Enforcement Triage as Constrained MPC: From FIFO to Foresight

Chinmay Koimuttum and Rohan C. Shekhar

***Abstract*—In a trust-and-safety review backlog, harm accrues while genuine cases wait to be actioned, cases carry consumer-protection deadlines, review capacity is scarce and below the mean arrival rate, and the magnitude of harm caused by a case is observable only through a noisy classifier score until it is eventually reviewed by a human. Production systems prioritize this backlog in a myopic fashion, using a first-in-first-out approach or by choosing those cases with the highest estimated harm based on a machine-learning classifier. These approaches ignore the forecastable quantities of future load and the rate at which pending harm accrues. The problem can be posed as a constrained model predictive control (MPC) over a marked-arrival backlog. A seeded, config-driven simulator anchored to public data provides the simulated environment to test MPC's efficacy.**

Evaluating against a series of baseline approaches (the $c\mu$ rule, a Whittle-inspired index, a full-information greedy benchmark, and a PPO reinforcement-learning agent), a fluid linear-program MPC controller is designed and tested on paired seeds. Three findings emerge and form the primary contribution of this work: Firstly, the cost of partial observability is large and dominant under a realistic signal. A factorial attribution assigns roughly 20% of it to unobserved harm *magnitude*, which improving a label-only classifier cannot address under our score-magnitude independence assumption but which can be learned online, and 80% to label uncertainty. Secondly, the forecast-aware controller beats the strong myopic rules only *conditionally*, namely when the deadline constraint carries weight and load is near-critical, and can be significantly *worse* than the $c\mu$ rule when its internal model is misleading. Finally, both failure modes share one cause and can be addressed by applying a dual-mode controller which executes the myopic rule when the plan's anticipatory value is found to be negligible. It reveals no statistically significant disadvantage against the better of the pure $c\mu$ and MPC policies in any of the tested regimes. In fact, it is significantly better than both pure rules in several regimes, and it retains MPC's advantages where foresight is particularly valuable.

***Index Terms*—Model predictive control, queuing control, restless bandits, content moderation, partial observability, false-positive-rate constraints.**

I. Introduction

An online platform's trust-and-safety operation can be viewed as a scheduling problem, where cases arrive continuously and are placed in a backlog for human review. Real-world examples include flagged sellers on a marketplace and flagged content on social media platforms. A case is deemed resolved when it completes human review. Until then however, if the case is genuinely harmful, buyers or users continue to be affected. The rate of arrivals routinely exceeds the rate at which a finite team of investigators can act, so a backlog forms. This raises the question of which of the many pending cases should the limited human reviewers examine next.

The complexity of addressing this issue manifests through four compounding reasons. 1. The decisive attributes are *hidden*. Before a case is investigated, it is unknown whether it is truly in breach or simply a noisy risk score, or *how much* harm the case causes. 2. Harm *accrues* while a case waits to be reviewed, therefore, prolonged deferral is increasingly costly. Moreover, cases carry *deadlines* (consumer-protection or policy service-level agreements, SLAs) whose breach possesses a separate penalty. Finally, cases arrive in *waves*, so competent present decisions must account for expected future load.

Production systems often prioritize myopically. First-in-first-out may seem to be a fair strategy but it is harm-blind, and ranking by classifier score only determines ordering and does not exploit the score magnitude. That is, it only predicts *whether* a case is in breach, and (for a label-only classifier) not the amount of harm it actually causes, nor whether a deadline risks being missed. Two forecastable quantities, namely the future load and the accrual of pending harm, remain unexploited in these approaches. This paper explores whether applying *model predictive control* (MPC) to select the order for actioning cases can improve upon the state-of-the-art. MPC forecasts future harm, finds the optimal review ordering over a finite horizon that minimizes predicted cost, and executes only the first action. Replanning occurs over the finite horizon shifted one time step into the future and the process is repeated.

In order to test the efficacy of MPC, this paper models the backlog of cases as a marked-point process with hidden per-case harm. It assumes a scarce review capacity, a deadline to avoid violating SLA, and an optional false positive rate-constrained auto-action mode. This approach forms the basis of modeling any scarce human servers attending to a backlog of jobs whose true severity is hidden until they are serviced and whose cost accrues with delay provides a suitable setting. Examples include a content-moderation review (flagged content awaiting a human reviewer), emergency-department triage (a patient's condition deteriorates while waiting, and true acuity is uncertain until proper examination), security-alert investigation, and preventive-maintenance backlogs, among others. These applications are structurally related, but it must be noted that several involve complex dynamics absent from the model presented in this work. For example, deteriorating states in emergency triage, heterogeneous or non-preemptive servers

C. Koimuttum is an independent researcher (e-mail: ckoimuttum@gmail.com).

R. C. Shekhar is with the College of Computational Sciences, Minerva University, San Francisco, CA, USA (e-mail: rshekhar@minerva.edu).

Corresponding author: R. C. Shekhar.

or domain-specific ethical constraints are not considered. The commonality is the basic structure of a marked point process with hidden per-case cost, scarce capacity, and deadlines. In this paper, marketplace enforcement is chosen as the candidate application since it is adequately supported by available public data (Sec. V).

The primary contributions of this work are as follows:

1) a seeded, config-driven simulator of harm-accruing, deadline-constrained, partially-observed enforcement triage with a standard reinforcement-learning interface and marked arrivals spanning seasonal, Markov-modulated, and self-exciting processes (Sec. III).
2) a comparable policy suite (myopic heuristics, two queuing-theoretic index rules, a full-information greedy benchmark, a PPO agent, and a fluid-LP MPC) with a methodology for calibrating the environment's observable layer to public data (Secs. IV and V).
3) a characterization of *when* MPC is truly helpful, and a decomposition showing that a substantial, classifier-irreducible part of the cost is unobserved harm *magnitude*, which is instead partly reclaimable by learning it online from the operation's own review log (Secs. VI–VII).
4) a dual-mode controller that removes the two conditions under which naive MPC is significantly worse than a myopic rule. On held-out seeds, it reveals no statistically significant disadvantage against the better of the pure $c\mu$ and MPC policies in any regime tested, instead showing a significant improvement in several regimes (Sec. VIII).

II. Related Work

It is not without precedent to cast a review pipeline as a sequential control problem over a harm-accruing, capacity-constrained queue. Prior art formalizes this problem and proves that classical myopic selection rules are suboptimal under evolving costs: Gocmen *et al.* [1] study queuing with uncertain, Markov-evolving holding costs and propose an opportunity-adjusted index rule; Lykouris and Weng [2] model an AI–human moderation pipeline in which true cost is revealed only on review, controlled by a drift-plus-penalty policy; and Lee *et al.* [3] design a near-optimal index policy for a queue whose job classes are predicted by a classifier. Each of these, however, solves the problem with a *myopic index or Lyapunov policy*. None utilize MPC, enforce explicit deadlines, forecast arrivals, or impose a false-positive-rate constraint.

The various components of the approach applied here are individually mature. The generalized $c\mu$ rule [4] and the $c\mu/\theta$ rule under abandonment [5] provide the reference myopic policies for holding-cost scheduling. The Whittle index for restless bandits [6], and its specialization to deadline scheduling [7], provide an anticipatory index baseline. Model predictive control of queues with arrival forecasting has also been applied to examples outside moderation, such as serverless scheduling [8]. Model predictive control itself is a mature technology that is widely used, with its theoretical guarantees established [9], [10], as is model-free reinforcement learning for scheduling [11], [12]. Finally, the EU DSA Transparency Database has been utilized in a descriptive and audit-oriented fashion [13], but its use as part of a control problem is unique, as far as is known to the authors of this paper.

This paper thoroughly investigates the application of MPC with arrival forecasting, per-category deadlines, partial observability, and a false-positive rate-constrained auto-action channel, anchored to public data. It also reveals *when* the MPC formulation adequately justifies its additional complexity, which arises from the need to solve an optimization problem at each time step.

