BEYOND DECISION AGREEMENT: EVALUATING REINFORCEMENT LEARNING POLICY BEHAVIOR AND OUTCOMES IN POWER-SYSTEM BALANCING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned decision-support policies are often evaluated by their agreement with historical operator decisions. However, historical agreement does not reveal how a policy behaves when its decisions are executed sequentially, because each decision changes the state for subsequent decisions. We study this issue in offline reinforcement learning for power-system balancing at a Belgian transmission system operator (TSO). We evaluate learned policies using three complementary measures: **Decision Agreement**, historical agreement with operator decisions; **Operational/Closed-Loop Behavior**, behavior during closed-loop execution; and **Operational Outcome**, the resulting system-level performance, measured by mean absolute Area Control Error (ACE). Across controlled changes to the reward, supervision, and decision-time deferral, we find that these measures can change independently. In particular, policies with similar historical agreement can produce substantially different closed-loop behavior, while deliberate improvements in agreement also alter closed-loop behavior. Yet, across all experiments, these behavioral differences correspond to only small, statistically undetectable changes in the resulting system imbalance. We conclude that historical decision agreement alone is insufficient to validate a learned decision-support policy. Instead, evaluating policy behavior alongside historical agreement and operational outcomes is essential for offline sequential decision support.

1 INTRODUCTION

In safety-critical domains such as electric power systems, learned decision-support policies are often developed from historical operational records because controlled experimentation with the real system is costly or unsafe. Historical operator actions provide a behavioral reference, but are not assumed to constitute ground-truth optimal decisions. Agreement with these actions can therefore measure behavioral fidelity, but it is not sufficient to establish how a learned policy behaves when its own decisions are executed sequentially, since each action changes the subsequent system state and thus the inputs to future decisions, or whether that behavior produces desirable operational consequences. This creates an evaluation gap for offline sequential decision support.

Offline reinforcement learning (RL) learns policies from fixed datasets without online exploration, while off-policy evaluation (OPE) provides tools for estimating policy performance from logged data (Kumar et al., 2022; Fu et al., 2021; Wang et al., 2024). However, these approaches do not directly address whether agreement with historical decisions predicts the operational behavior induced by a learned policy. Prior work has examined imperfect reward specifications (Pan et al., 2022) and differences in behavior among models with similar predictive performance (Hsu et al., 2024), but does not address this relationship in offline decision-support settings.

In the power systems domain, RL2Grid (Marchesini et al., 2025) standardizes environments, actions, and rewards for benchmarking RL algorithms in power-grid operations, asking how RL methods compare under realistic benchmark tasks. It focuses on benchmarking RL algorithms rather than on the relationship between historical decision agreement, operational behavior, and operational outcomes, leaving this gap open in the power-systems domain as well. We address this gap through

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.

a validation methodology for offline decision support, evaluated empirically on a balancing task of a transmission system operator (TSO).

We therefore distinguish three complementary validation dimensions: **Decision Agreement**, whether a policy reproduces historical decisions; **Operational/Closed-Loop Behavior**, the behavior induced when its decisions are executed sequentially; and **Operational Outcome**, the measured system-level consequence of that behavior. This complements existing OPE frameworks.

We study this problem in supervisory balancing decision support using real operational data from a Belgian TSO. The learned policy recommends whether and in which direction to activate manual Frequency Restoration Reserve (mFRR), while the final activation decision would remain with the system operator. Each activation decision carries a market cost, so a policy that reproduces historical labels on average but activates differently at specific decisions changes real operational cost and risk even when its agreement with history is unchanged. A transition model learned from historical operational data provides a common offline environment in which policies can be evaluated in closed loop before deployment.

We make three contributions:

- **A three-dimensional evaluation protocol for offline decision support.** We distinguish Decision Agreement, Offline Closed-Loop Behavior, and Operational Outcome as complementary validation dimensions for offline decision support.
- **Empirical evidence that these dimensions can dissociate.** A controlled reward intervention changes policy-induced behavior without a corresponding stable change in historical Decision Agreement or a change in measured Operational Outcome detectable at the evaluator's resolution. This dissociation persists when Decision Agreement is deliberately improved through targeted ambiguity-aware supervision.
- **A practical test of uncertainty-aware selective decision support.** Decision-time uncertainty is used to selectively defer locally ambiguous decisions across a range of autonomous coverage. For both CQL (conservative Q-learning) and AWAC (advantage-weighted actor-critic) and two fallback designs, deferral improves retained-decision agreement and changes closed-loop behavior, but does not produce an improvement in measured Operational Outcome.

Together, these experiments support an evaluation principle for this TSO decision-support setting: historical agreement is informative, but reliable assessment also requires evaluating the behavior induced by a learned policy and its measurable operational consequences.

2 METHODOLOGY

In this section, we formalize the decision problem and data (§2.1) and the offline RL benchmark (§2.2) used to train and compare policies along the three validation dimensions.

2.1 PROBLEM FORMULATION AND DATA

We use operational data from the Belgian transmission system operator (Elia) and the ENTSO-E Transparency Platform (Elia, 2026; ENTSO-E, 2026), harmonized into a single time series at 15-minute imbalance settlement period (ISP) resolution. The dataset covers system imbalance, balancing actions, reserve status, generation, load, market variables, and cross-border exchanges. Several variables, including Area Control Error (ACE), became available only from 22 May 2024 onward; therefore, all experiments use the window from 22 May 2024 to January 2026 rather than the full raw-data span.

We formulate supervisory TSO balancing as a discrete-time Markov decision process (MDP)

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma), \tag{1}$$

where $\mathcal{S}$ denotes the operational state space, $\mathcal{A}$ the supervisory action space, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ the transition dynamics, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, and $\gamma \in [0, 1)$ the discount factor.

The supervisory decision studied here is net mFRR activation, which represents the TSO's choice of whether and in which direction to call on manual reserves once automatic reserves alone have not

resolved an imbalance. We discretize the action into

$$a_t = \begin{cases} \text{HOLD} & \text{if } |\text{mFRR}_t^{\text{net}}| \leq \tau, \\ \text{UP} & \text{if } \text{mFRR}_t^{\text{net}} > \tau, \\ \text{DOWN} & \text{if } \text{mFRR}_t^{\text{net}} < -\tau, \end{cases} \tag{2}$$

where $\tau = 50$ MW, chosen to separate predominantly inactive periods from substantial manual activation in the observed data; Appendix A.1 reports on the sensitivity to this choice. The resulting activation volume varies across instances of the same action – in the observed data, net mFRR averages 166 MW for UP and 156 MW for DOWN – because volume is set by merit-order bid clearing rather than chosen bid-by-bid (Elia, 2026): only the activation direction is the discretionary decision modeled here; the transition model and policy use only this direction, while realized volume enters solely two reward primitives. Appendix A.2 provides further detail on this.

For policy learning, a 77-dimensional operational state combines physical system indicators, including deviation, reserve activation, cross-border flows, generation, and infrastructure variables, with their short-term trends; market signals, including forecasts and lagged prices; a short-horizon, pre-decision-only imbalance forecast; and a stress/regime summary obtained through PCA reduction and Gaussian-mixture clustering. Appendix A.3 contains the complete list of features.

The data are split into training, validation, and test sets using an episode-aware 70/15/15 split, chronological at the level of activation episodes rather than individual timesteps, so that no episode is severed across splits. The resulting action distribution is highly imbalanced, with HOLD, UP, and DOWN accounting for 83.71%, 12.11%, and 4.18% of observations, respectively.

