would have caught it (Section 6.2).
4. **A failure of well-posedness.** Faithfulness leaves undeclared the state distribution over which every reported importance is averaged. Recomputing the same decomposition on the policy's own states instead of a fixed reference trajectory inverts the top-ranked interaction in all twenty cells, while the rank-agreement statistic that ought to detect this remains high (Section 6.3).
5. **A dissociation of magnitude and effect.** There is no positive association between attributed magnitude and measured operational effect across the five cells evaluated (Section 6.4).

We mean something narrower than a claim of controller superiority. The study considers two synthetic configurations, a fixed interaction set, a fixed training budget, and the disruptions described here. Its value lies in separating the *form* of an explanation from its *content*, and its *reported precision* from its *evidential content*, under a protocol that exposes all three.

2 Background and related work

Interpretable policies and the faithfulness–usefulness gap. NAMs represent an output as a sum of univariate learned functions, enabling direct inspection of a feature contribution (Agarwal et al., 2021); Siems et al. (2023) apply NAM policies to multi-echelon inventory management. Explainable-RL surveys distinguish transparent policy classes from post-hoc explanation methods (Puiutta & Veith, 2020; Heuillet et al., 2021). A recurring theme is that an explanation can be technically correct yet unreliable as a basis for decisions: faithfulness must be defined and evaluated rather than assumed (Jacovi & Goldberg, 2020), saliency attributions can be insensitive to the model they

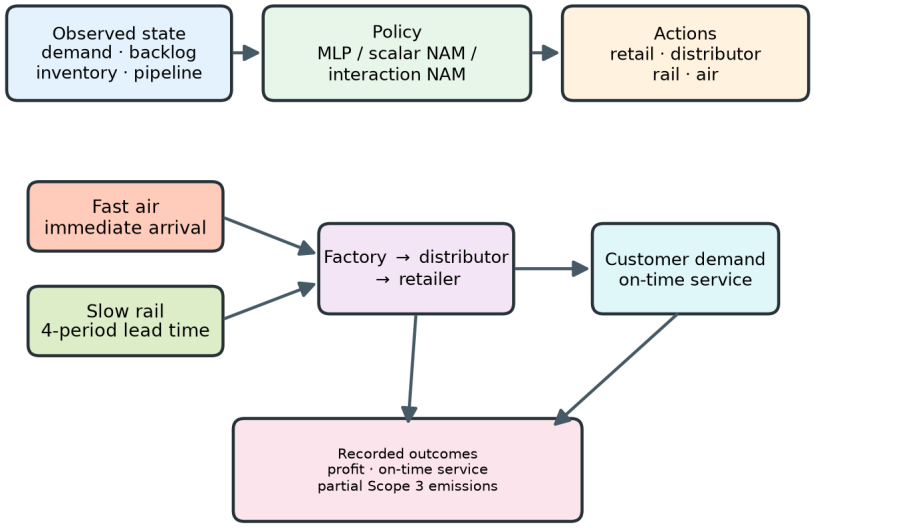


Figure 1: Dual-source inventory decision setting. The policy chooses four orders each period. Rail replenishment has a four-period lead time; air is immediate and has higher freight cost and emissions. The evaluator logs profit, on-time service, and the partial emissions ledger under held-out shifts.

explain (Adebayo et al., 2018), and inherently interpretable models are advocated precisely because post-hoc transparency can mislead (Rudin, 2019). We show that an *inherently* interpretable, exactly-decomposable policy exhibits failures of the same family: unidentifiable terms, tie-inflated stability, distribution-dependent rankings, and magnitudes that do not track effect. All four are detectable with the checks used in this audit.

Dual sourcing and carbon-aware inventory. Dual sourcing and transport-mode choice are established operations-research problems. Tailored base-surge and dual-index policies exploit the differing speed and cost of regular and expedited sources (Allon & Van Mieghem, 2010); emissions-aware mode selection has been studied directly (Drent et al., 2023). Recent work considers disruptions and deep RL in dual sourcing (Hamdan et al., 2025; Liu et al., 2025) and structure-informed RL for inventory control (Maggiar et al., 2025). Deployment-oriented RL further motivates evaluation under disturbances (Dulac-Arnold et al., 2019). We claim no novelty for carbon-aware sourcing, dual-index policies, or carbon penalties. Our question is whether an explicitly decomposable policy remains a reliable choice, and a trustworthy explainer, under a transparent audit.

3 Dual-source inventory environment

Dynamics and decisions. The finite-horizon simulator has retailer, distributor, and factory stages for one product. At period t the controller observes recent demand, retailer backlog, stage inventories, padded arrival schedules, time fraction, and a constant sentiment feature, and orders retailer replenishment, distributor replenishment, and slow and fast raw material. Slow material arrives after four periods, fast immediately; a supplier outage blocks slow ordering while the emergency route remains. Service is FIFO on-time demand fill, a stringent metric that does not credit late recovery.

One detail matters for Section 6.2. Retailer backlog is unmet retail demand, so by construction it becomes positive only once retailer inventory is exhausted. Backlog and retailer inventory therefore have *disjoint support*, a structural property of the dynamics that we exploit as a diagnostic rather than treat as a nuisance.

Carbon boundary. The ledger covers Category 1 purchased goods and Category 4 contracted freight. For purchased material we apply the EPA NAICS 325211 factor of 1.022 kg $\mathrm{CO_2e}$ per 2022 USD together with a documented synthetic material price of \$0.50 per unit. Rail, air, and truck factors come from the EPA GHG Emission Factors Hub (U.S. Environmental Protection Agency, 2024; 2025), and the freight factors are tank-to-wheel combustion values, the aircraft factor carrying no separately tabulated $\mathrm{CH_4}$ line. Left out are warehousing and production energy, upstream fuel, allocation to real firms, and any claim to a full corporate footprint. What we have is an accounting model for this simulator, not corporate reporting.

For each period, the balanced training reward is

$$r_t = \mathrm{profit}_t - 2\,\mathrm{unserved}_t - 0.5\,\mathrm{kgCO2e}_t. \tag{1}$$

We also report cost, resilience, and carbon objective variants; extended robustness comparisons use the balanced objective. The high-backlog configuration raises the per-period backlog penalty from 0.1 to 1.0. This reward is profit-weighted relative to the on-time service metric on which robustness is judged; Section 6.1 tests whether that mismatch explains the service gap and finds that it does not.