III. Problem Formulation

A. Nomenclature

Time is discrete with step $\Delta = 1\,\mathrm{h}$ over a horizon of $T = 336$ steps (two weeks). $\mathbb{R}$ and $\mathbb{Z}$ denote the reals and integers. $\mathbf{1}\{\cdot\}$ denotes the indicator function, which is equal to 1 when its argument holds and 0 otherwise. The state is the current *backlog*, which is a set of pending cases of variable cardinality (formally a random finite-set Markov process) with a clock mapping the step index to hour-of-day and day-of-week. A case i carries marks, which are as follows: a category c_i over a fixed taxonomy, a hidden Boolean label $y_i \in \{0, 1\}$ (in breach or not), an observable risk score $s_i \in [0, 1]$, a hidden harm rate $h_i \geq 0$, a service time τ_i (one step in all reported experiments), and a deadline d_i. Only c_i, s_i, and elapsed wait are observed as they occur, whereas y_i and h_i are hidden until review. The deadline d_i and service time τ_i are per-category constants, so they are known from c_i. The scarce resource is review capacity C. With unit service time, the per-step cap on the number of cases reviewed coincides with a workload budget, and the $\tau_i > 1$ case (a workload constraint $\sum_i \tau_i x_i \leq C\Delta$) is left to future work. Capacity is deliberately set below the mean arrival rate to allow a backlog to form.

B. Marked arrivals

At each step, the number of arrivals is drawn, with marks assigned to each arrival. The baseline arrival process is a non-homogeneous Poisson process (NHPP) with intensity $\lambda(t) = \lambda_0\,\phi(t)$, where the seasonal multiplier

$$\phi(t) = 1 + A_d \sin\!\left(\tfrac{2\pi\,\mathrm{hod}}{24} + \varphi_d\right) + A_w \sin\!\left(\tfrac{2\pi\,\mathrm{dow}}{7} + \varphi_w\right) \quad (1)$$

carries daily and weekly cycles, and the per-step count is $\mathrm{Poisson}(\lambda(t)\Delta)$. Two bursty processes, a Markov-modulated Poisson process (MMPP) with a hidden normal/surge state and a self-exciting Hawkes process, are normalized to the same mean load. This ensures that a bursty scenario, where a large burst of cases could arrive at a given time instant, differs from the NHPP in higher-order temporal structure (over-dispersion for MMPP, and self-excited clustering for the Hawkes) but not in mean load. The MMPP surge state and the Hawkes intensity are part of the environment's Markov state, though the controller never observes them.

C. Case marks and partial observability

The marks associated with each case are drawn as

$$c_i \sim \text{Multinomial}(p), \qquad y_i \sim \text{Bernoulli}(\pi_{c_i}),$$
$$s_i \mid y_i \sim \begin{cases} \text{Beta}(a_1, b_1), & y_i = 1, \\ \text{Beta}(a_0, b_0), & y_i = 0, \end{cases} \tag{2}$$

where p is the category mix. The overlap of the two class-conditional score distributions sets the signal quality. Let f_y denote the Beta score density given label y and F_y denote its cumulative distribution function,

$$f_y(u) = \frac{u^{a_y-1}(1-u)^{b_y-1}}{B(a_y, b_y)},$$
$$F_y(x) = \int_0^x f_y(u)\,\mathrm{d}u = I_x(a_y, b_y), \tag{3}$$

with (a_y, b_y) representing the parameters of (2), B denoting the Beta function, and I_x denoting the regularized incomplete Beta function. The harm rate is

$$h_i = \sigma_{c_i} \cdot E_i, \qquad E_i \sim \text{Lognormal}(\mu_e, \sigma_e), \tag{4}$$

the product of a per-category severity σ_{c_i} and a heavy-tailed *exposure* E_i. It is natural that two cases may possess identical scores, which presents an equal probability of being genuine violations. However, it is possible that each case generates vastly different amounts of harm. Assuming that the classifier has no reliable way to a prior estimate the exposure, it is modeled as statistically independent of the harm score.

D. Dynamics and cost

At each decision epoch, the environment advances the simulation clock, samples new arrivals, and appends them to the queue. Then, the selected policy reviews the selected cases and resolves them immediately. Reviewed cases that are genuine violations cease to accrue further harm, whereas reviewed legitimate cases are cleared without penalty. Harm accumulates at rate $h_i\Delta$ for every unresolved case that is in breach, and an SLA violation is recorded whenever a case's waiting time exceeds its deadline d_i. The per-step cost is

$$\ell_t = w_{\text{harm}}\,(\text{harm accrued}) + w_{\text{sla}}\,(\text{new breaches}) + w_{\text{fp}}\,(\text{wrongful actions}), \tag{5}$$

where the last term is only active in the auto-action mode discussed below. The objective is to minimize $\sum_t \ell_t$ over the horizon. The environment implements the standard `reset`/`step` reinforcement-learning interface of OpenAI Gym [14], and uses its maintained successor Gymnasium [15] (each `step` returns the next observation and reward $-\ell_t$) so that off-the-shelf RL algorithms can be applied without modification. All randomness is governed by an injected generator, ensuring a seed will reproduce a given run exactly.

E. Two action modes

In *Mode A* (prioritization, the primary setting) the policy selects at most C pending cases to review each step. In *Mode B*, the policy may additionally *auto-action* a case directly from its score without consuming review capacity. However, auto-actioning a legitimate case incurs the false-positive penalty w_{fp} and counts detrimentally against a cumulative *false-positive rate constraint* which bounds the wrongful-action rate on legitimate cases by ε.

IV. Policies

All policies act on the observable marks. From the class-conditional score model of (2), with Beta densities f_1, f_0 of (3), and the category prior, Bayes' rule gives the posterior

$$\Pr(y_i{=}1 \mid s_i, c_i) = \frac{\pi_{c_i} f_1(s_i)}{\pi_{c_i} f_1(s_i) + (1-\pi_{c_i}) f_0(s_i)}, \tag{6}$$

from which the *expected harm rate* is formed

$$\hat{h}(s_i, c_i) = \Pr(y_i{=}1 \mid s_i, c_i) \cdot \sigma_{c_i} \cdot \bar{E}, \tag{7}$$

where $\bar{E} = \exp(\mu_e + \frac{1}{2}\sigma_e^2)$ denotes the mean exposure. When the true exposure is hidden, this is the best estimate available.

A. Baselines, index rules, and the full-information reference

FIFO (oldest first) and *highest-score* (raw s_i) are harm-blind. *EDF* (earliest deadline first) is deadline-aware but harm-blind. The *$c\mu$ rule* ranks by $\hat{h}(s_i, c_i)/\tau_i$, which is the cost-aware myopic policy. The *Whittle-inspired* index adds a load-scaled anticipatory deadline term to the $c\mu$ index, which amortizes the one-off breach penalty over an estimated residence time $L \approx |\text{backlog}|/C$. This refers to the time necessary for clearing the current backlog at capacity (Little's law at balanced load [16]). Under heavy saturation, it reduces to $c\mu$. It is a heuristic index in the spirit of [7], but note that it is not a derived Whittle index established by a proven indexability result. The *full-information greedy* benchmark (FI-greedy) reads the true y_i and h_i and greedily ranks by true marginal cost. Since it is clairvoyant, it uses information no realizable policy would be able to access. Furthermore, as it is *greedy*, it is not necessarily a finite-horizon optimum, but provides a benchmark that bounds the true finite-horizon optimum from above. The suite of baselines concludes with the inclusion of a PPO agent [11], [17]. Its observation is a fixed-length backlog summary (per-category counts, a score histogram, oldest wait, near-deadline count, the current λ estimate, and hour/day sinusoids). Its action is a weight vector mapped to a per-case linear priority, of which the top C are reviewed. It uses the Stable-Baselines3 defaults (a $[64, 64]$ MLP, γ=0.99, GAE λ=0.95, 2048-step rollouts, learning rate 3×10^{-4}, entropy coefficient 0.005), trained for 2×10^5 steps on the base environment from a single seed.

B. The MPC controller

At each step, the MPC solves a finite-horizon optimization problem over horizon length H executing its first action. A seasonal forecaster supplies the expected future load used for predicting the impact of actioning cases in the future. The internal model used by MPC is deliberately approximate. It reasons about expected harm and mean arrivals while the true environment is stochastic and, under the bursty conditions, structurally different. The re-planning loop is thus tested under genuine model mismatch, characteristic of real-world MPC applications.