A learned transition model provides the environment for every offline closed-loop rollout. It is fitted using multi-output XGBoost (Chen & Guestrin, 2016) regression on historical state transitions, with stochasticity represented by resampling empirical residuals conditioned on the current stress regime. The model achieves a one-step test-set $R^2 = 0.973$ (RMSE $= 24.55$). Concretely, the deterministic component is

$$\hat{s}_{t+1} = \mathcal{T}_\theta(s_t, a_t), \tag{3}$$

where $\mathcal{T}_\theta$ is fitted on training-set transition tuples (s_t, a_t, s_{t+1}). Residuals are computed on held-out validation data and stored by stress regime, giving the generative rollout model

$$s_{t+1} = \mathcal{T}_\theta(s_t, a_t) + \epsilon_t, \qquad \epsilon_t \sim \hat{P}_{\text{regime}(s_t)}. \tag{4}$$

Here, $\hat{P}_{\text{regime}(s_t)}$ denotes the empirical distribution of validation residuals for the stress regime associated with s_t. The same transition model and stochastic residual sampling procedure are used for every policy, so all rollout-based comparisons are made under a common, disclosed environment. Appendix A.4 contains full detail on model comparisons.

Each closed-loop comparison uses 150 independent 96-step trajectories per seed, with starting states drawn once and shared across every policy so that comparisons are paired. Every significance test in §3 operates on a per-trajectory summary (e.g., a trajectory's own intervention rate or mean $|\text{ACE}|$), so the 150 trajectories, not their individual timesteps, are the unit of analysis.

2.2 Offline Policy Learning

All learned policies are trained using a common two-stage procedure. Stage 1 provides a shared initialization through behavioral cloning, while Stage 2 applies the method-specific learning objective. This separation ensures that the eight-method comparison in §3.1 evaluates differences in the learning procedure rather than differences in initialization. This comparison instantiates the three validation dimensions (§3) across training objectives.

The policy is represented by a dueling action-value network $Q_\theta(s, a)$ (Wang et al., 2016). Its softmax output defines the policy,

$$\pi_\theta(a \mid s) = \text{softmax}\big(Q_\theta(s, \cdot)\big)_a. \tag{5}$$

For the reinforcement-learning methods, Q_θ approximates the action-value function

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \,\middle|\, s_t = s,\ a_t = a \right]. \tag{6}$$

Stage 1: Behavioral cloning. We first pretrain the dueling network to reproduce the historical supervisory decisions via behavioral cloning (Pomerleau, 1988). The network is trained with focal loss (Lin et al., 2017) and class-balanced sampling to account for the strong imbalance between HOLD, UP, and DOWN decisions. The resulting checkpoint is used to initialize every Stage-2 method, providing a common representation of the historical decision policy before any reward-driven or algorithm-specific modification.

Stage 1 minimizes the focal-loss behavioral-cloning objective

$$\mathcal{L}_{\mathrm{BC}}(\theta) = \mathbb{E}_{(s,a)}\left[-(1 - p_\theta(a \mid s))^{\gamma_f} \log p_\theta(a \mid s)\right], \tag{7}$$

where the expectation is taken over class-balanced historical training samples, $\gamma_f = 2$, and $p_\theta(a \mid s) = \pi_\theta(a \mid s)$.

Stage 2: Offline policy learning. In Stage 2, all eight methods are initialized from the common Stage-1 checkpoint. Five are independently trained with their own Stage-2 learning procedure: behavioral cloning (BC), conservative Q-learning (CQL) (Kumar et al., 2020), advantage-weighted actor-critic (AWAC) (Nair et al., 2020), critic-regularized regression (CRR) (Wang et al., 2020), and a mixture-of-experts (MoE) gating approach. The remaining three are not independently trained: a five-seed Q-averaging ensemble, a confidence-aware BC/RL hybrid (Confidence Hybrid), and an adaptive per-class threshold method are each a post-hoc aggregation or decision rule computed directly from CQL's own Q-values, included as alternative ways of using an already-trained CQL model rather than as separate learned policies. Each of the eight is evaluated over five training seeds (42–46).

For the three RL-based methods (CQL, AWAC, and CRR), the usual Temporal Difference (TD) target is

$$y = r + \gamma(1 - d)\, Q_{\bar{\theta}}(s', a'), \tag{8}$$

where d indicates episode termination. Because each historical transition is treated as terminal in the main benchmark ($d = 1$), this reduces to $y = r$, giving

$$\mathcal{L}_{\mathrm{TD}}(\theta) = \mathrm{SmoothL1}\big(Q_\theta(s, a),\, r\big). \tag{9}$$

The three methods therefore share the same immediate-reward fitting term but differ in how they treat actions not taken by the operator. This makes Stage 2 closer to offline contextual-bandit learning than to multi-step value backup; sequential effects enter through the closed-loop rollout evaluation.

CQL adds an explicit pessimism penalty on alternative actions together with a Stage-2 behavioral-cloning regularizer,

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{TD}}(\theta) + \alpha\, \mathcal{L}_{\mathrm{CQL}}(\theta) + \beta\, \mathcal{L}_{\mathrm{BC}}(\theta), \tag{10}$$

where

$$\mathcal{L}_{\mathrm{CQL}}(\theta) = \mathbb{E}_s\left[\log \sum_{a' \in \mathcal{A}} \exp Q_\theta(s, a') - Q_\theta(s, a)\right]. \tag{11}$$

This is a discrete-action CQL variant with an added BC regularizer, a common practical addition for training stability. The CQL term penalizes high values assigned to alternative actions while retaining the observed action as the reference action. The Stage-2 $\mathcal{L}_{\mathrm{BC}}$ term is a standard cross-entropy loss, rather than the focal-loss objective used in Stage 1, and regularizes the policy toward the historical action distribution.

AWAC and CRR instead re-weight the observed historical action according to its estimated advantage. Specifically, the state value is computed under the current softmax policy as

$$\bar{V}_\theta(s) = \sum_{a' \in \mathcal{A}} \pi_\theta(a' \mid s)\, Q_\theta(s, a'), \tag{12}$$

with corresponding advantage

$$\hat{A}_\theta(s, a) = Q_\theta(s, a) - \bar{V}_\theta(s). \tag{13}$$

AWAC converts this advantage into an exponential per-example weight,

$$w(s, a) = \min\left(\exp\left(\frac{\hat{A}_\theta(s, a)}{\lambda}\right), C\right), \tag{14}$$

and minimizes

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{TD}}(\theta) - \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[w(s,a)\log\pi_\theta(a \mid s)\right]. \tag{15}$$

Thus, historical actions with above-average estimated advantage receive greater weight, while those with below-average advantage receive less weight; C limits the influence of any individual sample. CRR uses a binary advantage filter instead of AWAC's continuous weighting scheme. Hyperparameters α, β, λ, and C are reported in Appendix A.5.

Reward formulations. Because the operational objective is not directly observable from the historical decision record, we consider two reward formulations. R_1 is inferred from historical behavior via Bayesian optimization over five penalty primitives and is strongly biased toward low activation and inaction. R_2 is a fixed, state-independent reward assigning a small positive value to HOLD and larger values to UP and DOWN, providing a transparent intervention that removes R_1's inaction preference while holding architecture, data, initialization, and training procedure fixed. Full reward definitions and optimization details are provided in Appendix A.2.

3 RESULTS

We present our empirical results for evaluating learned policies along the three validation dimensions we propose: **Decision Agreement**, **Offline Closed-Loop Behavior**, and **Operational Outcome**. We proceed in four steps: baseline agreement (§3.1); a reward intervention (§3.2); a specificity check (§3.3); and targeted improvement and deferral (§3.4–§3.5).

3.1 HISTORICAL DECISION AGREEMENT ACROSS METHODS

Before intervening on the learning objective, we first ask what standard evaluation based on agreement with historical decisions shows on its own. Table 1 compares the eight benchmark methods introduced in §2.2 using macro F1, balanced accuracy, and per-class F1 on the held-out test split. Each stochastic method is evaluated over five seeds; the CQL-seed ensemble is a deterministic aggregation and therefore has no independent seed variation. The strongest methods achieve similar levels of historical agreement: AWAC (0.669), Confidence Hybrid (0.667), CQL (0.665), and the adaptive-threshold method (0.665) span only 0.004 in macro F1 – smaller than any individual method's own across-seed standard deviation (0.009–0.011) – so these four are not distinguishable from each other at this resolution, while all four sit clearly above BC (0.643), CRR (0.641), and MoE gating (0.593). This does not determine whether differences in Decision Agreement correspond to differences in Offline Closed-Loop Behavior or Operational Outcome, which we test next.