4 Policies, explanations, and baselines

We train PPO (Schulman et al., 2017) with a dense MLP, scalar NAM, or interaction-NAM actor. For action a and state x, the scalar NAM mean is

$$\mu_a(x) = b_a + \sum_j f_{j,a}(x_j). \tag{2}$$

The interaction actor adds six domain-specified pairs,

$$\mu_a(x) = b_a + \sum_j f_{j,a}(x_j) + \sum_{(i,j)\in P} g_{ij,a}(x_i, x_j), \tag{3}$$

where P contains lagged demand $\times$ retailer inventory, backlog $\times$ retailer inventory, retailer $\times$ distributor inventory, distributor $\times$ factory inventory, lagged demand $\times$ factory inventory, and backlog $\times$ factory inventory. We zero-anchor an interaction as

$$g_{ij,a}(x_i, x_j) = h_{ij,a}(x_i, x_j) - h_{ij,a}(x_i, 0) - h_{ij,a}(0, x_j) + h_{ij,a}(0, 0), \tag{4}$$

so it vanishes at either physical zero anchor. Equations 2–3 reconstruct the unconstrained action mean exactly; action clipping occurs after this mean and is assigned to no learned term. Exact reconstruction is a *definitional* property of an additive actor, so we treat zero reconstruction error as a code-correctness check rather than a finding. The substantive questions are whether the interaction terms are identifiable, whether their attribution is stable and well-posed, and whether they matter operationally (Sections 6.2–6.4).

The structured baseline is a validation-tuned dual-index rule, compared with nominal receding-horizon linear MPC observing the current state and a lagged-demand forecast but not future shocks, soft 0.80 and 0.95 service-target MPC variants, and matched-budget downside-penalized PPO. These are comparisons, not optimal-control certificates: the MPC is a linear surrogate, the heuristic is tuned only on validation data, and the absence of a base-stock floor and hindsight ceiling bounds how far the comparison can be pushed.

5 Experimental protocol

Each learned architecture has five independent training seeds and 200,704 environment transitions per seed. MLP, scalar NAM, and interaction-NAM have 46,649, 46,695, and 46,443 parameters; mean training times are 65, 110, and 171 seconds, so matching parameters and transitions does not make compute identical. Validation chooses checkpoints and heuristic/MPC settings without using held-out results.

The audit evaluates 50 held-out episode seeds per condition: early-long demand spike, severe spike, severe lead-time shock, long supplier outage, low initial stock, and a longer horizon. All contrasts use a hierarchical bootstrap (10^4 resamples) resampling training seeds independently per policy and

Table 1: Original held-out demand spike. Higher profit/service is better, lower emissions better. The dense MLP, not only the transparent policies, trails the heuristic.

Policy	Profit	Service	Shock service	kg CO_2e
Dual-index	3022	**0.914**	**0.781**	**1919**
MLP PPO	3176	0.680	0.407	2102
Scalar NAM PPO	3098	0.641	0.362	2134
Interaction-NAM PPO	2982	0.642	0.392	2031
Nominal MPC	**3407**	0.425	0.019	2927

episode seeds jointly, so an architecture contrast is paired at the episode level and carries training-seed variance. We apply this to every architecture contrast we discuss, since asserting equivalence from point estimates alone is a failure mode we specifically avoid, and we test rank-agreement statistics against explicit permutation nulls.

For explanation content we report (i) per-term mean absolute interaction contributions, (ii) the interaction share of total additive magnitude, (iii) seed-pair Spearman correlation of interaction-importance rankings as a function of how many top-magnitude terms are retained, with a bootstrap interval over training seeds and a permutation null, and (iv) a cutoff-free top-weighted rank statistic. These are attribution-stability metrics, not causal feature importance. Every such quantity is an average over a state distribution. Unless stated otherwise we report the original convention, a fixed reference trajectory, and Section 6.3 measures how much that choice matters by recomputing all of them on the policies' own states.

6 Results

6.1 Exact faithfulness does not warn of fragility

The nominal result already favors the structured baseline. On the original held-out demand spike (Table 1) the dual-index rule has substantially higher service and lower emissions than *every* learned policy, including the dense MLP; interaction-NAM improves shock service over nominal MPC but not over the heuristic. These numbers do not support a learned-policy superiority claim.

The robustness gap is stable across held-out episodes. Appendix D summarizes service across six shifts and both configurations. In the original configuration, Interaction-NAM is below dual-index by 0.303 service on an early-long spike (95% interval $[-0.474, -0.189]$), 0.320 with low initial stock ($[-0.487, -0.206]$), and 0.133 on severe spikes ($[-0.241, -0.047]$). The high-backlog configuration repeats the early-long, low-stock, and severe-spike deficits: 0.306, 0.356, and 0.111. Interaction-NAM is closest under the severe lead-time shock, but still negative in the original configuration ($[-0.171, -0.035]$). Downside-penalized PPO is worse than the other learned methods and often diverges into large backlog costs (Appendix D); we report it as a failed rather than a strong robust-RL baseline.

What the gap does and does not attribute. Two design facts bound the interpretation. *First*, the dense MLP also trails the heuristic, so the gap is not specific to additive structure and we do not claim that "interpretability costs robustness" as a policy-class property. In the high-backlog configuration the transparent actor in fact *outperforms* the dense MLP by 0.201 service ($[+0.002, +0.417]$, Appendix E); we do not lean on this, since the interval barely excludes zero among several unadjusted contrasts and the high-backlog MLP runs are degenerate (episodic reward -9300 to $-75{,}000$), making it MLP training collapse rather than an interpretability benefit. *Second*, the gap might be a scalarization artifact, since the balanced reward (Eq. 1) is profit-weighted relative to the service metric while the heuristic is tuned toward service. The logs rule this out. On *episodic training reward*, the policies' own objective, the dual-index rule scores higher than every learned policy in all twelve condition $\times$ configuration cells (Appendix F). Beaten on the quantity they optimized, the learned policies exhibit an optimization-and-budget failure, not an objective–metric mismatch. The defensible claim is narrow and not about interpretability: *within this budget PPO does not reach a hand-tuned*

structured rule on its own objective, transparent or dense, and the transparency apparatus gives no warning of this. Ten of twelve service contrasts have intervals excluding zero (Appendix D).