1) Compression: The backlog is compressed into G cohorts, indexed $g \in \{1,\dots,G\}$, by (category $\times$ score-bin $\times$ slack-bin). Cohort g possesses its mass m_g, its mean expected harm rate $\hat{h}_g$ (the average of (7) over its members), and a breach step σ_g, which is the earliest step at which a member would breach its deadline ($\sigma_g > H$ if no SLA breaches occur within the horizon; $\sigma_g \le 0$ if the cohort has already breached the deadline). Since each cohort utilizes a single σ_g, the SLA term becomes an approximation, since it credits every reviewed unit of the cohort with averting the earliest member's breach. Slack-binning ensures the deadlines of members are kept close, and finer slack bins naturally shrink the approximation. The cohorts are the *current* backlog, which is known when planning the cases to action. Anticipated future arrivals are similarly compressed into J classes. A class $j \equiv (c,\varsigma)$ pairs a category c with a score-bin ς and possesses a representative harm rate $\hat{h}_j = \hat{h}(\bar{s}_\varsigma, c)$ (equation (7) at the bin center $\bar{s}_\varsigma$) with a cumulative forecast mass, which arrives by horizon step k,

$$A_{j,k} = p_c\, q_{c,\varsigma} \sum_{\kappa=1}^{k} \hat{\lambda}_{t+\kappa}, \qquad (8)$$
$$q_{c,\varsigma} = \pi_c\, Q_\varsigma^1 + (1-\pi_c)\, Q_\varsigma^0,$$

where the category c and score-bin ς are the two components of class j. p_c is the category mix, $\hat{\lambda}_{t+\kappa}$ is the forecaster's expected arrival count κ steps ahead (the seasonal mean $\lambda(t+\kappa)\Delta$ of Sec. III-B), and $Q_\varsigma^y = F_y(e_\varsigma) - F_y(e_{\varsigma-1})$ is the probability that a class-y arrival's score falls in bin ς (edges $[e_{\varsigma-1}, e_\varsigma]$), with F_y being the Beta score CDF of (3). Thus, $q_{c,\varsigma}$ is the probability that a category-c arrival's score lands in bin ς, marginalizing the two class-conditional score distributions over the breach prior (harmful with probability π_c, benign otherwise). $p_c\, q_{c,\varsigma}$ is the probability an arrival belongs to class $j = (c,\varsigma)$, so $A_{j,k}$ is the expected number of class-j arrivals by horizon step k. Both $A_{j,k}$ and $\hat{h}_j$ are not decision variables, but fixed forecast *inputs*, and the sum starts at $\kappa{=}1$ given that current-step arrivals are already in the backlog and can be considered as the initial state for the finite-horizon optimal control problem. The future predicted arrivals forecast in class j may only be reviewed upon arrival so are considered separately from the current backlog. This distinction of current backlog and anticipated future arrivals allows the plan to reserve future review slots for load that is expected rather than commit them to the present backlog, which is the predictive element that distinguishes the MPC from the myopic baselines. As detailed below, only the first-step of the review plan, which is a selected case from the current backlog, is ever executed. At the subsequent iteration, the cases that actually arrive are folded into the backlog cohorts. Therefore, anticipated arrivals influence the executed action without themselves being acted on. Harm is valued over an estimated residence time

$$L = \max\big(H,\ \omega\,|\text{backlog}|/C\big), \qquad (9)$$

where $|\text{backlog}|/C$ is an estimate of the necessary steps to drain the current backlog at capacity if cases stopped arriving and $\omega \ge 1$ is a tuned inflation factor. The estimated drain time undervalues the true residence time since new arrivals will replenish the queue and continue competing for the same available capacity. Hence, a deferred case will await review for a longer period of time than the current backlog alone can imply. This downward bias is corrected by ω, and the floor function applied to $\max(\cdot, H)$ ensures L at least spans the horizon. Since L may exceed H, the harm term behaves as a terminal cost-to-go that potentially extends beyond the plan window rather than a cost that is strictly within the horizon. Moreover, ω is not an environmental parameter. Rather, it is a policy-side hyperparameter that is tuned on the base scenario.

2) The finite-horizon control problem: The decision variables are the mass of each cohort reviewed at each horizon step, $x_{g,k} \ge 0$ for $g \in \{1,\dots,G\}$ and $k \in \{0,\dots,H-1\}$, and the mass of each forecast arrival class reviewed at each step, $v_{j,k} \ge 0$. The controller minimizes the predicted cost (5) summed over horizon H. The cost of reviewing no cases over the horizon is a constant, which is independent of (x, v). Therefore, the minimization problem can be recast as *maximizing* the total predicted cost *reduction* achieved by the reviews, which is a linear function of the decision variables. Reviewing one unit of cohort g at step k ceases its harm for the remaining residence time, which saves $w_{\text{harm}}\hat{h}_g\Delta(L-k)$. Moreover, if such a review is done strictly before an SLA breach within the horizon ($k < \sigma_g \le H$), the one-off breach penalty w_{sla} is averted. If however, a cohort has already breached the SLA or will breach beyond the horizon, no SLA term is contributed. The resulting linear program (LP) is defined by

$$\max_{x,\,v\,\ge\,0} \quad \sum_{g=1}^{G}\sum_{k=0}^{H-1} \big[w_{\text{harm}}\hat{h}_g\Delta(L-k) + w_{\text{sla}}\mathbf{1}\{k<\sigma_g\le H\}\big]\, x_{g,k} + \sum_{j=1}^{J}\sum_{k=0}^{H-1} w_{\text{harm}}\hat{h}_j\Delta(L-k)\, v_{j,k} \qquad (10)$$

$$\text{s.t.} \quad \sum_{g=1}^{G} x_{g,k} + \sum_{j=1}^{J} v_{j,k} \le C, \quad \forall k, \qquad (11)$$

$$\sum_{k=0}^{H-1} x_{g,k} \le m_g, \quad \forall g, \qquad (12)$$

$$\sum_{k'=0}^{k} v_{j,k'} \le A_{j,k}, \quad \forall j,k, \qquad (13)$$

$$x_{g,k} \ge 0, \quad v_{j,k} \ge 0, \quad \forall g,j,k. \qquad (14)$$

Constraint (11) caps reviews per step at the capacity C, constraint (12) prevents reviewing more of a cohort than the number of cases that exist within it, and constraint (13) enforces arrival causality, so cumulative reviews of a class cannot exceed the cumulative amount expected to have arrived by that step. Arrival classes do not possess an SLA term in (10) because new arrivals cannot violate the SLA within the horizon under the requirement that $H < \min_c d_c$.

3) Execution and properties: Problem (10)–(14) is solved with the open-source HiGHS LP solver [18]. By the MPC control law, the first-step allocation $x^\star_{g,0}$ alone is executed. Following this, it is rounded to an integer count per cohort. Within each cohort, the best actual cases are drawn first

(by highest $\hat{h}$, then soonest deadline). The rounding step rounds down the fractional allocation and delivers the leftover capacity to the largest fractional parts, capped by cohort mass to ensure the executed number of reviews never exceeds C. In Mode B, the auto-action counts are also rounded down, which keeps $\sum_g \Pr(y{=}0)_g\, a_g$ within the false-positive-rate budget after rounding and each cohort's mass is split, ensuring that any given case is never both reviewed and auto-actioned. The remainder of the plan is discarded and at the next step, the entire program is re-solved. The fluid relaxation restores a fixed-dimension state for a variable-cardinality backlog, and the option to review no cases ($x = v = 0$) is always feasible. Therefore, a feasible plan will always exist. This formulation is most similar to a standard economic-MPC problem rather than requiring the terminal set constructions of [9], [19].

4) Mode B: When auto-action is allowed, a per-cohort variable $a_g \geq 0$ (mass auto-actioned at the current step) is added. Though it does not consume review capacity, it saves the same harm and breach cost as a review would; moreover, it incurs the expected false-positive penalty $w_{\text{fp}} \Pr(y{=}0 \mid s, c)$ per unit. It is bounded by a single cumulative *false-positive-rate* row $\sum_g \Pr(y{=}0 \mid \cdot)\, a_g \leq \beta$, where β is the wrongful-action budget remaining under the bound ε. The evaluated fixed-threshold index baseline cannot represent this coupling constraint, whereas the LP satisfies it exactly. The comparisons in this paper are conducted against the standard fixed-threshold rule, though constraint-aware index methods (Lagrangian or primal-dual schemes) can also incorporate such coupled budgets.