Table 1: Historical Decision Agreement on the held-out test split for the eight benchmark methods. Values are mean ± standard deviation over five seeds, except for the deterministic ensemble of five CQL seeds.

Method	Macro F1	Bal. Acc.	UP F1	DOWN F1	HOLD F1
Ensemble (avg. of 5 CQL seeds)	0.671	0.795	0.603	0.486	0.924
AWAC	0.669 ± 0.011	0.784 ± 0.009	0.597 ± 0.010	0.485 ± 0.018	0.925
Confidence Hybrid (BC+RL)	0.667 ± 0.009	0.784 ± 0.010	0.598 ± 0.012	0.480 ± 0.012	0.924
CQL	0.665 ± 0.011	0.792 ± 0.008	0.590 ± 0.009	0.483 ± 0.019	0.921
Adaptive threshold	0.665 ± 0.011	0.791 ± 0.008	0.590 ± 0.009	0.482 ± 0.019	0.921
BC	0.643 ± 0.004	0.722 ± 0.004	0.579 ± 0.004	0.425 ± 0.008	0.925
CRR	0.641 ± 0.024	0.807 ± 0.018	0.563 ± 0.020	0.455 ± 0.035	0.905
MoE gating	0.593 ± 0.011	0.838 ± 0.009	0.544 ± 0.027	0.371 ± 0.011	0.865

3.2 CONTROLLED REWARD INTERVENTION

To examine whether historical Decision Agreement captures changes in a learned policy's behavior and consequences, we introduce a controlled intervention on the Stage-2 reward, comparing the behavior-inferred reward R_1 with the fixed intervention reward R_2 while keeping data, architecture, and initialization fixed. R_1, optimized to reproduce historical operator behavior, is expected to favor low intervention, consistent with the historical record's strong HOLD bias; R_2 is constructed to favor intervention over inaction, so a resulting shift in behavior is expected relative to R_1; the intervention's purpose is to test whether Decision Agreement and Operational Outcome track that engineered

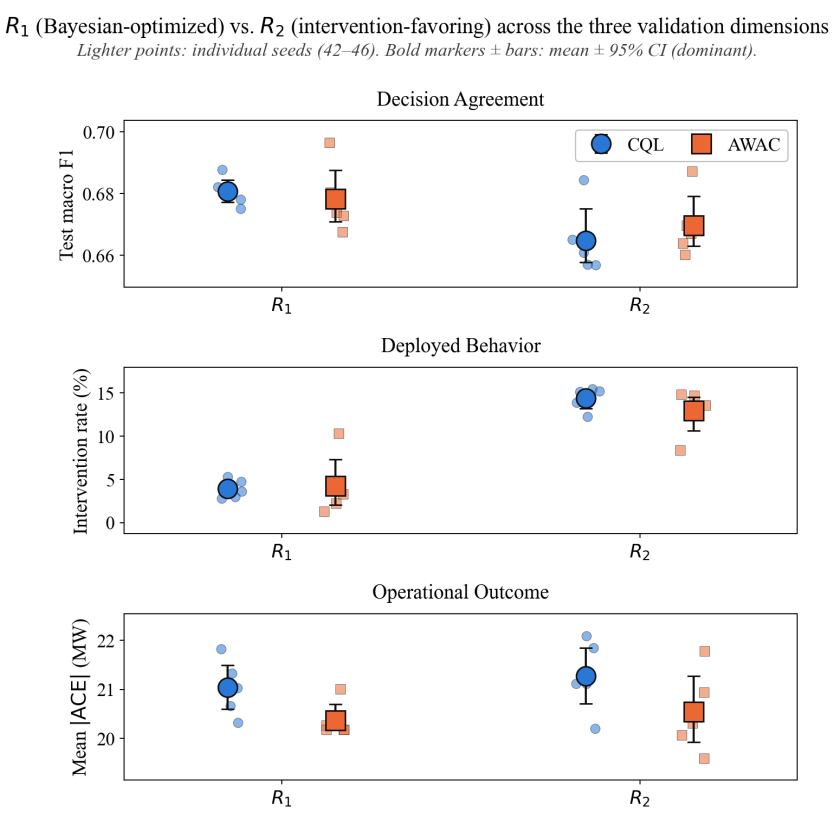

Figure 1: The controlled reward change ($R_1 \rightarrow R_2$) across the three validation dimensions. Each panel shows the R_1 and R_2 values for individual training seeds (42–46), together with the across-seed mean and 95% confidence interval (CI).

shift. CQL and AWAC are carried forward: among the five independently trained methods, they are the two strongest on Decision Agreement.

Figure 1 summarizes the intervention's effects across the three validation dimensions.

We first examine the effect on Decision Agreement. Across the five seeds, changing the reward shifts macro F1 by at most 2.7 percentage points for either algorithm. The direction is not fully consistent either: R_1 yields higher macro F1 at four of five seeds, R_2 at the remaining one. A selective-prediction check at seed 42 shows that any edge R_1 has is shallow: the accuracy gap between R_1 and R_2 narrows from 3.7 percentage points at full coverage to 0.3 points once only the most confident 70% of decisions are retained, concentrated in the easiest cases rather than the hardest ones (see Table 9, Appendix A.6).

The same intervention produces a clear, consistent shift in Offline Closed-Loop Behavior: the intervention rate under R_2 exceeds R_1 for every algorithm–seed pair, consistent with R_2's construction, which explicitly reduces R_1's strong inaction preference (full paired statistics in Table 10, Appendix A.7).

Operational Outcome, measured by simulated mean $|\mathrm{ACE}|$, shows no corresponding change: effect sizes are small and inconsistent in sign, and every 95% bootstrap CI includes zero (full paired statistics in Appendix A.7). Thus, the reward intervention produces a substantial, consistent change in Offline Closed-Loop Behavior without a change in Operational Outcome detectable at the evaluator's resolution: the outcome effect, if any, is smaller than the trajectory-to-trajectory variability in the evaluation. For a TSO, intervention frequency – which maps directly to activation cost – can therefore change without a demonstrated change in balance quality, so behavior and outcome must each be assessed directly rather than inferred from one another or from agreement.

3.3 Is the Dissociation Specific to the Reward Manipulation?

We next examine whether the dissociation observed in §3.2 extends beyond the controlled reward intervention, using two complementary analyses.

Across the five policies with closed-loop rollouts (BC, Confidence Hybrid, AWAC, CQL, CRR) plus the historical Operator (Table 2), Offline Closed-Loop Behavior varies substantially without a corresponding difference in Operational Outcome detectable at this resolution. The reported means are close, but the underlying trajectory-level distributions are not identical: CRR, the most interventionist policy, also has the largest across-trajectory variance and maximum $|\mathrm{ACE}|$ (full distribution in Table 12, Appendix A.7). This is not a repeat of the controlled experiment, but it shows the same pattern: policies can differ substantially in behavior without a corresponding difference in outcome.

Table 2: Closed-loop evaluation of five policies and the historical Operator (150 paired 96-step trajectories). Intervention rate is computed over the rollout trajectories (§2.1), not the raw historical test period; vs. Operator is the relative change in that rate (p: paired Wilcoxon). Mean $|\mathrm{ACE}|$ is averaged over all steps of each trajectory; no policy differs from the Operator in mean $|\mathrm{ACE}|$ (paired t-test, all $p > 0.27$).