6.2 FAITHFULNESS DOES NOT IMPLY AN IDENTIFIABLE EXPLANATION

One domain-specified interaction cannot be non-zero. The decomposition reports backlog $\times$ retailer inventory as numerically zero in every seed, objective, and configuration: 10^{-12} in the original and 10^{-10} in high-backlog, floating-point noise around an exact zero. The natural reading, "the policy learned this joint effect does not matter," is wrong. Backlog becomes positive only once retailer inventory is exhausted (Section 3), so the features never occur off-anchor together; under Eq. 4, if $x_j = 0$ whenever $x_i \neq 0$ then $g_{ij,a} \equiv 0$ on the whole reachable state space whatever $h_{ij,a}$ learned. The term is *unidentifiable by construction*. Two empirical checks agree (Appendix B): a random-policy rollout records zero within-period co-occurrences, and across all $11{,}100$ grid episodes the count of backlog-positive periods never exceeds the count of zero-inventory periods. Invariance across seeds, objectives, and configurations is the signature of a structural zero; the neighbouring small term backlog $\times$ factory inventory varies over 10^{-4}–10^{-3} as a learned quantity does.

Rarity does not explain this. Under the trained policies backlog is non-zero in a mean of 46.7 periods per episode, so the feature is well exercised even though the *pair* is not. Nothing in six named terms tells a practitioner that one of them was never a free parameter, and our domain-based selection of the pairs did not check this.

The headline stability statistic is inflated by that tie. The mean pairwise Spearman correlation of interaction-importance rankings is 0.863 ($[0.771, 0.943]$) across five seeds in the original balanced configuration and 0.823 ($[0.729, 0.929]$) in high-backlog. Both read as "reasonably stable." But a constant-zero term is ranked last by every seed, so it contributes a guaranteed concordant pair to every seed-pair correlation. Removing that one term, and nothing else, drops the statistic to 0.64–0.80, so the structural tie alone accounts for 0.09–0.15 of the headline. The remainder is carried by the other low-magnitude terms. Figure 2a shows mean seed-pair Spearman climbing from between -0.20 and 0.20 at $k = 2$ to 0.79–0.89 at $k = 6$ as they are added back, and in every cell, at every $k < 6$, the statistic lies below its $k = 6$ value. Agreement is concentrated in the tail rather than among the terms that carry weight, and the effect is not an artifact of the particular cutoff chosen.

What is and is not reproducible. Restricting to the four terms carrying non-negligible magnitude gives mean seed-pair Spearman 0.44–0.64 (Table 2). This agreement is real, since a permutation test against five seeds ranking four terms at random rejects chance ordering in every cell ($p = 0.006$–0.031). It is nonetheless much weaker than advertised and imprecisely estimated, with bootstrap intervals as wide as $[0.00, 1.00]$. A cutoff-free top-weighted statistic on the five identifiable terms agrees, at 0.46–0.65.

The quantity a practitioner reads first fares worse. Asked which interaction is largest, seeds return the modal answer in 60% of cases in four of five cells, not significantly above chance for five seeds over six terms ($p = 0.21$). Only carbon, at 80%, exceeds chance ($p = 0.02$), and as the largest of five cells chosen post hoc it does not survive Holm correction ($p \approx 0.10$). The decomposition thus supports a weak claim about the overall ordering and not the headline claim at all. A reader taking 0.86 to mean "the ranking reproduces" would be wrong about the ranking, and the top-ranked interaction is not demonstrably reproducible above chance in four of five cells.

6.3 FAITHFULNESS DOES NOT PIN DOWN THE EXPLANATION

Equations 2–4 hold pointwise: on *every* state the terms sum to the action mean, and reconstruction error is exactly zero on all five distributions below. But the numbers a practitioner reads, meaning per-term importances, the interaction share and the ranking, are not pointwise. They are averages of $|g_{ij,a}|$ over a state distribution that the architecture does not determine and faithfulness does not touch. Exactness is silent about where the average is taken, and the choice is easy to make without noticing. Our own pipeline evaluates every importance on a fixed reference trajectory, a constant action $(20, 20, 15, 5)$ under the undisrupted setting with shocks off. The convention looks reasonable, being deterministic, policy-independent and therefore "comparable," and it is precisely wrong, because the policy being explained never generated those states. We recomputed the identical

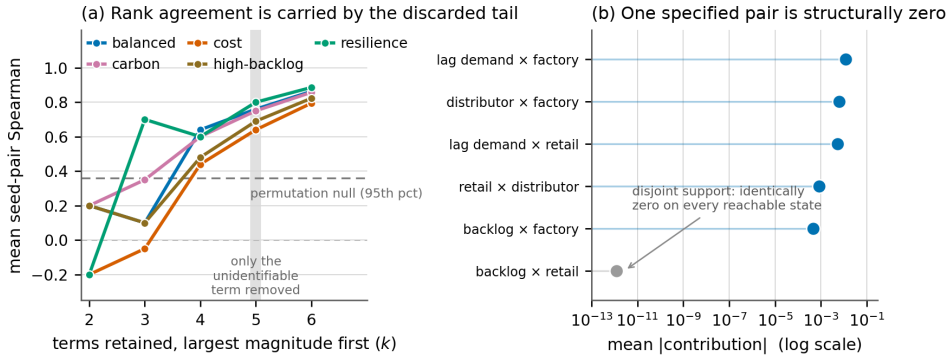


Figure 2: Attribution dissociation. (a) Mean seed-pair Spearman as a function of the number of top-magnitude terms retained. Agreement rises sharply as near-zero terms are added back, so the headline statistic is carried by the discarded tail. The shaded column at $k = 5$ removes only the unidentifiable term, which alone costs 0.09–0.15. Restricted to the terms carrying magnitude ($k = 4$), agreement remains above the permutation null (dashed; $p < 0.04$ in every cell) but far below the headline. (b) Per-term mean absolute contribution (log scale, balanced objective). Backlog × retailer inventory is structurally zero: its two features have disjoint support, so the zero-anchored term vanishes on every reachable state.