V. Data Anchoring

The model is separated into an *observable* layer that can be validated by real data, and a *hidden* layer that cannot. The observable layer is calibrated using public data with the classifier cross-validated over several folds. Specifically, the observable layer (the category mix) is anchored to a one-week sample of *non-automatically-detected* moderation decisions from the EU DSA Transparency Database [13] (Amazon Store Statements of Reasons, 2026-07-07 to 07-13, filtered on `automated_detection`="No", 390,599 records. Since this is the *detection* field, it identifies cases that were flagged by a human, although not necessarily actioned as such. It is used as a proxy for selecting those cases that were subject to human review. Since the database records moderation decisions rather than an incoming queue, the anchored quantity is the *category proportions* across low/mid/high-severity tiers ($0.34/0.29/0.37$). The per-tier severity *magnitudes* $[1, 3, 10]$, arrival timing, deadlines, prevalence, per-case harm ground truth and capacity are not present in publicly-available data, so they are assigned nominal values and are swept across a plausible range. It will be seen that the paper's conclusions are supported by this robust sweep. The hidden layer's *shapes* (signal quality and the heavy-tailed exposure) are anchored to a public fraud dataset [20] by training a classifier and utilizing its out-of-fold score distribution and the tail representing the transaction amount, which is an analog for harm magnitude. The score-magnitude independence assumption is tested by omitting the amount from the classifier features. The results show that the fitted score is weakly *negatively* associated with the transaction amount. Thus, treating score and magnitude as independent is a mild idealization. A sweep of this correlation (base environment, 20 seeds) while ensuring the marginal exposure remains fixed confirms the conclusions are indeed robust to the previously stated assumption: as the score becomes more predictive of magnitude, the realizable policies improve, and the gap to FI-greedy approach narrows ($3.3\times$ at correlation -0.6, $3.1\times$ at independence, to $2.4\times$ at $+0.6$), while the FI-greedy cost itself is unchanging. At the observed sign (≈ -0.15), the gap ($3.3\times$) is slightly *wider* than under independence, making the assumption mildly optimistic, but the policy ordering remains unchanged throughout. As will be seen, the harm-aware rules beat highest-score by about $2\times$ and the MPC tracks $c\mu$ at every correlation. Hence, although the coupling alters the size of the aforementioned gap, it does not affect the qualitative result.

The quality of the classifier signal is measured by two threshold-free metrics, namely the area under the ROC curve (AUC) and expected calibration error (Brier score). Since the AUC is invariant to any monotone transformation of the scores, it is blind to both calibration and *resolution* (the spread of the predicted probabilities). The Brier decomposition of Murphy [21], $\text{Brier} = \text{Reliability} - \text{Resolution} + \text{Uncertainty}$, isolates the resolution term that a magnitude-planning controller depends on, but is discarded by AUC (reported per classifier in Table V). Since the MPC controller utilizes the magnitude signal as an estimate of risk and not just the relative ordering, the signal must necessarily be characterized by both the AUC and Brier score.

VI. Experimental Study

A. Protocol

All comparisons use *paired* arrivals. As such, for a given seed, the environment is reset with that particular seed, so every policy faces the identical arrival stream and mark realization. This causes the large common-mode harm variance to be canceled out by the paired per-seed difference. The absolute costs per policy (Tables I, IV) are reported as seed means. The paired per-seed differences that reveal the comparisons are reported with 95% confidence intervals. All intervals use the Student-t quantile $t_{.975,n-1}$ at the stated seed count n, which is the best metric given the small sample size. A difference is called *significant* (*) when its interval excludes zero. The dual-mode non-inferiority claim (Sec. VIII) is evaluated on held-out data using a pre-specified margin, with Bonferroni-simultaneous bounds across the regime panel. The broader regime-sweep and ablation studies are exploratory, so their results are interpreted as estimations of effect size rather than a set of controlled tests. Since the harm magnitudes are heavy-tailed, the headline MPC$-c\mu$ comparison and all six dual-mode regimes are cross-checked with a paired percentile bootstrap ($10{,}000$ resamples) and a leave-one-seed-out sweep. The bootstrap 95% intervals agree with the Student-t intervals on every comparison, and when any single seed is removed, every significant result remains sign stable. Thus, no conclusion is dependent upon one or two extreme episodes. The

only sign changes that occur under leave-one-out correspond to statistically insignificant comparisons. Each episode commences from an empty backlog and runs for T=336 steps. The initial steps therefore include a fill-in transient, and ρ=1.00 denotes finite-horizon critical load (not long-run stability). The SLA-violation rate is computed over all arrived cases; a case is counted as a violation once its wait exceeds d_i, irrespective of whether it is subsequently resolved. Mean and p_{95} wait are only computed over *resolved* cases, so cases that are still pending at T are right-censored and excluded from the wait statistics. Harm and SLA costs are booked within the episode alone, while the residence heuristic (9) supplies the terminal continuation value utilized by the planner. The reported RL rows arise from a seed-pinned trained model. Unlike the deterministic heuristic, index, FI-greedy, and MPC policies, PPO training is not guaranteed to be bit-reproducible, therefore, the RL numbers must be interpreted as subject to training noise.

Prior to any policy comparison, environment parameters are determined and are never tuned in favor of any policy. Following Sec. V, the category/severity mix and the signal/harm *shapes* are anchored to public data, whereas the queuing, deadline, and harm-accrual *dynamics* are treated as assumptions and swept. Hence, the conclusions drawn from this work are robust as opposed to being valid only for any single operating point. The primary metric is total episode cost (5). Additionally, SLA-violation rate, waiting-time tail, and the multiplicative gap to the FI-greedy baseline are also reported.

B. Base scenario: the harm-aware/harm-blind chasm

Table I and Fig. 1 provide the base operating point (assumed parameters, a moderately informative synthetic signal, 30 seeds). The dominant finding is an order-of-magnitude separation: the harm-blind rules (FIFO, EDF), which allocate reviews based on deadlines regardless of harm, incur costs that are 30–50× higher than those that result from the harm-aware rules. Among the harm-aware policies, the MPC, $c\mu$, and Whittle are within about a factor of three of FI-greedy (Whittle at 3.1×). It is seen that the paired test is decisive where the marginal intervals are not. MPC *ties* $c\mu$ (paired MPC$-c\mu$ $= -214$, not significant) but modestly, and significantly beats Whittle (-711^{*}). Under saturation, Whittle's one-off breach term slightly misfires. This provides the ordering $c\mu \approx$ MPC $<$ Whittle. The PPO agent matches raw highest-score on cost, but it achieves the lowest wait time by far. It is treated as an *illustrative* baseline, rather than a tuned competitor. This low waiting time is consistent with aggressively draining the backlog while under-weighting the hidden, heavy-tailed harm that the reward exposes only noisily.

C. Where the MPC wins: load, SLA weight, and burstiness

A sweep over review capacity (equivalently, load ρ) and the SLA weight on the fixed environment reveals a clean structure (Table II, Fig. 2). The MPC ties with $c\mu$ at the harm-dominated base point (ρ=1.33, low SLA weight). The MPC significantly surpasses both index rules (by as much as $\sim$94%) as breaches become *preventable with foresight* (capacity near the arrival rate) and the SLA weight grows. The myopic $c\mu$ rule's cost rises sharply on unmitigated breaches near critical load, whereas the MPC and Whittle remain controlled. Across the preventable-breach band, the MPC follows the shape of FI-greedy most closely.

TABLE I
BASE SCENARIO, 30 SEEDS.

Policy	Total cost	SLA rate	p_{95} wait	gap
FIFO	830,025	0.381	91.6	145×
EDF	500,482	0.360	97.9	88×
Highest score	32,346	0.258	76.5	5.7×
RL (PPO)	31,527	0.246	14.5	5.5×
Whittle	17,733	0.230	53.7	3.1×
$c\mu$	17,236	0.254	76.6	3.0×
MPC	17,022	0.220	60.6	3.0×
FI-greedy	5,701	0.168	72.6	–

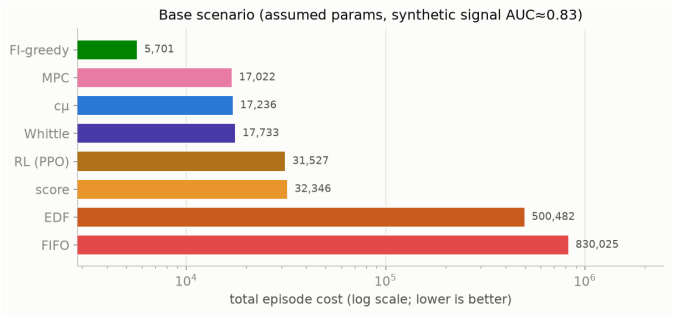

Fig. 1. Base scenario, total cost per policy (log scale, 30 seeds). The order-of-magnitude split between harm-aware (FI-greedy, MPC, $c\mu$, Whittle) and harm-blind (EDF, FIFO) policies is the dominant effect. RL and raw score sit between.