Policy	Interv. rate	vs. Operator	Mean $\lvert\mathrm{ACE}\rvert$ (MW)
Operator	11.1%	—	21.20
BC	11.3%	$+1.9\%$ ($p = 0.46$)	21.04
Confidence Hybrid	13.1%	$+17.4\%$ ($p = 0.070$)	22.07
AWAC	14.1%	$+27.2\%$ ($p = 0.0032$)	21.41
CQL	15.0%	$+35.2\%$ ($p = 0.0036$)	21.11
CRR	19.1%	$+72.1\%$ ($p = 1.2 \times 10^{-8}$)	21.82

A separate, descriptive check compares historical intervention against the counterfactual of always choosing HOLD at each pre-decision imbalance-severity quintile (Q1, calmest, to Q5, most extreme imbalance). Only the most extreme-imbalance quintile (Q5) – the highest-stakes cases, where a poor decision carries the largest consequences – shows a statistically significant association favoring intervention (Appendix A.7, Figure 5).

Five generic training interventions – optimization, reward weighting, state representation, calibration, and class imbalance – left Decision Agreement essentially unchanged for either algorithm (see Appendix A.8), indicating the metric's resistance to indirect adjustment. This motivates a targeted intervention, tested next.

3.4 Ambiguity-Aware Supervision: A Targeted Intervention

Unlike the generic interventions in the previous section, ambiguity-aware supervision is a targeted intervention designed specifically to improve Decision Agreement, directly testing whether the earlier separation was simply a matter of metric difficulty. It modifies the Stage-2 behavioral-cloning target using local label ambiguity: each hard action label is blended with the empirical label distribution among its $k = 25$ nearest neighbors in proportion to that ambiguity. The intervention is evaluated for CQL and AWAC over three training seeds each, along all three validation dimensions.

Decision Agreement improves. Ambiguity-aware supervision increases macro F1 by $+2.1$ percentage points for CQL and $+1.7$ points for AWAC, driven primarily by HOLD and UP; the smaller DOWN-class gain is not similarly supported by the smoothing controls. The consequences of this improvement for policy behavior and operational outcome are shown across severity quintiles in Figure 2. The historical operator is included only as a reference; the two learned policies are evaluated through the same closed-loop rollout procedure.

Offline Closed-Loop Behavior changes substantially. Despite the Decision Agreement improvement, intervention rates decrease by approximately $2.5\times$ for both algorithms ($p < 10^{-7}$ at every seed), a reduction that persists across severity quintiles and is largest in absolute terms in Q5 – the stratum with the most extreme imbalance, and the same one where §3.3's historical-data check found the strongest associational evidence favoring intervention over HOLD.

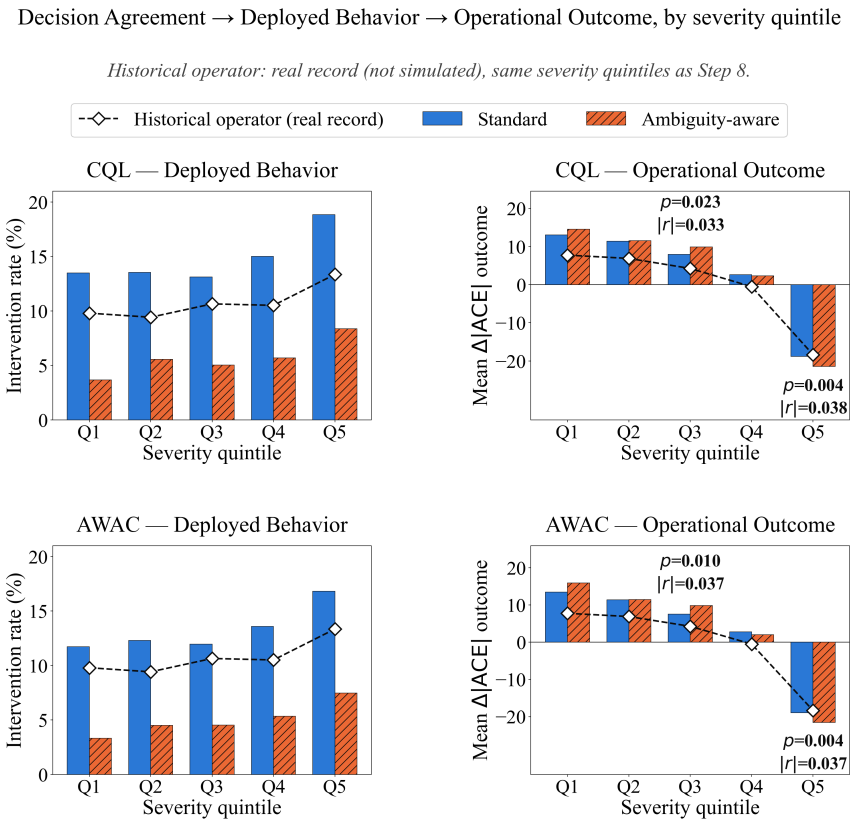

Figure 2: Ambiguity-aware versus standard training across severity quintiles for CQL and AWAC. The left panels show Offline Closed-Loop Behavior as intervention rate, while the right panels show the measured Operational Outcome as mean $\Delta|\mathrm{ACE}|$. The historical operator is shown as a reference based on the real historical record.

Operational Outcome does not establish that the behavioral change is better. We therefore evaluate whether this behavioral change is accompanied by an improvement in Operational Outcome, measured by mean $|\mathrm{ACE}|$. Across severity quintiles, differences between ambiguity-aware and standard training are small, with rank-biserial effect sizes below 0.04; some quintiles show nominal significance, but the effects are too small to provide consistent evidence of improved outcome: below the conventional small-effect threshold ($\sim$0.1) and far below the market-mechanism-driven $|\mathrm{ACE}|$ shift in Appendix A.9 (19.75 to 14.47 MW), hence not operationally meaningful (full per-seed, per-class, and per-quintile detail in Appendix A.8).

3.5 Selective Deferral: Changing Autonomous Behavior

We next test the generality of the observed dissociation by intervening directly at decision time rather than through retraining. States with high local ambiguity are selectively deferred to a fallback action while the underlying classifier remains unchanged. The ambiguity score is computed from the state using the 25 nearest historical neighbors and requires no ground truth; it predicts classification error with AUROC $= 0.836$. We sweep the target deferral rate – the fraction of states routed to a fallback instead of the policy's own decision – from 0% (no deferral) to 30%, so autonomous coverage correspondingly falls from 100% to 70%. Two fallback designs are compared: HOLD, the reward-maximal default action, and nearest-neighbor consensus, the majority action among the state's 25 nearest historical neighbors; thresholds are selected on validation data only.

As autonomous coverage decreases, retained-set classification error decreases monotonically and closed-loop intervention rate decreases substantially for both CQL and AWAC and both fallback designs. This pattern holds across all severity quintiles, including Q5, as shown in Figure 3. Selective deferral therefore changes both retained Decision Agreement and Offline Closed-Loop Behavior as intended.

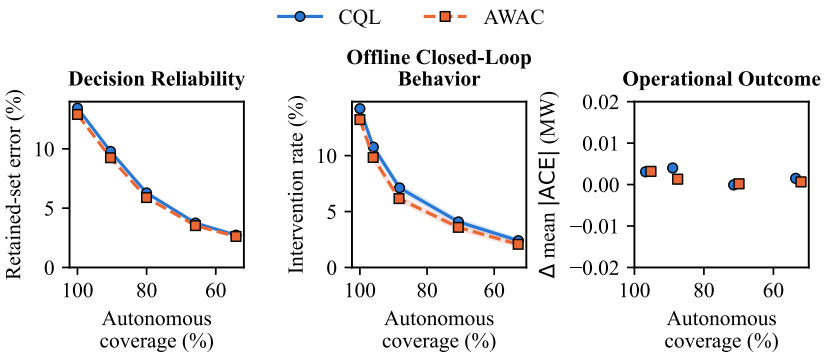

Figure 3: Uncertainty-aware selective decision support across autonomous coverage. CQL and AWAC are shown together to emphasize replication. As autonomous coverage decreases, retained-set classification error falls and closed-loop intervention behavior changes, while the estimated change in mean |ACE| remains near zero. Both HOLD and nearest-neighbor consensus fallbacks were evaluated.