Table 2: Attribution stability across five training seeds, on the original reference distribution. "Inter./main" is the interaction share of additive magnitude (mean ± s.d.). Spearman columns give mean seed-pair rank agreement over all six terms, over the five identifiable terms (structurally zero pair removed), and over the four terms carrying magnitude, the last with a bootstrap interval over training seeds and a permutation p. "Top-1" is the fraction of seeds sharing the modal largest interaction; under the null its mean is 0.43 and 0.60 is reached 21% of the time. A cutoff-free top-weighted statistic agrees with the top-4 column (Appendix C).

			Spearman			
Config.	Objective	Inter./main	all 6	5 ident.	top-4 [95%] (p)	Top-1 (p)
Original	Cost	.029 ± .013	.79	.64	.44 [.03, .80] (.031)	.60 (.21)
Original	Resilience	.017 ± .007	.89	.80	.60 [.30, .80] (.007)	.60 (.21)
Original	Carbon	.018 ± .007	.86	.75	.60 [.00, 1.00] (.007)	.80 (.02)
Original	Balanced	.016 ± .005	.86	.76	.64 [.37, .90] (.006)	.60 (.21)
High backlog	Balanced	.009 ± .002	.82	.69	.48 [.20, .89] (.023)	.60 (.21)

decomposition from the same checkpoints on the states each policy visits, holding configuration and episode seeds fixed so that *only the action source differs*.

The ranking inverts. On the reference trajectory the largest interaction is lagged demand × factory inventory; on the policy's own states it is lagged demand × retailer inventory, 6.5× larger there (Figure 3a). Across five cells and four on-policy distributions the modal largest interaction differs from the reference answer in *all twenty* comparisons, always in the same direction. Backlog × factory inventory moves 435×, for the reason Section 6.2's diagnostic anticipates: the constant action drives backlog off-anchor in 16.3% of periods against the trained policy's 0.17%, and in 24.3% under demand spikes, where the robustness claims live. The attributed share moves with it, by up to 2.6× within a cell (Figure 3b).

The statistic that should catch this does not. Rank agreement between the reference and on-policy explanations is 0.83 mean Spearman, within noise of the 0.86 seed-stability figure Section 6.2 already showed to be tie-inflated. The aggregate stays reassuring while the quantity it is read to certify flips in 20 of 20 cases. Rank correlation over six terms cannot see a swap at the top when the tail is long and stable, which is the regime these decompositions produce.

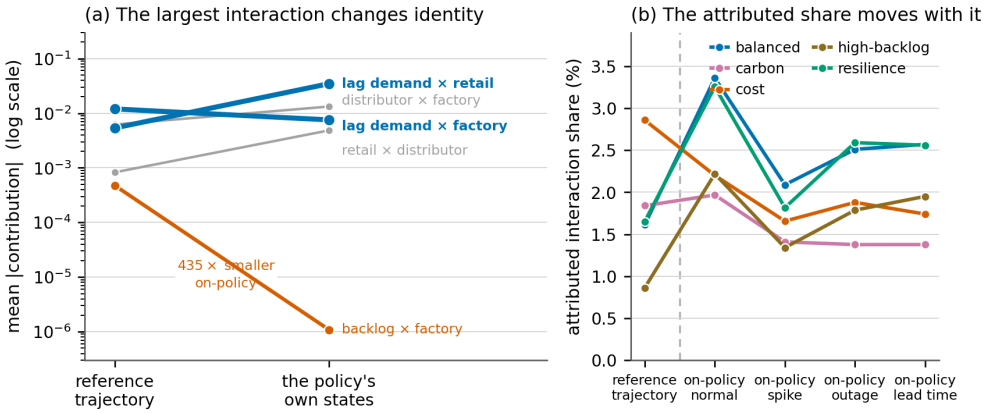


Figure 3: The same exactly-faithful decomposition evaluated in two places (balanced objective; identical configuration and episode seeds, only the actions differ). (a) The two largest terms exchange rank, and backlog × factory inventory falls $435\times$ because the constant reference action drives backlog off-anchor $98\times$ more often than the trained policy does. The structurally zero pair is omitted; see Figure 2b. (b) Attributed interaction share by evaluation distribution, all five cells.

The point is not that this particular convention is bad. A reader may object that a constant-action probe is an obviously poor choice. Two things answer that. It was our own original pipeline, adopted for reasons that look principled, since a deterministic and policy-independent probe makes importances "comparable" across seeds and architectures. That is precisely why the failure is worth reporting rather than a strawman we built. The sensitivity also does not disappear once the reference convention is abandoned. Among the four *on-policy* distributions alone, the attributed share moves by $1.6\times$ within the balanced cell and the four-term rank statistic ranges from 0.48 under normal operation to -0.08 under demand spikes in the cost cell. We report honestly that those four agree with each other on which term is largest, so the inversion itself is reference-versus-on-policy; the magnitudes and the stability are unstable among them regardless. There is no distribution-free reading of these numbers, only a declared one.

This qualifies our own analysis and survives it. Table 2 uses the reference set, so we recomputed it on all four on-policy distributions (Appendix C). Section 6.2's conclusions do not depend on the choice and sharpen under it. The six-term minus four-term Spearman gap, averaged over cells, is 0.28–0.40 across the four on-policy distributions against 0.29 on the reference set, and on the demand-spike distribution the four-term statistic falls to -0.08–0.50. The ranking among terms that carry weight is least reproducible exactly where it would be consulted.

An attributed importance is a property of a model *and* a distribution. Faithfulness fixes the first argument exactly and leaves the second undeclared, which is how a decomposition can be at once exact, seed-reproducible, and about the wrong states. The fix is cheap. Declare the distribution, and average over the policy's own states under the conditions at issue. What the fix is not is discoverable by inspecting the explanation, which looks equally clean either way. We found it by recomputing the explanation somewhere else.

6.4 Attributed magnitude is not evidence about operational effect

The decomposition reports that the six interactions carry 0.9–2.9% of the additive magnitude, and it is tempting to read a small share as "these terms do little." That reading is directly testable, by comparing separately trained scalar NAM and interaction-NAM policies. Their held-out difference measures an architecture contrast, rather than an ablation of a fixed policy.