Under the MMPP and Hawkes processes at equal mean load, the seed count is increased to n=60 to adequately power the smaller MPC−Whittle comparison (Fig. 3). At a high SLA weight, the MPC significantly outperforms both index rules under *all three* arrival processes, including the smooth NHPP. Therefore, the advantage is attributable to the SLA weight, as opposed to burstiness. The bursty point estimates trend mildly larger, however, their intervals overlap that of the smooth process. Therefore, no significant burstiness-specific effect can be established. The results are considerably nuanced at low SLA weight. MPC ties $c\mu$ except under MMPP surges, where its performance is significantly *worse*. This failure mode is analyzed below.

D. A realistic signal: the cost of exposure blindness

As mentioned, the category mix and the signal/harm shapes are anchored to public data. When this anchoring is carried out for a real classifier at AUC 0.65 in particular, the study's headline effects are produced (Table IV, Fig. 4). The gap to FI-greedy widens from the synthetic $\sim 3\times$ to roughly 19×. This is sharpened by two observations. Ranking by the raw score performs almost as poorly as ignoring the signal altogether (highest-score costs $\sim$1.28M, near FIFO). Ranking by the Bayesian posterior (6), on the other hand, is $\sim 6.4\times$ cheaper (the value of the Bayes step alone). Category-severity weighting adds a further $\sim$ 1.2×, demonstrating that the

TABLE II
REGIME SWEEP (BASE ENVIRONMENT, 20 SEEDS, PAIRED).

ρ	w_{sla}	$c\mu$	MPC$-c\mu$	MPC$-$Whittle
1.33	5	18,038	-78 ± 620	$-858 \pm 740^*$
1.11	5	6,639	$-950 \pm 311^*$	$-1{,}331 \pm 749^*$
1.11	20	20,245	$-5{,}353 \pm 887^*$	$-3{,}886 \pm 575^*$
1.11	50	47,456	$-22{,}938 \pm 1169^*$	$-21{,}742 \pm 1293^*$
1.00	20	8,618	$-7{,}359 \pm 454^*$	$-478 \pm 196^*$
1.00	50	20,415	$-19{,}157 \pm 1018^*$	$-992 \pm 272^*$

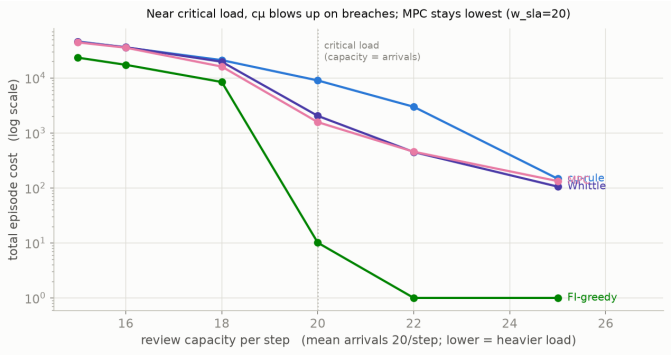

Fig. 2. Total cost versus review capacity at a high SLA weight (log scale; lower capacity is heavier load). As the load approaches the critical point, $c\mu$'s cost greatly escalates due to SLA breaches, while the MPC and Whittle remain an order of magnitude lower. The MPC closely tracks FI-greedy across the near-critical band.

posterior harm rule costs $\sim 8\times$ less than raw-score ranking. Secondly, by improving the classifier from AUC 0.65 to 0.90 (a gradient-boosted model, Table V), the realizable cost is lowered by $\sim 34\%$, and the information gap is decreased to $\sim 13\times$. A substantial gap persists, since part of it remains inherently beyond the reach of any label classifier. It reflects unobserved harm *magnitude* (hidden exposure), instead of uncertainty about the label itself. To allocate the gap for this policy family, a factorial information attribution, built around a single fixed priority rule, is utilized: the marginal-cost index $\mu_i(w_{\text{harm}}\hat{h}_i + b_i)$ where b_i is the load-scaled deadline term shared by the Whittle-inspired rule and FI-greedy. This rule is run under four clairvoyant states that only change the harm estimate $\hat{h}_i$. With the posterior $\hat{h}_i$, this index reduces exactly to the Whittle-inspired rule, so the "neither" row equals its Table I value. And when it is set to the true $y_i h_i$, the index becomes FI-greedy (the "both" row). The cells only differ in information. When the true label alone is observed, the entire gap is almost closed. Observing the magnitude alone closes far less of it. The two interact strongly. The interaction is allocated by a Shapley attribution whose shares sum to the gap, and the fully-observed state reproduces FI-greedy. This attributes $\sim 80\%$ of the gap to label uncertainty and $\sim 20\%$ to magnitude. Thus, magnitude constitutes a smaller share, but it cannot be reduced by a classifier. A better classifier shrinks the label component, whereas magnitude can only be mitigated by learning it directly. (Sec. VII).

TABLE III
WITHIN-POLICY FACTORIAL INFORMATION ATTRIBUTION OVER THE INFORMATION GAP (30 SEEDS).

Information state	Base	Real-anchored
neither (posterior, mean mag.)	17,733	162,911
label only	5,862	10,663
magnitude only	12,997	98,253
both (FI-greedy)	5,701	8,513
gap (neither $-$ both)	12,032	154,398
Shapley label share	80%	78%
Shapley magnitude share	20%	22%
label$\times$magnitude interaction	38%	40%

TABLE IV
REAL-ANCHORED OPERATING POINT (AUC 0.65, 30 SEEDS).

Policy	Total cost	SLA rate	gap
FIFO	1,900,105	0.536	223$\times$
Highest score	1,284,138	0.282	151$\times$
EDF	1,242,371	0.589	146$\times$
RL (PPO)	187,573	0.250	22$\times$
MPC	165,486	0.229	19$\times$
Whittle	162,911	0.249	19$\times$
$c\mu$	162,573	0.254	19$\times$
FI-greedy	8,513	0.206	–

E. When the MPC loses, and why

The experiments show that MPC does not consistently perform at least as well as the myopic rules. In two regimes, it emerges to be significantly *worse* than $c\mu$. The first pertains to the crushed real signal at heavy overload. An AUC sweep at two loads (Fig. 5) reveals that the MPC's advantage over $c\mu$ is robust and grows with AUC at *critical* load. At *heavy overload* however, it straddles zero, and with the real classifier's crushed, near-zero score, the MPC is significantly worse ($+5{,}136$ per episode). The MPC manages to achieve a tie at the same AUC but with a higher-resolution score model. The culprit is the score distribution's *resolution*, not its AUC (Sec. V). Resolution collapses from 0.012 at AUC 0.90 to 0.0006 at AUC 0.65 (Table V) while calibration remains intact. This is precisely the metric that AUC cannot capture. The second regime is an MMPP surge at zero SLA weight (Fig. 3, left), where the MPC's smooth-mean forecast misplans against an unanticipated load. In both regimes, the plan's *anticipatory* value of averting breaches within the horizon is negligible, for heavy overload makes breaches impossible to avert, and zero SLA weight makes them futile. Hence, the MPC is effectively chasing the same harm as FI-greedy, but it achieves this through cohort binning, a fluid relaxation, and a forecast. Their combined approximation error allows the exact greedy rule to outperform the MPC. A myopic rule possesses no model, so there is no scope for it to be wrong. This is a practical example of a plan being undone by its own faulty assumptions, and it becomes the constructive starting point of Sec. VIII.