Operational Outcome, however, remains flat across the coverage range for both algorithms and both fallback designs, with no difference between HOLD and consensus detectable at the evaluator's resolution (Table 15, Appendix A.8). Thus, reducing autonomous coverage improves Decision Agreement among retained decisions and substantially changes Offline Closed-Loop Behavior without producing a change in Operational Outcome detectable at the evaluator's resolution.

4 Discussion

Our results show that Decision Agreement, Offline Closed-Loop Behavior, and Operational Outcome capture distinct aspects of a learned decision-support policy. Across the controlled reward intervention, ambiguity-aware supervision, and selective deferral, changes in one dimension do not necessarily translate into corresponding changes in the others. In particular, policies can exhibit substantially different closed-loop behavior despite similar historical agreement, while substantial changes in agreement and behavior need not produce changes in the measured operational outcome detectable at this resolution. Historical agreement is therefore informative as a diagnostic of correspondence with recorded decisions, but does not by itself validate the policy.

Agreement evaluates whether a policy reproduces historical decisions; closed-loop evaluation examines the behavior induced when those decisions are repeatedly executed; and operational evaluation asks whether that behavior produces measurable system-level consequences. Validating a policy therefore requires evaluating each dimension separately rather than inferring one from another. In our experiments, selective deferral provides a direct example: reducing autonomous coverage improves agreement among retained decisions and substantially changes closed-loop intervention behavior, but does not improve the measured |ACE| outcome. Likewise, the controlled reward intervention changes behavior without producing a corresponding change in agreement. The absence of an outcome effect should not be interpreted as evidence that the policies are operationally equivalent in general, but as evidence that the present outcome measure does not distinguish them under the evaluated conditions.

Several limitations define the scope of these findings. The study considers one TSO balancing task and evaluates closed-loop behavior in a learned transition environment rather than on the physical grid. The data also span a mid-study market-mechanism change and a resulting seasonal skew in the validation/test splits, both quantified in Appendix A.9. The operational assessment is also centered on |ACE|, and other system-level metrics or longer-horizon effects may reveal differences not captured here. The results should therefore be interpreted as an empirical demonstration of the validation gap in the studied setting, rather than as a general claim about offline RL. For decision-support applications, the findings nevertheless motivate evaluating historical agreement alongside policy-induced behavior and operational outcomes.

AI USE STATEMENT

The large language models (LLMs) ChatGPT and Claude were used in the preparation of this work for the following purposes: (i) to aid and polish the writing, including language editing; (ii) for retrieval and discovery, such as finding related work; (iii) for research ideation and execution, including programming assistance for the experiments and analysis; and (iv) to draft sections of the paper. All AI-assisted content, code, and analyses were reviewed, adapted, and verified by the authors, who assume full responsibility for the accuracy, integrity, and final content of the manuscript.

REFERENCES

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Elia. Elia transmission belgium gains access to MARI, the European platform for exchanging tertiary balancing energy. https://www.elia.be/en/press/2025/05/20250523_elia-transmission-belgium-gains-access-to-mari, 2025. Accessed: September 2026.

Elia. Elia Open Data Platform, 2026. [Online]. Available: https://www.elia.be/en/grid-data. Accessed: May 26, 2026.

ENTSO-E. ENTSO-E Transparency Platform, 2026. [Online]. Available: https://transparency.entsoe.eu/. Accessed: May 26, 2026.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, et al. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations (ICLR)*, 2021.

Hsiang Hsu, Guihong Li, Shaohan Hu, and Chun-Fu Chen. Dropout-based rashomon set exploration for efficient predictive multiplicity estimation. In *International Conference on Learning Representations*, 2024.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. When should we prefer offline reinforcement learning over behavioral cloning? In *International Conference on Learning Representations (ICLR)*, 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Enrico Marchesini, Benjamin Donnot, Constance Crozier, Ian Dytham, Christian Merz, Lars Schewe, Nico Westerbeck, Cathy Wu, Antoine Marot, and Priya L. Donti. Rl2grid: Benchmarking reinforcement learning in power grid operations. *arXiv preprint arXiv:2503.23101*, 2025.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. AWAC: Accelerating Online Reinforcement Learning with Offline Datasets, June 2020. arXiv:2006.09359 [cs.LG].

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations (ICLR)*, 2022.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

Jie Wang, Rui Gao, and Hongyuan Zha. Reliable off-policy evaluation for reinforcement learning. *Operations Research*, 72(2):699–716, 2024.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pp. 1995–2003. PMLR, 2016.

Ziyu Wang, Alexander Novikov, Konrad Zolna, Jost Tobias Springenberg, Scott Reed, Bobak Shahriari, Noah Siegel, Josh Merel, Caglar Gulcehre, Nicolas Heess, and Nando de Freitas. Critic Regularized Regression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7768–7778. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/588cb956d6bbe67078f29f8de420a13d-Abstract.html.

A Appendix

A.1 Action Threshold Sensitivity

The threshold $\tau = 50$ MW marks a physical separation between automatic and supervisory balancing in the data (Elia, 2026). Below it, net mFRR is effectively absent (mean 2 MW) while automatic aFRR is active almost throughout (mean 60 MW); above it, mean net mFRR rises sharply to 162 MW, isolating deliberate operator activation from automatic regulation. Class balance is stable to this choice: the HOLD fraction varies smoothly from 81.6% at $\tau = 30$ MW to 86.7% at $\tau = 75$ MW.

A.2 Reward Formulation Details

The stress and regime variables (Table 4, Appendix A.3) are constructed from six normalized stress dimensions – deviation, balancing, variability, uncertainty, cross-border, and infrastructure stress – reduced via PCA to three components (76% cumulative explained variance; Figure 4). PC1, dominated by cross-border stress (loading 0.92), is adopted as the Operational Stress Index (OSI); PC2 loads jointly on deviation (0.59), balancing (0.44), and uncertainty (0.43) stress; PC3 is dominated by forecast-uncertainty stress (0.76). Infrastructure stress loads near zero on all three components. A Gaussian Mixture Model with $K = 3$ components, selected via BIC and silhouette score, partitions the latent stress space into a *stable* regime (44.7% of observations), a *balancing-stressed* regime (35.2%), and a *critical-uncertainty* regime (20.1%). The extreme-regime indicator ϕ_{ext} below is 1 for the least-frequent regime (critical-uncertainty) and 0 otherwise.

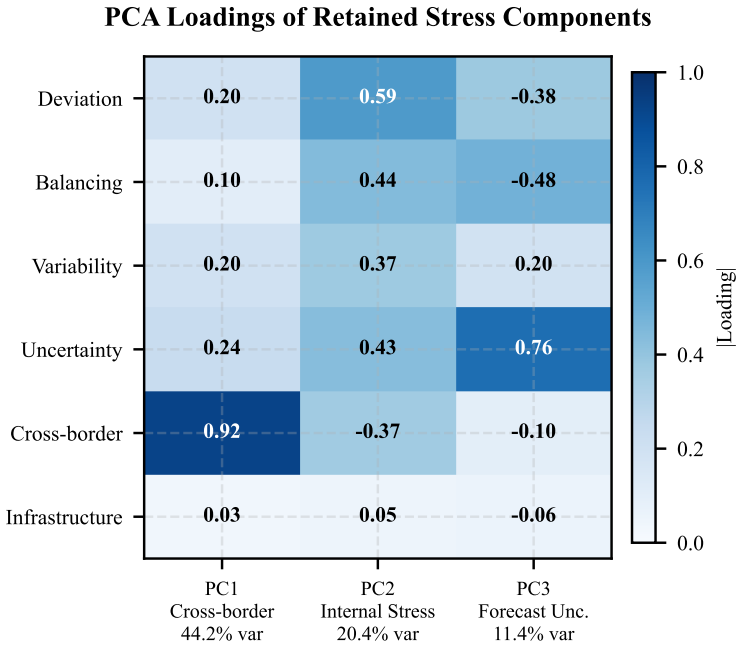

Figure 4: PCA loadings of the three retained stress components. Shading reflects loading magnitude (darker = larger); signed values are printed in each cell. Cross-border stress dominates PC1 (0.92); PC2 loads jointly on deviation, balancing, and uncertainty; PC3 is dominated by uncertainty (0.76). Infrastructure stress loads near zero throughout.