That interpretation is not supported. Figure 6 (Appendix E) shows attributed share against measured effect across the five cells. There is no positive association. The rank correlation is -0.70, but with five cells ($p = 0.19$) this establishes neither the direction nor the absence of an association, and the cells are not fully exchangeable. The design supports the weak claim: *the attributed share*

provides no evidence of operational effect, enough to undermine reading it as one. The cell with the largest share (cost, 2.9%) has an effect indistinguishable from zero and negative in point estimate (-0.004, $[-0.036, +0.028]$); the cell with the smallest share (high-backlog, 0.9%) has the largest effect ($+0.058$, $[-0.015, +0.120]$). The two-significant-figure presentation hides the fact that the shares themselves are imprecise across seeds (cost, 0.029 ± 0.013).

The explanation is more precise than the evidence. On the held-out robustness grid the same contrast is $+0.005$ ($[-0.119, +0.119]$) in the original configuration and $+0.069$ ($[-0.074, +0.191]$) in high-backlog, where the point estimate favors the interaction actor in all six conditions but only one interval excludes zero (Appendix E). What we can conclude is limited: *at this sample size, the operational effect remains unresolved*, with the interval ranging from mildly harmful to substantially beneficial.

Symmetrically, the equivalence that near-zero magnitude invites is not supported: a 0.19 service difference cannot be ruled out either. That is the dissociation. The explanation gives contributions to two significant figures and a stability coefficient of 0.86, while the operational quantity these numbers seem to describe is unresolved in the same experiment. The precision comes from the arithmetic: an additive model yields exact shares, but these are not measurements of operational effect. A practitioner reading 1.6% as an effect size is reading evidence that is not there.

6.5 Reported emissions are a function of accounting, not policy

For frozen policies, we recompute the ledger under factor multipliers of 0.5 and 2.0, and light/short versus heavy/long mass-and-distance assumptions. Reported kg CO_2e changes by up to $2\times$ while profit and service remain unchanged (Appendix G). This is accounting sensitivity, not adaptive carbon control, since factors are not present in the observations and policies are not retrained. It is the lesson of Section 6.3 in a second register: an inspectable number looks precise while resting on a modeling choice.

7 Limitations and conclusion

This study is bounded: one synthetic domain, two configurations differing only in a backlog penalty, a partial Scope 3 ledger, six interaction terms, a modest fixed budget, and attribution stability rather than causal importance. The MPC is a linear surrogate. We do not include a base-stock floor or a hindsight ceiling, which would separate "interpretability cost" from "under-budget RL" more sharply than we can. Claims about what a practitioner would conclude follow from properties of the artifact rather than from measurements of human inference; a reader study is the natural next step. Nothing here is meant to suggest that NAM explanations are generally unhelpful, that a real firm has this footprint, or that dual-index policies dominate RL.

In this setting, an exactly decomposable actor still fails four different tests: it does not match a tuned heuristic under disruption and gives no warning of it; one of its six named interactions could never have been non-zero, and the stability statistic certifying the explanation is inflated by exactly that term; the explanation is not well-posed until an evaluation distribution is declared, and the conventional choice inverts the headline ranking in all twenty cells while the diagnostic statistic stays high; and the attributed magnitudes show no positive association with an operational effect this experiment cannot resolve.

The constructive reading is a checklist on which our own protocol would have failed on every line. Check that the features of a domain-specified interaction have overlapping support, or the term is an artifact of the anchoring. Declare the distribution an importance averages over, preferring the policy's own states under the conditions at issue. A decomposition that is exact everywhere is still a summary of somewhere. Report rank stability against a permutation null and as a function of terms retained, so that ties do not certify agreement that does not exist. Measure an effect before reading an attributed share as one: the share is arithmetic, the effect is evidence. Exact faithfulness is necessary for inspectability and not sufficient for robustness, identifiability, well-posedness, or relevance.

ETHICS STATEMENT

This work uses a synthetic simulator and no human data. It makes no deployment recommendation or real-company emissions claim. Its main practical risk is overinterpreting a partial accounting boundary or a synthetic robustness result; Sections 3 and 6 state these boundaries explicitly.

REPRODUCIBILITY STATEMENT

The supplementary materials contain the manuscript source, figure assets, and plotting scripts. The simulator, training configurations, seed banks, checkpoints, and raw episode outputs are not included. Section 5 specifies the training budget, held-out seeds, bootstrap structure, baselines, and selection protocol; Appendix A records additional details.

AI ASSISTANCE DISCLOSURE

AI-based tools assisted with literature discovery, language editing, drafting and revising portions of the manuscript, code review, experimental planning, and software implementation. The authors are responsible for verifying the code, data sources, accounting assumptions, results, citations, and claims.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E. Hinton. Neural additive models: Interpretable machine learning with neural nets. In *Advances in Neural Information Processing Systems*, 2021.

Gad Allon and Jan A. Van Mieghem. Global dual sourcing: Tailored base-surge allocation to near-and offshore production. *Management Science*, 56(1):110–124, 2010.

Melvin Drent, Poulad Moradi, and Joachim Arts. Efficient emission reduction through dynamic supply mode selection. *European Journal of Operational Research*, 2023. doi: 10.1016/j.ejor.2023.06.004.

Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Sadeque Hamdan, Youssef Boulaksil, Kilani Ghoudi, and Younes Hamdouch. Simplicity or flexibility? dual sourcing in multi-echelon systems under disruption. *Operations Research Perspectives*, 14: 100333, 2025. doi: 10.1016/j.orp.2025.100333.

Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021. doi: 10.1016/j.knosys.2020.106685.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020.

Defeng Liu, Ying Liu, and Carson Eisenach. Deep rl dual sourcing inventory management with supply and capacity risk awareness. *arXiv preprint arXiv:2507.14446*, 2025.

Alvaro Maggiar, Sohrab Andaz, Akhil Bagaria, Carson Eisenach, Dean Foster, Omer Gottesman, and Dominique Perrault-Joncas. Structure-informed deep reinforcement learning for inventory management. *arXiv preprint arXiv:2507.22040*, 2025.

Erika Puiutta and Eric M. S. P. Veith. Explainable reinforcement learning: A survey. In *Machine Learning and Knowledge Extraction*, pp. 77–95. Springer, 2020. doi: 10.1007/978-3-030-57321-8_5.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

Julien Siems, Maximilian Schambach, Sebastian Schulze, and Johannes S. Otterbach. Interpretable reinforcement learning via neural additive models for inventory management. In *AI4ABM Workshop at ICLR*, 2023.