F. Robustness of findings

Checks are used to guard against over-reading the point estimates. The first is a one-factor-at-a-time sweep over the

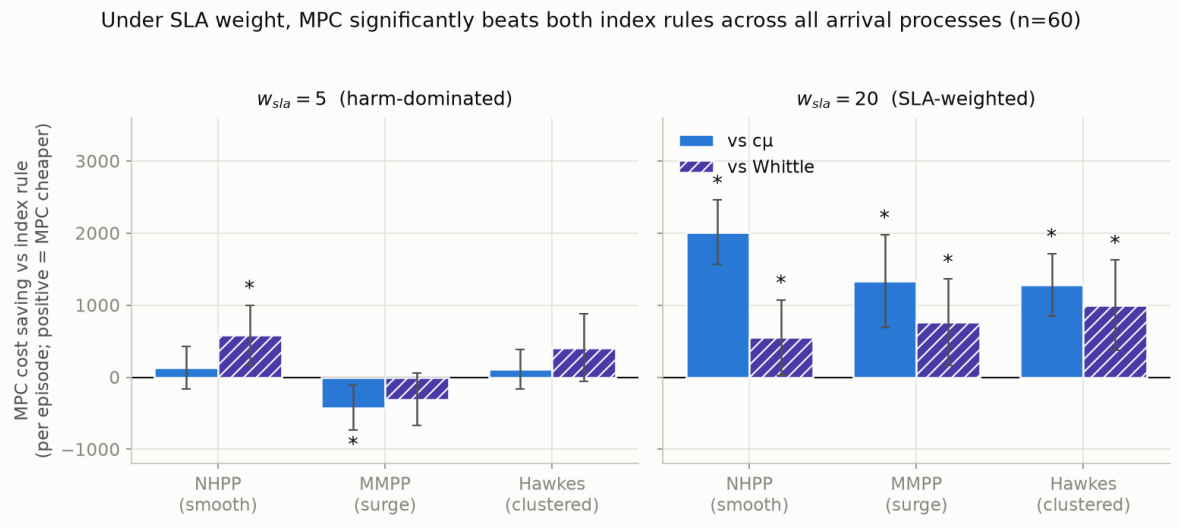

Fig. 3. MPC cost saving over each index rule (positive means MPC cheaper; * marks a 95% Student-t CI clear of zero, either direction), n=60. *Right* (w_{sla}=20): the MPC significantly beats both index rules under all three arrival processes, the robust result, not burstiness-specific. *Left* (w_{sla}=5): mostly ties, except the MPC edges Whittle under NHPP and is significantly *worse* than $c\mu$ under MMPP surges (the starred bar below zero).

TABLE V
CLASSIFIER QUALITY ON THE PUBLIC FRAUD DATA, OUT-OF-FOLD (5-FOLD CV); MURPHY BRIER DECOMPOSITION, WITH UNCERTAINTY (BASE-RATE VARIANCE) 0.029 FOR ALL ROWS.

Classifier	AUC	Brier	ECE	Reliab.	Resol.
Logistic, 12 features	0.65	0.028	0.002	0.0000	0.0006
Logistic, 200 features	0.82	0.024	0.004	0.0001	0.0053
GBM, 200 features	0.90	0.017	0.003	0.0002	0.0116

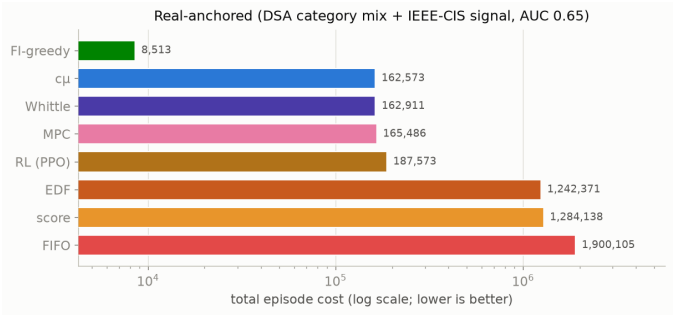

Fig. 4. Real-anchored total cost per policy (AUC 0.65, log scale, 30 seeds). The information gap opens to ≈ 19×; the posterior harm rules beat raw score ≈ 8×. Here the MPC sits marginally *behind* the index rules, as the crushed signal blunts its magnitude planning (Sec. VI-E).

nominally assumed parameters (signal AUC, severity, exposure spread, harm prevalence). This sweep shows that the qualitative ordering of the policies is preserved (clairvoyant benchmark below all; harm-aware ≫ harm-blind) in every cell. The MPC's point-estimate advantage persists throughout, except at the harm-dominated extremes, where its SLA-awareness edge is expected to fade. The second check is a nested selection of the MPC's one tuned hyperparameter (the residence-time inflation), which is fit on validation seeds, and is reported on *disjoint* test seeds. This confirms that the MPC's advantage over Whittle holds as a broad plateau, but also

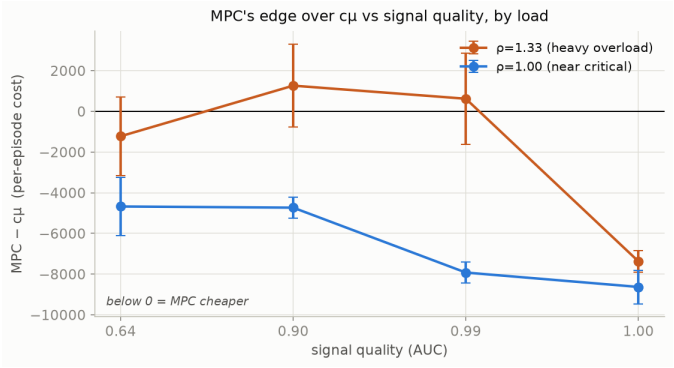

Fig. 5. Paired MPC$-c\mu$ cost versus signal quality at two loads (negative means MPC cheaper; bars are 95% CIs). At near-critical load the MPC's edge is robust and grows with AUC; at heavy overload it straddles zero, winning clearly only at near-perfect signal. This is the MPC caveat made visual.

reveals that the $c\mu$ margin at the moderate regime is directional and *not* significant off the tuning seeds. On the real-anchored signal, the same nested selection shows the MPC attaining a tie with both index rules on held-out seeds. A third check constitutes a horizon sweep ($T \in \{168, 336, 720\}$). The qualitative ordering remains stable, with the harm-aware policies clustered at 1.7–2.7× FI-greedy at every tested horizon. Thus it may be ensured that the results are not merely a two-week cutoff artifact. The MPC's paired margin over $c\mu$ at this moderate regime however, presents another narrative. It becomes indistinguishable at T=720 as episode-cost variance grows (point estimate +2452, CI ±5721). This is a reminder that MPC's decisive advantages occur near-critical load, and not in this particular regime.

G. Constrained (Mode-B) control

When auto-action is permitted under a cumulative false-positive-rate budget (ε=0.05), the constrained MPC exploits

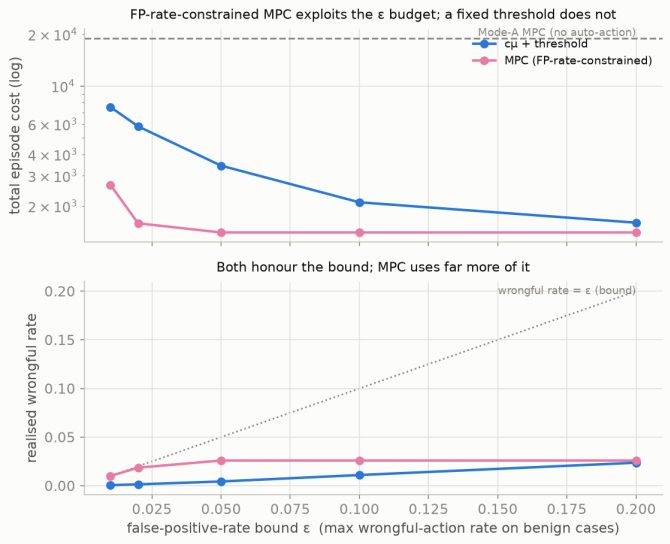

Fig. 6. Mode B versus the false-positive-rate bound ε. *Top*: the constrained MPC (pink) reaches far lower cost than the fixed-threshold $c\mu$ baseline (blue) and the Mode-A MPC (dashed). *Bottom*: both honor the realized wrongful-action bound, but the MPC uses far more of the available budget: it auto-actions up to the constraint, which a fixed threshold cannot track.

the free channel to reduce cost while honoring the wrongful-action bound. It surpasses a Mode-A MPC (no auto-action) and a fixed-threshold $c\mu$ baseline (Fig. 6). $c\mu$ acts as the representative fixed-threshold index baseline. Whittle would share the limitation, as it sets a threshold on each case's score individually while the bound is an aggregate over cases. Index methods that are aware of constraints can incorporate coupled budgets. However, the baseline evaluated here is the standard fixed-threshold rule. The fixed threshold either violates ε on account of excess permissiveness, or squanders the channel instead. The LP encodes the aggregate expected-count budget as a single coupling row, therefore, it is capable of exhausting it exactly. In this setting, MPC possesses a *structural* advantage, not merely a regime-specific one. A fixed-threshold index rule, unlike the LP, cannot represent an aggregate constraint that couples decisions across cases. The realized wrongful-action rate is the number of auto-actioned benign ($y{=}0$) cases divided by the true number of benign cases (only revealed after the fact). Since the controller cannot observe this denominator online, the LP optimizes over expectation. At step t, the remaining budget is $\beta_t = \varepsilon \sum_{i \in \mathcal{S}_t} \Pr(y_i{=}0 \mid s_i, c_i) - w_t$, where $\mathcal{S}_t$ is the set of cases seen so far and w_t the expected wrongful count already spent. The online denominator is therefore the posterior-expected benign population observed to date. Accordingly, the constraint bounds an expected *count*.