The Bayesian-optimized reward R_1 is defined as a weighted sum of five standardized penalty primitives representing imbalance, activation, stress, extreme-regime conditions, and activation cost:

$$\phi(s,a) = [\phi_{\text{imb}}(s), \phi_{\text{act}}, \phi_{\text{stress}}(s), \phi_{\text{ext}}(s), \phi_{\text{cost}}], \tag{16}$$

Each primitive is standardized using training-set mean and standard deviation,

$$\hat{\phi}_j(s,a) = \frac{\phi_j(s,a) - \mu_j}{\sigma_j}, \tag{17}$$

giving

$$R_1(s,a) = -w^\top \hat{\phi}(s,a), \tag{18}$$

where the weights w are given in Table 3. The weights are obtained by Bayesian optimization to reproduce the historical TSO actions, constrained to the probability simplex ($w_i \geq 0$, $\sum_i w_i = 1$), so Table 3 reports relative importance among the five primitives rather than independently scaled coefficients.

Table 3: Bayesian-optimized reward weights. The five weights are constrained to sum to 1 and multiply standardized penalty primitives; since $R_1 = -w^\top \hat{\phi}(s,a)$, a larger weight means a stronger relative penalty.

Primitive	Weight
Imbalance	0.098
Activation	0.203
Stress	0.113
Extreme	0.324
Cost	0.263

Three primitives depend only on the state,

$$\phi_{\text{imb}}(s) = |\text{ACE}_t|, \qquad \phi_{\text{stress}}(s) = |\text{stress_deviation}_t|, \tag{19}$$

and $\phi_{\text{ext}}(s) \in \{0,1\}$ denotes the extreme-regime indicator. The remaining two primitives are computed from realized continuous activation rather than from the discretized supervisory action:

$$\phi_{\text{act}} = |\text{mFRR}_t^{\text{net}}| + |\text{IGCC}_t^{\text{net}}|, \tag{20}$$

$$\phi_{\text{cost}} = |\text{mFRR}_t^{\text{net}}|\Delta t|\text{marginal price}_t|, \tag{21}$$

with $\Delta t = 0.25$ h. The activation term therefore includes realized IGCC activation, although IGCC is not part of the supervisory policy's action space.

Because all reward weights are non-negative, the activation and cost terms favor lower realized activation. Thus, R_1 has a strong structural preference toward low-activation behavior, but this does not imply that HOLD strictly maximizes the reward for every state. The strength of this bias is also supported by an independent doubly robust policy optimization experiment trained on R_1, which converged to an all-HOLD policy. Its estimated value exceeded the historical operator's mean reward on the test set.

The second reward, R_2, is fixed and state-independent:

$$R_2(a) = \begin{cases} 0.05, & a = \text{HOLD}, \\ 1.0, & a = \text{UP}, \\ 1.2, & a = \text{DOWN}. \end{cases} \tag{22}$$

Unlike R_1, R_2 explicitly favors intervention over inaction and is used as the controlled reward intervention in the main experiment.

A.3 Full State Feature List

Table 4 lists all 77 state variables, grouped by content type as in §2.1.

A.4 Transition Model

Table 5 compares candidate transition-model families on the full feature set. XGBoost achieves the best fit and is used throughout the paper.

Table 5: Transition-model family comparison on the full feature set.

Model	**RMSE (Val)**	R^2 **(Val)**	**RMSE (Test)**	R^2 **(Test)**
Persistence (baseline)	114.27	0.479	–	–
RF	42.95	0.926	40.82	0.924
LGBM	26.35	0.972	25.46	0.970
XGB	**25.65**	**0.974**	**24.55**	**0.973**
ElasticNet	38.03	0.942	50.82	0.882

Table 4: Full state variable list by content-type group.

Group	#	Variables
Deviation	12	ACE, ACE_lag1, ACE_ramp, ACE_ramp_curvature, ACE_rolling_std, ACE_x_aFRR, system_imbalance_lag1, system_imbalance_lag2, imbalance_ramp, imbalance_ramp_curvature, imbalance_rolling_std, imbalance_x_net_position
Balancing (aFRR)	5	aFRR_net, aFRR_net_lag1, aFRR_ramp, aFRR_ramp_curvature, aFRR_rolling_std
Cross-border	12	flow_de/fr/nl/uk_lag1, flow_de/fr/nl/uk_ramp, net_position_lag1, net_position_ramp, implicit_net_position_lag1, implicit_net_position_ramp
Generation	9	gen_gas_lag1, gen_nuclear_lag1, gen_solar_lag1, gen_wind_onshore_lag1, wind_ramp, wind_rolling_std, solar_ramp, solar_rolling_std, renewable_x_load
Infrastructure	7	gen_unavailability_ramp, gen_unavailable_change_lag1, gen_unavailable_ratio_lag1, high_gen_outage_lag1, high_tx_stress_lag1, tx_reduction_ramp, tx_reduction_ratio_lag1
Load (actual)	3	load_actual_lag1, load_ramp, load_rolling_std
Temporal	4	hour_sin, hour_cos, dow_sin, dow_cos
Market: forecasts	9	solar_forecast_lag1, solar_forecast_error, solar_forecast_error_ramp, wind_onshore_forecast_lag1, wind_forecast_error, wind_forecast_error_ramp, load_forecast_lag1, load_forecast_error, load_forecast_error_ramp
Market: lagged prices	3	alpha_lag1, imbalance_price_lag1, marginal_incremental_price_lag1
Imbalance nowcast	1	imbalance_nowcast_causal
Stress & regime	12	OSI, regime_id, regime_probability, stress_pc1/2/3, stress_deviation, stress_balancing, stress_variability, stress_uncertainty, stress_crossborder, stress_infrastructure
Total	**77**	

A.5 HYPERPARAMETERS

Table 6 lists the Stage-1 hyperparameters shared by every method. Table 7 lists the Stage-2 hyperparameters specific to CQL and AWAC, the two algorithms used throughout §3.2 and §3.3.

Table 6: Shared Stage 1 hyperparameters.

Parameter	Value
Pretraining epochs	20
Optimizer	AdamW
LR schedule	Cosine annealing

Table 7: Stage 2 hyperparameters, CQL and AWAC.

CQL		AWAC	
Parameter	Value	Parameter	Value
BC regularizer weight β	1.0	Temperature λ	1.0
CQL weight α	10^{-4}	Weight clip	20.0

A.6 Decision Agreement: Selective Prediction

Table 8 is the full per-seed record behind §3.2's Decision Agreement summary. Unlike Offline Closed-Loop Behavior and Operational Outcome (Appendix A.7), no significance test is reported here: each cell is a single training run's held-out metric rather than a distribution over paired rollouts.

Table 8: Reward-ablation Decision Agreement (held-out test macro F1), all 5 seeds, both algorithms.

Method, seed	**Macro F1** (R_1)	**Macro F1** (R_2)	**Diff** ($R_1 - R_2$)**, pp**
CQL, 42	0.680	0.684	-0.42
CQL, 43	0.675	0.657	$+1.83$
CQL, 44	0.682	0.661	$+2.13$
CQL, 45	0.678	0.657	$+2.10$
CQL, 46	0.688	0.665	$+2.25$
AWAC, 42	0.667	0.687	-1.97
AWAC, 43	0.674	0.664	$+1.00$
AWAC, 44	0.680	0.667	$+1.34$
AWAC, 45	0.673	0.660	$+1.26$
AWAC, 46	0.696	0.670	$+2.67$

Table 9 reports selective-prediction accuracy at 100/90/80/70% coverage for the reward-ablation pair (seed 42), using a softmax-over-Q pseudo-probability as the confidence score for both R_1 and R_2.