U.S. Environmental Protection Agency. Supply chain greenhouse gas emission factors v1.3 by naics-6, 2024. URL `https://catalog.data.gov/dataset/supply-chain-greenhouse-gas-emission-factors-v1-3-by-naics-6`.

U.S. Environmental Protection Agency. Ghg emission factors hub, 2025. URL `https://www.epa.gov/climateleadership/ghg-emission-factors-hub`.

A Expanded protocol and compute matching

The experiment uses a common observation, action transform, PPO optimization budget, checkpoint-selection procedure, and evaluation seeds. Parameter counts are 46,649, 46,695, and 46,443 for MLP, scalar NAM, and interaction-NAM, and mean training times are 65.0, 110.1, and 171.5 seconds. The wall-clock difference is a transparency cost that parameter matching alone does not erase.

The nominal MPC validation grid varies horizon (8, 16, 24), backlog multiplier (1, 3, 8), and a soft service target (none, 0.80, 0.95). Horizon 16 without a target is selected by validation. In the present linear surrogate, service-target versions reduced both validation objective and observed service, so they are reported as unsuccessful refinements rather than as stronger baselines. The MPC is a weak surrogate under spikes (shock service 0.013–0.539 across the six held-out conditions); the credible non-learned competitor is the dual-index rule, and a stochastic or robust MPC would strengthen the comparison.

All contrasts use a hierarchical bootstrap with 10^4 resamples. Training seeds are resampled with replacement independently for each policy, since training runs are independent; episode seeds are resampled once and shared across the two policies in a contrast, since both are evaluated on the same held-out episodes, which pairs the comparison and removes episode-difficulty variance from the difference.

B Support diagnostic for domain-specified interactions

For a zero-anchored interaction (Eq. 4) to be identifiable, the two features must take non-anchor values simultaneously on the reachable state space. If $x_j = 0$ whenever $x_i \neq 0$, then substituting into Eq. 4 gives $g_{ij,a} = h_{ij,a}(x_i, 0) - h_{ij,a}(x_i, 0) - h_{ij,a}(0, 0) + h_{ij,a}(0, 0) = 0$, and symmetrically in the other case, so the term vanishes identically whatever $h_{ij,a}$ has learned.

The analytic argument above is the primary evidence; two empirical checks are consistent with it, and we are explicit that neither is sufficient alone. First, a random-policy rollout of $12{,}000$ steps across the four scenario families records the two features jointly off-anchor in 0 steps. This is the only check that observes *within-period* co-occurrence directly, but it has little power, because backlog is non-zero in just 0.03% of random-policy steps. Second, under the trained policies, where backlog is well exercised and non-zero in a mean of 46.7 periods per episode, the episode logs record, for each of the $11{,}100$ grid episodes, the number of backlog-positive periods and the number of zero-inventory periods. Disjoint support implies the former never exceeds the latter, and it never does, for any policy or configuration; the counts are equal in all but 133 episodes, where zero-inventory periods exceed backlog periods by at most two. These per-episode counts do not establish within-period containment; they are consistent with the disjoint support established by the analytic argument. The failure is specific to the pair, not to the feature.

The diagnostic is cheap and we recommend it as a precondition for interpreting any anchored interaction: estimate $\Pr[x_i \neq 0 \wedge x_j \neq 0]$ on rollouts *of the trained policy* before the term enters

a ranking. Logging the joint indicator per period, which our episode logs did not, is the cheap fix that would have made our own check decisive rather than merely consistent. Terms failing it should be removed from the specification rather than reported as learned zeros, and in particular must be excluded from rank-stability statistics, where they act as guaranteed concordant pairs.

C ATTRIBUTION UNDER ALTERNATIVE EVALUATION DISTRIBUTIONS

Section 6.3 recomputes the decomposition of Eq. 4 on the states each policy visits, from the same checkpoints, holding the environment configuration and the five episode seeds fixed so that only the action source differs between the reference trajectory and the on-policy sets. Reconstruction error is exactly 0 on every distribution, as it must be: the dissociation is not a failure of faithfulness but of what faithfulness covers. As a correctness check, the pipeline reproduces the original reference importances to $< 10^{-6}$ for all twenty-five checkpoint/cell combinations before any on-policy state is generated.

Table 3 reports the consequences. Three readings matter. First, the attributed interaction share is not a property of the fitted policy alone: within the balanced cell it ranges from 1.6% to 3.4% depending only on where it is averaged. Second, the conclusions of Section 6.2 hold on every distribution and are sharpened on most: the six-term minus four-term Spearman gap, which measures how much of the reported agreement is carried by near-zero terms, averages 0.29 over cells on the reference set against 0.28–0.40 across the four on-policy sets. Third, agreement among the terms that carry magnitude is worst on the demand-spike distribution, at -0.08 for cost and 0.12 for carbon, and that is the distribution the robustness claims of Section 6.1 concern. The ranking is least reproducible exactly where it would be consulted.

The mechanism is visible in the support rates. The domain-specified pair backlog $\times$ factory inventory has both features off-anchor in 16.3% of reference periods, 0.17% of on-policy periods under all three non-spike settings, and 24.3% under demand spikes. Its attributed magnitude tracks that rate rather than anything about the policy's logic, falling $435\times$ between the reference trajectory and on-policy normal operation. A constant action is not a neutral probe: it drives the system into states the trained policy avoids, and the explanation then reports the model's behaviour there.

The cutoff-free statistic referenced in Table 2 is a top-weighted rank correlation on the five identifiable terms, which weights agreement by magnitude instead of imposing a hard k. On the reference distribution it gives 0.46, 0.64, 0.65, 0.56, and 0.52 for the cost, resilience, carbon, balanced, and high-backlog cells. These sit close to the top-4 column and well below the all-six headline, confirming that the drop is not an artifact of choosing $k = 4$.

D FULL ROBUSTNESS COMPARISONS

Table 4 reports all condition means. The corresponding hierarchical service intervals use five independent training seeds and 50 held-out episode seeds. A negative value in the final column denotes lower interaction-NAM service than dual-index.