The realized rate in a finite episode is therefore a random draw. Across 30 seeds, the constrained MPC's realized rate remains below ε throughout (mean 0.025, p_{95} and maximum 0.030; no seed exceeds $\varepsilon{=}0.05$), while using far more of the budget than the conservative fixed threshold (realized 0.004). Guaranteeing this bound on every single episode would necessitate a chance-constrained or robust-budget formulation; this is an extension left for future work.

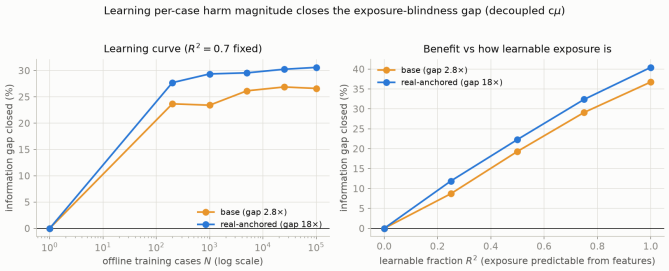

Fig. 7. Decoupled harm-magnitude learning, base and real-anchored environments. *Left*: fraction of the information gap closed versus offline training size (benefit saturates quickly). *Right*: gap closed versus the learnable fraction R^2: monotone, zero at $R^2{=}0$, ceiling $\sim 40\%$ (magnitude's main-effect gap closure; Table III gives the attributed share).

VII. Learning the Harm Magnitude

Given that a portion of the cost arises from unobserved magnitude and cannot be addressed by an improved classifier, a natural question is whether magnitude is nonetheless *learnable*. In the base model, exposure is purely independent noise, which is provably unlearnable; he best possible estimate is the mean. To study learnability, exposure is given observable structure: an observable covariate x_i with $\log E_i = \mu_e + a\,x_i + b\,z_i$, $a^2 + b^2 = \sigma_e^2$, so the marginal law is unchanged for any *learnable fraction* $R^2 = a^2/\sigma_e^2$. This is measurable by a deployer. One simply regresses realized harm on observable case features and reads off R^2.

Reviewing a case reveals its harm, which yields a training pair (x_i, E_i). A nonparametric estimator (a gradient-boosted regression of $\log E_i$ with Duan's smearing retransformation [22], assuming nothing about the exposure family) is fitted, and its per-case prediction is substituted for $\bar{E}$ in (7). The resulting improvement rises monotonically with R^2, and it is exactly zero at $R^2{=}0$ as the learner invents no benefit where none exists, and it closes up to $\sim 40\%$ of the information gap at $R^2{=}1$ (Fig. 7).

This $\sim 40\%$ represents magnitude's *main effect* against the fully-blind baseline, and it overlaps with label uncertainty. The factorial decomposition (Table III) reveals a large label–magnitude interaction. When that interaction is properly attributed via Shapley values, the actual share of magnitude decreases to $\sim$ 20%, against $\sim$ 80% for label uncertainty. Magnitude is the smaller component, yet it cannot be reduced by a classifier. Online learning recovers most of what remains.

A coupled learner that trains online from the policy's own review log is tested next. The log is selection-biased, for it preferentially reviews high expected-harm cases. Therefore, its training log oversamples high-exposure cases. The coupled learner tracks the clean offline learner to within one to two percentage points at every R^2, and it converges within a single episode (Fig. 8). In practice, the selection bias is not particularly consequential. A triage system can learn the needed harm magnitude directly from its own operation; it does not require a separate data-collection phase. The claim is not one of real-world efficacy, rather, it bridges simulation and real-world deployment through the deployer-measurable R^2.

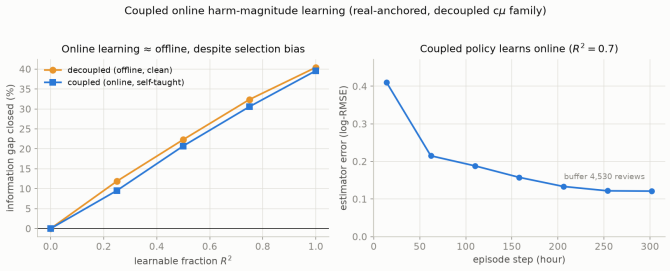

Fig. 8. Coupled (online) versus decoupled learning, real-anchored. *Left*: the self-taught online learner tracks the clean offline learner to within 1–2 points at every R^2, despite its selection-biased log. *Right*: the coupled policy's estimator error falls within a single episode as its review log grows.

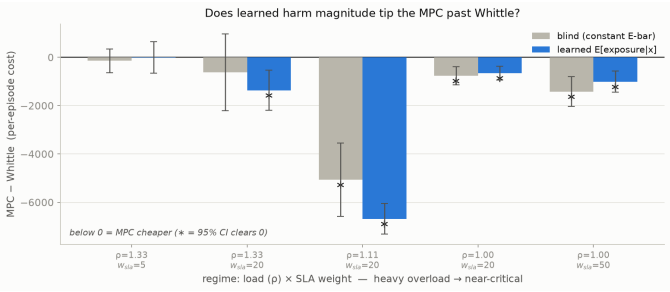

Fig. 9. Paired MPC−Whittle cost across regimes, blind (constant mean exposure) versus learned per-case magnitude (R^2=0.7; negative means MPC cheaper, * significant). Learned magnitude tips the MPC from a tie to a significant win at overload with SLA weight and widens its lead at moderate overload; at pure critical load, where the edge is timing rather than magnitude, it slightly narrows but keeps the existing lead.

Feeding the learned per-case magnitude to the MPC alters its behavior in two ways. It enters the cohort harm rate directly, and it becomes an extra cohort bin, providing the LP a lever (allocating by magnitude) that a simple rank rule does not possess. It expands the range of conditions under which the MPC defeats Whittle. At overload, with SLA weight, the learned-magnitude MPC converts a tie into a significant advantage. At moderate overload, it widens the lead further (Fig. 9). This does not indicate that MPC is superior in all circumstances, for in the harm-dominated regime with no SLA weight, both rules still tie. This is consistent with the broader pattern seen throughout the paper: MPC's advantage can only materialize when there exists a preventable outcome for foresight to act on.

VIII. A Dual-Mode Controller

The MPC's two failure regimes share a single cause that can be computed online. In both, the plan's anticipatory value is negligible. After each solve, the controller reports

$$\phi = \frac{\text{SLA-term savings in the solved plan}}{\text{total savings}}, \qquad (15)$$

which measures the share of predicted savings attributable to averting breaches. When $\phi < \tau$, the controller executes the exact $c\mu$ action. Otherwise, it executes the full MPC action. The gate only keys on the plan's internal structure, so it cannot be tuned to the test set. It fires precisely when foresight is idle, since heavy overload makes breaches impossible to avert

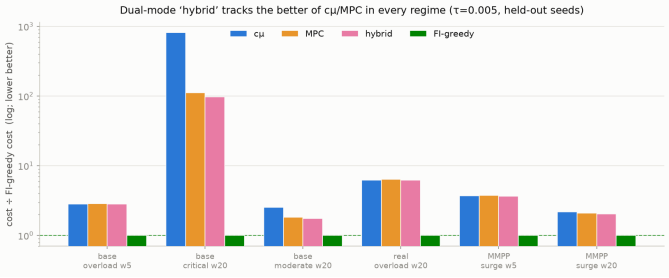

Fig. 10. Cost relative to FI-greedy (log scale) per regime, held-out seeds. The dual-mode controller ("hybrid") tracks the lower of $c\mu$ and MPC in every regime: it keeps the order-of-magnitude MPC win at near-critical load, where $c\mu$ blows up, and recovers to $c\mu$ in the two regimes where naive MPC loses.

within the horizon, driving $\phi \to 0$; a low SLA weight has the same effect. Gating on the naively observable breach count would be incorrect, as breaches are frequent under overload, but they are *not* preventable. Preventability is the valuable property, and the within-horizon avertable term alone captures that.