Table 9: Selective prediction, reward-ablation pair (seed 42): accuracy retained at each coverage level.

Policy	**100%**	**90%**	**80%**	**70%**
Reward R_2	0.876	0.915	0.953	0.980
Reward R_1	0.913	0.951	0.971	0.983

The accuracy gap between the two rewards narrows as coverage decreases – from 3.7pp at 100% coverage to 0.3pp at 70% – consistent with R_1's higher overall Decision Agreement being concentrated in the easier, high-confidence majority of cases rather than the hardest, most ambiguous ones.

A.7 Reward-Ablation Outcome Detail

Table 10 breaks §3.2's Offline Closed-Loop Behavior summary down to the full per-cell record. Every cell shows the same pattern: intervention rate is higher under R_2 than under R_1, with large effect sizes and 95% CIs that exclude zero.

Table 10: Reward-ablation Offline Closed-Loop Behavior (closed-loop intervention rate), all 5 seeds, both algorithms. Every cell shows a large, consistent divergence ($R_2 > R_1$); all effect sizes are large in magnitude and every CI excludes zero.

Method, seed	**Diff ($R_1 - R_2$), pp**	**p (Wilcoxon)**	**rank-biserial**	**95% CI**
CQL, 42	-12.44	7.5×10^{-10}	-1.00	$[-14.77, -10.27]$
CQL, 43	-9.42	2.4×10^{-9}	-0.98	$[-11.31, -7.62]$
CQL, 44	-9.81	2.3×10^{-8}	-0.92	$[-12.02, -7.58]$
CQL, 45	-10.29	6.8×10^{-9}	-0.95	$[-12.75, -8.02]$
CQL, 46	-10.42	2.5×10^{-9}	-1.00	$[-12.58, -8.38]$
AWAC, 42	-11.38	1.8×10^{-9}	-0.99	$[-13.65, -9.19]$
AWAC, 43	-7.10	3.0×10^{-9}	-0.98	$[-8.87, -5.54]$
AWAC, 44	-10.48	1.8×10^{-9}	-1.00	$[-12.83, -8.23]$
AWAC, 45	-9.96	2.0×10^{-9}	-0.97	$[-12.10, -7.94]$
AWAC, 46	-4.48	1.1×10^{-3}	-0.55	$[-6.81, -2.15]$

Table 11 similarly breaks down §3.2's Operational Outcome summary to the full per-cell record.

Table 11: Reward-ablation Operational Outcome (simulated mean $|\mathrm{ACE}|$), all 5 seeds, both algorithms. No cell is significant; every CI includes zero.

Method, seed	**Diff ($R_1 - R_2$), MW**	**p (Wilcoxon)**	**rank-biserial**	**95% CI**
CQL, 42	$+0.21$	0.521	$+0.11$	$[-2.46, 2.82]$
CQL, 43	-1.77	0.521	-0.11	$[-4.74, 0.81]$
CQL, 44	-1.18	0.901	-0.02	$[-4.22, 1.78]$
CQL, 45	$+0.70$	0.625	-0.08	$[-1.48, 3.09]$
CQL, 46	$+0.83$	0.143	$+0.24$	$[-1.81, 3.49]$
AWAC, 42	$+0.94$	0.118	$+0.26$	$[-2.32, 3.62]$
AWAC, 43	$+0.58$	0.449	$+0.13$	$[-1.76, 2.80]$
AWAC, 44	-0.67	0.893	$+0.02$	$[-3.69, 2.35]$
AWAC, 45	-1.60	0.378	-0.15	$[-4.55, 1.12]$
AWAC, 46	-0.13	0.546	$+0.10$	$[-2.14, 1.66]$

Figure 5 shows the full quintile-by-quintile test underlying §3.3's finding for the most extreme-imbalance cases: historical intervention is compared against the counterfactual of always choosing HOLD, separately at each pre-decision imbalance-severity quintile. Only Q5, the quintile with the most extreme imbalance, crosses the significance threshold.

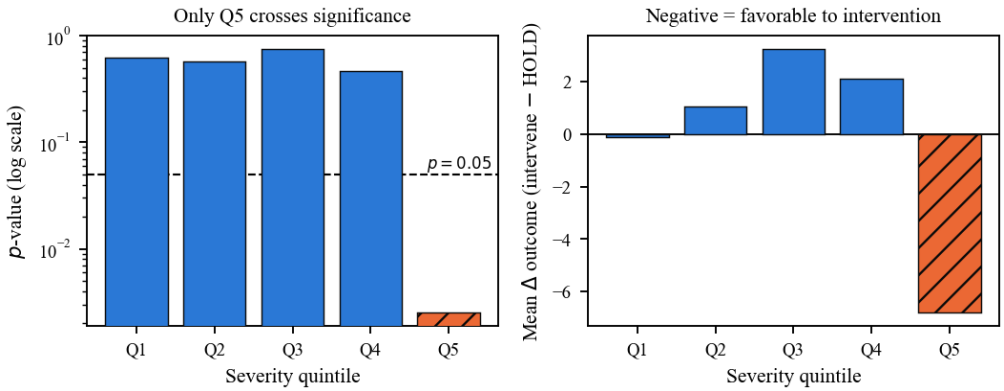

Figure 5: Historical intervention vs. HOLD across pre-decision imbalance-severity quintiles, from calmest (Q1) to most extreme imbalance (Q5). Left: one-sided Mann-Whitney p (log scale) against the $p = 0.05$ reference; only the most extreme-imbalance quintile (Q5) crosses it. Right: the same quintiles' mean outcome delta, negative meaning intervention outperformed HOLD. Q5 is the only quintile with a significant historical association favoring intervention – the quintile where a real link between historical agreement and operational outcome should be easiest to detect.

Closed-loop outcome distribution (Table 2). Table 2 reports mean $|\mathrm{ACE}|$ per policy; Table 12 gives the full across-trajectory distribution. Mean $|\mathrm{ACE}|$ is the mean of $|\mathrm{ACE}_t|$ over all 96 steps of each trajectory, not restricted to intervention steps; each step's state, and its ACE, results from that policy's own preceding action applied through the shared transition model. While the means are close, the distributions are not: CRR, the most interventionist policy, also has the largest standard deviation and the largest maximum $|\mathrm{ACE}|$ across trajectories. A breakdown by intervention direction (UP vs. DOWN) was not retained from the original rollouts and is not reported here.

Table 12: Distribution of per-trajectory mean $|\mathrm{ACE}|$ (MW) across the 150 rollouts behind Table 2.

Policy	Mean	Std	Min	Max	Q95
Operator	21.20	6.62	11.85	52.48	39.50
BC	21.04	6.37	11.23	49.98	29.27
CQL	21.11	7.20	12.59	67.16	30.59
AWAC	21.41	7.75	12.36	53.39	44.10
CRR	21.82	8.99	11.96	80.04	42.20
Confidence Hybrid	22.07	7.88	11.48	64.06	42.22

A.8 Robustness Interventions: Per-Seed and Per-Class Detail

This subsection gives the full statistical detail behind two claims from the main text: §3.3's finding that five generic training interventions do not move Decision Agreement, and §3.4/§3.5's two further interventions – one that changes Decision Agreement, one that changes Offline Closed-Loop Behavior – neither of which produces a change in Operational Outcome detectable at the evaluator's resolution.

Test 1: moving the model's decisions. Main text §3.3 reports that Decision Agreement resists five generic training interventions: a larger optimization budget, reward re-weighting, a reserve-margin state feature, logit adjustment (Menon et al., 2021), and LDAM-DRW class re-weighting (Cao et al., 2019). This rules out that the gap is simply immovable. Here is the full detail: none produces a meaningful change for either algorithm – the first four leave macro F1 flat or worse, and LDAM-DRW raises balanced accuracy only by lowering macro F1.