E ARCHITECTURE CONTRASTS ON THE ROBUSTNESS GRID

Table 5 reports the architecture contrasts that Sections 6.1 and 6.4 rely on, pooled over the six held-out conditions with hierarchical intervals. Two readings are worth stating. First, the Interaction-NAM-minus-scalar-NAM contrast, which measures the effect of including interactions in the architecture, is positive in point estimate in both configurations and in all six high-backlog conditions individually, but its pooled interval contains zero, so the effect is unresolved rather than absent; symmetrically, an equivalence claim is unsupported. Second, in the high-backlog configuration the interaction-NAM exceeds the dense MLP by an interval that excludes zero, which is further evidence against reading Section 6.1 as an interpretability penalty. As noted there, though, the high-backlog MLP runs are degenerate, so this reflects MLP training collapse rather than a benefit of decomposability, and the interval barely excludes zero among several unadjusted contrasts.

Table 3: Attribution content by evaluation distribution. "Share" is the attributed interaction percentage of additive magnitude; "all 6" and "top 4" are mean seed-pair Spearman over all six terms and over the four carrying magnitude; "gap" is their difference, the portion of reported agreement carried by near-zero terms. The reference row is the original convention used in Table 2.

Cell	Evaluation distribution	Share (%)	all 6	top 4	gap
Balanced	reference trajectory	1.61	.86	.64	.22
Balanced	on-policy, normal	3.36	.83	.40	.43
Balanced	on-policy, demand spike	2.09	.62	.26	.36
Balanced	on-policy, supplier outage	2.51	.86	.50	.36
Balanced	on-policy, lead-time shock	2.57	.81	.34	.47
High backlog	reference trajectory	0.86	.82	.48	.34
High backlog	on-policy, normal	2.22	.86	.50	.36
High backlog	on-policy, demand spike	1.34	.76	.50	.26
Cost	reference trajectory	2.86	.79	.44	.35
Cost	on-policy, normal	2.20	.85	.48	.37
Cost	on-policy, demand spike	1.66	.55	−.08	.63
Resilience	reference trajectory	1.65	.89	.60	.29
Resilience	on-policy, demand spike	1.82	.81	.50	.31
Carbon	reference trajectory	1.84	.86	.60	.26
Carbon	on-policy, demand spike	1.41	.55	.12	.43

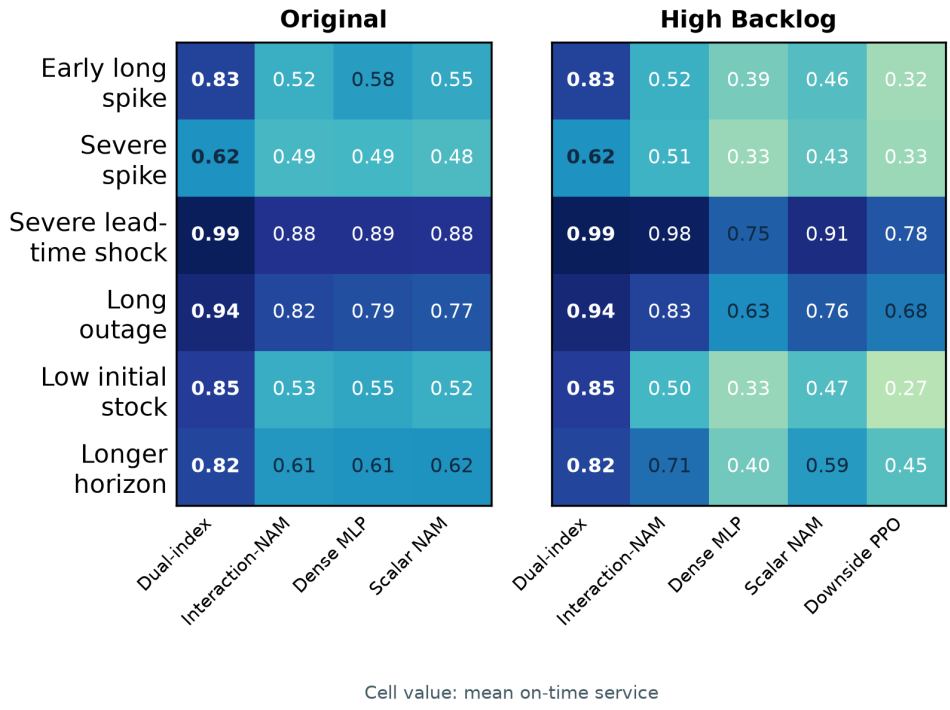


Figure 4: Mean held-out on-time service for six disruption conditions. Each cell averages five training seeds and 50 held-out episode seeds.

F THE HEURISTIC ALSO WINS ON THE TRAINING OBJECTIVE

Section 6.1 rules out a scalarization explanation for the service gap. Table 6 gives the evidence: mean episodic reward under each configuration's own training objective (Eq. 1), which is the quantity PPO maximizes and which the dual-index rule never sees. The heuristic is highest in all twelve condition × configuration cells. In the high-backlog configuration the learned policies are not merely