The threshold τ is a single hyperparameter. It is selected on validation seeds over a six-regime panel spanning the MPC's wins and both of its failure modes (minimizing mean regret against the better of $\{c\mu, \text{MPC}\}$), and results are reported on *disjoint* test seeds. The selected value is $\tau = 0.005$. Table VI presents the held-out result.

When the dual-mode controller is evaluated under a pre-specified non-inferiority test (margin 2% of the better pure policy's cost, one-sided Bonferroni-simultaneous 95% bounds across the six regimes), it is established as non-inferior in four regimes, and shown to be significantly cheaper than *both* pure rules in three (base moderate and the two MMPP-surge regimes). The gain in the moderate regime arises because a single episode is not uniformly one regime, therefore, per-step switching rides the transition from drainable, where foresight is beneficial, to saturated, where foresight is idle. The hybrid is not significantly worse than either pure rule in any regime. Base heavy-overload and base near-critical are two regimes that did not achieve non-inferiority. Both are low absolute-cost regimes where the 2% relative margin is small in absolute terms (317 and 21 cost units respectively) and the paired difference, within that margin in point estimate ($+114$ and -13 against the better pure policy), is simply not resolved at this power. Neither result represents a significant disadvantage for the dual-mode controller. On the contrary, Tthe failures associated with naive-MPC are removed. At real overload, the loss against the MPC is erased ($-4{,}898$). What remains against $c\mu$ is a non-significant tie. FIFO, EDF, highest-score, and Whittle are dominated by $c\mu$ or the MPC in every tested regime. Therefore, displaying no disadvantage against the better of these two implies no disadvantage against any *realizable* pure policy at all. FI-greedy, being clairvoyant, is excluded from this claim as it is not realizable. The gate must be understood as a principled limiting-case correction, as opposed to a tuned regime switch, since the controller should reduce to the myopic optimum exactly when the horizon carries no information. Under this logic, the gate converts

TABLE VI
DUAL-MODE HELD-OUT EVALUATION (30 TEST SEEDS, PAIRED). *: HYBRID SIGNIFICANTLY CHEAPER. δ: PRE-SPECIFIED NON-INFERIORITY MARGIN (2% OF THE BETTER PURE POLICY'S COST); UB: ONE-SIDED BONFERRONI-SIMULTANEOUS 95% UPPER BOUND ON (HYBRID − BETTER PURE); NI PASSES WHEN UB$<\delta$.

Regime	hybrid$-c\mu$	hybrid$-$MPC	δ	UB	NI
base overload, low SLA	-114 ± 161	$+114 \pm 507$	317	$+745$	
base near-critical, SLA	$-7{,}396 \pm 352^*$	-13 ± 129	21	$+147$	
base moderate, SLA	$-5{,}798 \pm 572^*$	$-375 \pm 136^*$	286	-206	✓
real overload *(fail 1)*	$+77 \pm 1148$	$-4{,}898 \pm 1507^*$	3,712	$+1{,}634$	✓
MMPP surge, low SLA *(fail 2)*	$-142 \pm 128^*$	$-848 \pm 457^*$	569	$+16$	✓
MMPP surge, SLA	$-2{,}039 \pm 822^*$	$-728 \pm 679^*$	1,074	$+116$	✓

two negative results into no detected disadvantage, with strict improvement where the interval is precise enough.

IX. CONCLUSION

This paper has shown that it is certainly valuable to treat an enforcement backlog as constrained model predictive control problem, but the efficacy is conditional. Partial observability is the dominant cost. Within it, unobserved harm *magnitude* is a smaller but classifier-irreducible share of the gap ($\sim 20\%$ by a factorial attribution, and a larger $\sim 40\%$ main effect that overlaps with label uncertainty). Under the score-magnitude independence assumption, this share cannot be mitigated by a better *label* classifier. It is, however, partly reclaimable by learning it online from the operation's own reviews. When the deadline constraint carries weight and load is near-critical, model predictive control beats the strong myopic rules. However, when its internal model is misleading, it may perform worse than a myopic rule. A dual-mode controller that reverts to the myopic rule precisely when foresight is idle is, on held-out seeds, never significantly worse than the better of the pure $c\mu$ and MPC policies. It also retains the controller's advantage where foresight is valuable. Future work includes wiring the learned magnitude estimator and the false-positive-rate-constrained auto-action mode into the dual-mode controller, and extending the real-data anchor across multiple platforms.

REFERENCES

[1] C. Gocmen, T. Lykouris, D. Sinha, and W. Weng, "Scheduling in queueing systems with uncertain and evolving holding costs," *arXiv preprint arXiv:2505.21331*, 2025.

[2] T. Lykouris and W. Weng, "Learning to defer in congested systems: The AI–human interplay," in *Proc. 25th ACM Conf. on Economics and Computation (EC)*, 2024.

[3] J. Lee, H. Namkoong, and Y. Zeng, "Design and scheduling of an AI-based queueing system," *arXiv preprint arXiv:2406.06855*, 2024.

[4] J. A. Van Mieghem, "Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule," *The Annals of Applied Probability*, vol. 5, no. 3, pp. 809–833, 1995.

[5] R. Atar, C. Giat, and N. Shimkin, "The $c\mu/\theta$ rule for many-server queues with abandonment," *Operations Research*, vol. 58, no. 5, pp. 1427–1439, 2010.

[6] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, pp. 287–298, 1988.

[7] Z. Yu, Y. Xu, and L. Tong, "Deadline scheduling as restless bandits," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2343–2358, 2018.

[8] C. Nguyen, M. Bhuyan, and E. Elmroth, "Taming cold starts: Proactive serverless scheduling with model predictive control," in *Proc. 33rd IEEE Int. Symp. on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, 2025, arXiv:2508.07640.

[9] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[10] J. B. Rawlings, D. Q. Mayne, and M. M. Diehl, *Model Predictive Control: Theory, Computation, and Design*, 2nd ed. Nob Hill Publishing, 2017.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[12] G. Wilson, G. Goh, Y. Jiang, A. Gupta, J. Wang, D. Freeman, and F. Dinuzzo, "Predictive response optimization: Using reinforcement learning to fight online social network abuse," in *Proc. 34th USENIX Security Symposium (USENIX Security)*, 2025, pp. 3239–3256.

[13] R. Kaushal, J. van de Kerkhof, C. Goanta, G. Spanakis, and A. Iamnitchi, "Automated transparency: A legal and empirical analysis of the DSA transparency database," in *Proc. ACM Conf. on Fairness, Accountability, and Transparency (FAccT)*, 2024, pp. 1121–1132.

[14] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI Gym," *arXiv preprint arXiv:1606.01540*, 2016.

[15] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG *et al.*, "Gymnasium: A standard interface for reinforcement learning environments," *arXiv preprint arXiv:2407.17032*, 2024.

[16] J. D. C. Little, "A proof for the queuing formula: $l = \lambda w$," *Operations Research*, vol. 9, no. 3, pp. 383–387, 1961.

[17] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.

[18] Q. Huangfu and J. A. J. Hall, "Parallelizing the dual revised simplex method," *Mathematical Programming Computation*, vol. 10, no. 1, pp. 119–142, 2018.

[19] R. C. Shekhar and C. Manzie, "Optimal move blocking strategies for model predictive control," *Automatica*, vol. 61, pp. 27–34, 2015.

[20] IEEE Computational Intelligence Society and Vesta Corporation, "IEEE-CIS fraud detection," Kaggle competition, 2019, https://www.kaggle.com/c/ieee-fraud-detection.

[21] A. H. Murphy, "A new vector partition of the probability score," *Journal of Applied Meteorology*, vol. 12, no. 4, pp. 595–600, 1973.

[22] N. Duan, "Smearing estimate: A nonparametric retransformation method," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 605–610, 1983.