§3.4 reports that a targeted intervention succeeds where generic pressure does not: ambiguity-aware supervision blends each hard label with the empirical label distribution among its $k = 25$ nearest neighbors, in proportion to local ambiguity, directly targeting the mechanism rather than applying generic pressure. Macro F1 rises by $+2.1$pp for CQL and $+1.7$pp for AWAC. Table 13 gives the full detail, including a smoothing control isolating how much of this gain is ambiguity-specific rather than generic.

Table 13: Ambiguity-aware supervision: per-class classification effects and the contribution of a generic label-smoothing control.

Measure	**CQL**	**AWAC**
Per-class ΔF1: HOLD / UP / DOWN	$+2.1$ / $+3.5$ / $+0.8$pp	$+1.9$ / $+2.6$ / $+0.6$pp
Total macro-F1 gain	$+2.11$pp	$+1.68$pp
Gain reproduced by generic label smoothing	$+0.64$pp ($\sim 30\%$)	$+0.63$pp ($\sim 37\%$)
DOWN F1: ambiguity-aware vs. smoothing-only	0.491 vs. 0.488 (indistinguishable)	0.494 vs. 0.492 (indistinguishable)

§3.4 also reports that this change in decisions does not reach Operational Outcome. Table 14 gives the full detail: the deployment shift and outcome significance by quintile, where rank-biserial effect sizes stay below 0.04 for both algorithms, and only Q3 and Q5 reach nominal significance; these are uncorrected across the five quintiles, and a Holm correction leaves only Q5 significant for both algorithms (Table 14).

Decision Agreement changes; Operational Outcome does not detectably. Because the decisions are independently verified to have moved, the missing outcome effect cannot be attributed to the policies failing to change; it is bounded by the evaluator's resolution.

Table 14: Ambiguity-aware supervision: changes in model decisions, deployed behavior, and operational outcome. Values are reported for CQL and AWAC.

Measure	CQL	AWAC
Change in model decisions		
HOLD: low → high ambiguity	$0.27\% \rightarrow 14.47\%$	$0.26\% \rightarrow 13.63\%$
UP: low → high ambiguity	$0\% \rightarrow 21.27\%$	$0\% \rightarrow 22.15\%$
DOWN: low → intermediate → high	$0\% \rightarrow 23.0\% \rightarrow 14.5\%$	$5.7\% \rightarrow 21.8\% \rightarrow 13.7\%$
Change in deployed behavior		
Intervention reduction across severity quintiles	55.6%–72.8%	55.5%–71.7%
Intervention reduction in Q5	10.47 pp	9.33 pp
Operational outcome		
Outcome significance, Q3 / Q5	$p = 0.023$ / 0.004	$p = 0.0104$ / 0.0044
Holm-corrected p, Q3 / Q5 (family of 5 quintiles)	$p = 0.090$ / 0.021	$p = 0.042$ / 0.022
Maximum rank-biserial effect size $\lvert r \rvert$	0.038	0.037

Test 2: moving deployed behavior directly. §3.5 reports a second test of the same diagnosis: rather than retraining on ambiguous states, uncertainty-aware selective deferral avoids committing to a decision on them. The classifier itself is not retrained; only the decision-time routing changes, as described in §3.5. Both algorithms use the same rollout protocol as §3.3 (3 seeds each) across five swept deferral rates; AWAC reuses CQL's thresholds, since the ambiguity score depends only on the state, not the policy.

As coverage decreases, two things happen consistently: retained-set classification error falls monotonically, and Offline Closed-Loop Behavior's intervention rate drops sharply. This holds across every severity quintile, including Q5 – the most extreme-imbalance stratum, where §3.3's historical-data check found the strongest evidence favoring intervention – and across both algorithms and both fallback designs (Figure 3). Operational Outcome, by contrast, is flat across the entire swept range, and HOLD and consensus do not differ at any threshold (Table 15).

Table 15: Selective-deferral coverage sweep: Delta|ACE| outcome test (Mann-Whitney, pooled 3 seeds) vs. each algorithm's standard policy, at each swept threshold. Coverage and intervention rate are rollout-time actuals, not validation targets; HOLD vs. Consensus (last column) compares the two fallbacks directly. See Figure 3 in the main text.

Alg.	Target defer. / Design	Coverage	Interv. rate	Δ mean |ACE|	p vs. Standard	p, HOLD vs. Cons.
CQL	Standard (0%)	100%	14.2%	–	–	–
CQL	5% / HOLD	96.0%	10.3%	+0.008	0.879	0.922
CQL	5% / Consensus	96.0%	11.2%	−0.002	0.959	
CQL	10% / HOLD	88.3%	6.7%	+0.004	0.999	0.982
CQL	10% / Consensus	88.1%	7.5%	+0.004	0.982	
CQL	20% / HOLD	70.7%	3.8%	+0.002	0.912	0.921
CQL	20% / Consensus	70.6%	4.4%	−0.002	0.837	
CQL	30% / HOLD	52.9%	2.1%	+0.003	0.988	0.885
CQL	30% / Consensus	52.7%	2.7%	+0.000	0.873	
AWAC	Standard (0%)	100%	13.2%	–	–	–
AWAC	5% / HOLD	96.0%	9.4%	+0.008	0.847	0.889
AWAC	5% / Consensus	96.1%	10.3%	−0.001	0.958	
AWAC	10% / HOLD	88.5%	5.8%	+0.002	0.989	0.990
AWAC	10% / Consensus	88.2%	6.5%	+0.000	0.978	
AWAC	20% / HOLD	70.7%	3.2%	+0.001	0.928	0.929
AWAC	20% / Consensus	70.7%	3.9%	−0.001	0.860	
AWAC	30% / HOLD	52.9%	1.8%	+0.002	0.969	0.908
AWAC	30% / Consensus	52.8%	2.4%	−0.001	0.876	

A.9 Data Regime: Market Mechanism and Seasonal Coverage

Market mechanism change (MARI). Roughly midway through the study window, on 21 May 2025, Elia connected to MARI, the pan-European platform for exchanging mFRR balancing energy (Elia, 2025). Before this date, mFRR bids cleared against Belgian demand in isolation; after it, bids clear against a shared cross-European merit-order list, with cross-border netting and a shared cross-border marginal price. All results reported in the main text pool data from both regimes; Table 16 quantifies the difference.

Table 16: Descriptive comparison of the data before and after Elia's 21 May 2025 connection to the MARI platform.

Measure	Before MARI	After MARI
HOLD	81.8%	89.7%
UP	13.8%	5.4%
DOWN	4.4%	4.8%
Mean activation when active (MW)	168.5	145.2
Mean marginal price	159.94	109.36
Mean $\lvert\mathrm{ACE}\rvert$	19.75	14.47

Pooling across both regimes, rather than restricting to one, is deliberate. First, every comparison in this paper is a paired, within-experiment comparison – reward ablation, method comparison, ambiguity-aware supervision, and selective deferral all evaluate every compared policy on the same pooled data, the same transition model, and the same rollout starting states. A regime characteristic shared identically across every compared condition does not bias the differences between them, which is what each experiment in this paper actually reports. Second, restricting to a single regime would substantially shrink an already modest single-TSO dataset – particularly for the rarer DOWN class – weakening the statistical power of the significance tests reported throughout §3. Third, this mechanism change is a property of the real operational setting rather than an artifact of data handling: any deployment window of comparable length would be expected to span some evolution in market design, and disclosing this transparently is preferable to selecting an artificially "clean" sub-period that would not be representative of any future regime either.

Seasonal composition of the splits. The chronological split also concentrates validation and test in different parts of the year – validation mostly June–October 2025, test mostly November–December 2025 – rather than each spanning a full annual cycle as training does. The state includes generation and load variables (Appendix A.3) that correlate with seasonal conditions, but no explicit month-of-year feature. As with the market-mechanism change above, this composition is shared identically across every compared method, so it does not differentially bias the paper's comparative claims.