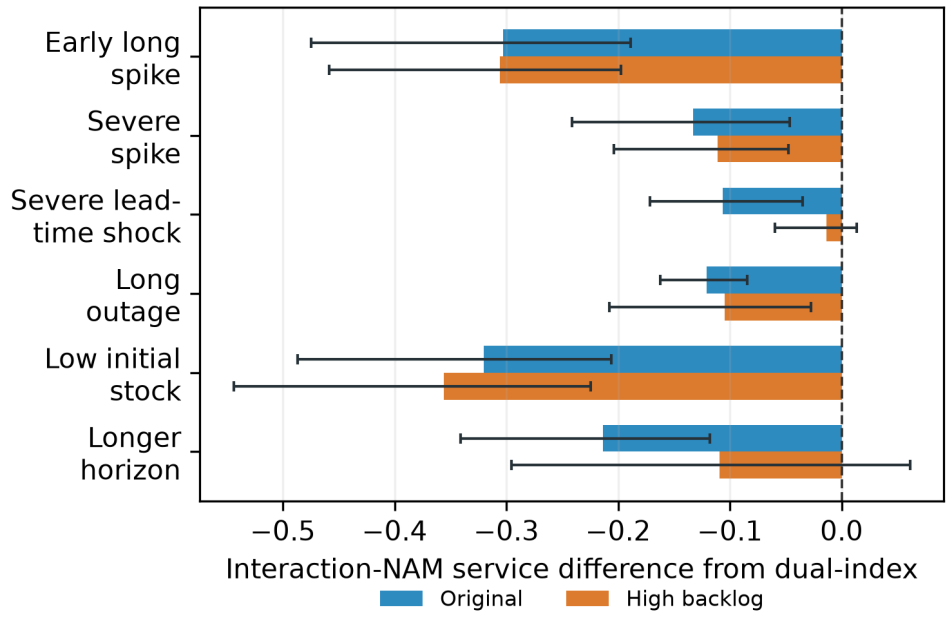


Figure 5: Interaction-NAM service gap relative to dual-index. Bars show mean gaps and whiskers show hierarchical 95% bootstrap intervals over training and episode seeds. Negative values favor dual-index.

Table 4: Held-out robustness results. Service difference is interaction-NAM minus dual-index; brackets are hierarchical 95% bootstrap intervals.

Config.	Shift	Dual svc.	Int.-NAM svc.	Δ service [95% interval]
Original	Early-long spike	.825	.522	−.303 [−.474, −.189]
Original	Severe spike	.620	.487	−.133 [−.241, −.047]
Original	Severe lead time	.990	.884	−.106 [−.171, −.035]
Original	Long outage	.936	.815	−.121 [−.162, −.084]
Original	Low initial stock	.851	.531	−.320 [−.487, −.206]
Original	Longer horizon	.819	.605	−.213 [−.341, −.118]
High backlog	Early-long spike	.825	.519	−.306 [−.459, −.197]
High backlog	Severe spike	.620	.509	−.111 [−.204, −.047]
High backlog	Severe lead time	.990	.977	−.013 [−.060, .013]
High backlog	Long outage	.936	.831	−.104 [−.208, −.027]
High backlog	Low initial stock	.851	.495	−.356 [−.544, −.225]
High backlog	Longer horizon	.819	.710	−.109 [−.295, .061]

behind but divergent, trailing the heuristic by absolute margins of 1.1×10^4 to 7.5×10^4 reward, which is also why we report downside-penalized PPO as a failed run rather than a robust-RL baseline.

The implication constrains what this paper may claim. A gap on a metric the policy was not trained for would be a scalarization artifact; a gap on the training objective itself is an optimization or budget failure. Since the dense MLP is beaten too, neither is evidence about interpretability, and our robustness dissociation should be read as: *exact decomposability did not buy robustness here*, not as: *decomposability cost robustness*.

G Carbon sensitivity details

Under the base accounting, dual-index, interaction-NAM, MLP, scalar NAM, and nominal MPC report 1924, 2035, 2105, 2139, and 2948 kg CO_2e respectively. Multiplying emission factors by two

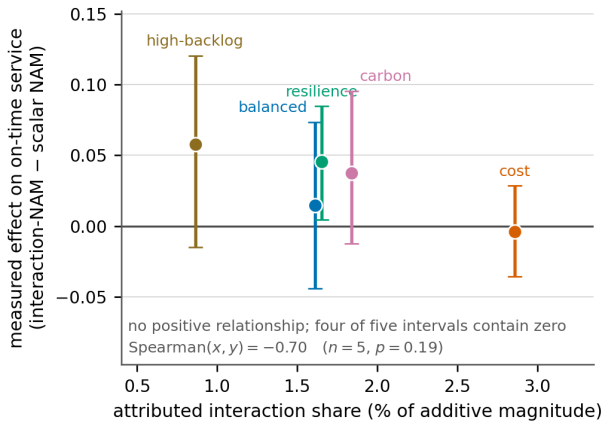


Figure 6: Attributed interaction share against the measured effect of including those terms (interaction-NAM minus scalar NAM), hierarchical 95% intervals, five objective/configuration cells. Effects pool 50 held-out episode seeds across the three disrupted scenario families. Attributed magnitude shows no positive association with operational effect across these five cells; four intervals contain zero.

Table 5: Pooled architecture contrasts on the six-condition held-out grid, on-time service, hierarchical 95% bootstrap intervals over five training and 50 episode seeds.

Config.	Contrast	Δ service	95% interval
Original	Interaction-NAM − scalar NAM	+.005	[−.119, +.119]
Original	Interaction-NAM − MLP	−.012	[−.142, +.103]
Original	scalar NAM − MLP	−.018	[−.122, +.087]
High backlog	Interaction-NAM − scalar NAM	+.069	[−.074, +.191]
High backlog	Interaction-NAM − MLP	+.201	[+.002, +.417]
High backlog	scalar NAM − MLP	+.132	[−.055, +.346]

Table 6: Mean episodic reward under the configuration's own training objective, six held-out conditions. Higher is better. The validation-tuned dual-index rule is highest in every cell despite never optimizing this objective.

Config.	Condition	Dual-index	interaction-NAM	Scalar NAM	MLP
Original	Early-long spike	**1167**	−1434	−1070	−495
Original	Severe spike	**−409**	−2325	−2404	−2053
Original	Severe lead time	**1375**	1079	1133	957
Original	Long outage	**1132**	832	583	711
Original	Low initial stock	**1080**	−1187	−1129	−782
Original	Longer horizon	**1494**	−1549	−980	−866
High backlog	Early-long spike	**351**	−15,948	−15,307	−39,582
High backlog	Severe spike	**−5549**	−20,236	−25,403	−53,037
High backlog	Severe lead time	**1349**	751	475	−9327
High backlog	Long outage	**646**	−519	−2376	−12,834
High backlog	Low initial stock	**374**	−12,693	−11,023	−33,798
High backlog	Longer horizon	**−67**	−15,205	−20,795	−75,017

produces 3848, 4070, 4209, 4279, and 5896 kg CO_2e. Under the heavy/long physical assumption, the corresponding values are 2214, 3019, 3211, 3305, and 6038 kg CO_2e. These changes demonstrate dependence on the declared accounting model; they do not alter the frozen controller actions. The base values differ by $<0.8\%$ from Table 1 because the sensitivity sweep re-evaluates on a 20-seed episode subset rather than the 50-seed held-out set.