WHEN SHOULD A HUMAN TAKE BACK CONTROL? OPTIMAL DELEGATION UNDER TURBULENT AI RISK

Anonymous authors
Paper under double-blind review

ABSTRACT

Deploying AI systems requires deciding when to delegate tasks and when humans should intervene to monitor and mitigate risk induced by AI operations. These decisions become particularly challenging when failures cluster: a hallucination or harmful output can trigger further errors, creating periods of elevated risk. We introduce a continuous-time framework for learning adaptive human oversight under such turbulent AI risk. Existing oversight and delegation formulations condition on history but do not model incident clustering, or its suppression by supervision effort, jointly with the delegation decision and this study fixes this gap. The self-exciting dynamics capture how risk events increase the likelihood of subsequent events, making their timing and history central to decision-making. We formulate a stochastic control problem that combines human actions, monitoring effort, and switching between human–AI-assisted operation and full AI delegation, balancing operational rewards against oversight costs, and cascading AI-failures and induced uncertainty. Human participation is therefore an endogenous component of risk management: the policy determines both when oversight is needed and how much effort to allocate. We study a relaxed switching formulation and propose Hawkes-PPO, a policy-gradient method that uses a bank of exponential filters of observed incident times as a finite-dimensional summary of the history. In a synthetic environment it attains a higher risk-adjusted objective than either fixed regime and approaches an approximate full-information oracle. We illustrate our results with numerical simulations by examining how cascade risks influence intervention and delegation, connecting reinforcement learning with adaptive human oversight of AI systems. In particular, we illustrate the benefit of our switching strategy and Hawkes-PPO algorithm to monitor the project efficiently along time, reducing turbulent risks occurrences and costs.

1 INTRODUCTION

Language models are increasingly deployed as agents that retrieve information, plan, invoke tools, and act on the results over extended horizons. This shift makes the reliability of extended agent execution an increasingly important safety concern and revives the problem of scalable oversight (Engels et al., 2025): how a limited and costly human supervisor should allocate attention to a system that is cheaper, faster, and possibly more capable, but imperfectly reliable. The difficulty is that agent failures can be dependent across successive steps. A coding agent that misreads a failing test edits the wrong file, and the misleading output of that edit degrades its next diagnosis. For a single evolving task, the operational question is:

> *When should a costly but reliable human controller delegate a continuously evolving project to a cheaper black-box AI controller, and when should the human take control back because accumulated incident risk has become too large?*

Turbulent AI risk. Recent controlled experiments provide evidence that earlier errors can increase the likelihood of subsequent errors: an early incorrect claim can be reused as a premise and induce further incorrect claims (Zhang et al., 2024), and injecting erroneous prior steps into a model's context lowers its accuracy on later steps, an effect that persisted across model sizes and was absent in the tested thinking-enabled variants (Sinha et al., 2026). Execution traces suggest cross-excitation

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.

mechanisms, where a hallucinated state assessment induces an incorrect action that alters the task history (Yang et al., 2024; Cemri et al., 2025; Liu et al., 2026).

An erroneous output thus contaminates the information the agent uses next, so past incidents raise the likelihood of further incidents, of the same type (self-excitation) or of another type (cross-excitation); effective supervision can weaken this dependence; and supervision is costly. We use the term turbulent AI risk for incident dynamics with this incident-driven dependence across time: cascade of self-excited incidents, propagating into a cyber-system while raising uncertainty of the outcome.

Memory-dependent and cascade modeling. The empirical evidences mentioned above advocate for avoiding the use of memoryless processes to model AI risk and incidents, like Poisson processes, rather focusing on self-excited processes. We therefore represent incidents by a self- and mutually exciting Hawkes intensity (Hawkes, 1971) which encompasses the memory-dependent impact and cascade AI risk and we treat this specification as a falsifiable reduced-form hypothesis. That the influence of a past incident decays, and that supervision weakens propagation and cascade effect, are further assumptions. The closest support for this assumption is that external feedback, including human-written feedback, improved code repair in the settings evaluated by Olausson et al. (2024), with gains depending on feedback quality, whereas intrinsic self-correction did not improve reasoning accuracy in Huang et al. (2024).

The model adds two simplifying assumptions that are not defining properties. First, it holds the severity distribution of each incident fixed, so excitation changes how often incidents occur rather than how large they are, although expected losses still rise with intensity. Then, it assumes an additional disturbance to the agent's output by increasing uncertainty and volatility with incident intensity, so that active periods make outcomes less predictable beyond their direct losses. For example, Hao et al. (2024) shows that hallucinated content can propagate through a system and its spread is itself uncertain. We study observable, history-dependent incidents whose propagation can be reduced through verification and corrective supervision, with state and action turbulence (e.g. hallucination) in the broad agentic sense of Section 3.2 as the worked formulation; the framework optimizes the response to incident signals available to the supervisor, not their detection.

Deployed systems show the decision a supervisor faces. In July 2025, a Replit coding agent deleted a production database despite repeated instructions restricting changes, then incorrectly reported that recovery was impossible, after earlier fabricating data and test results.[1] In Anthropic's month-long trial of a model running a small shop, a fabricated interaction was followed by related confabulations, human corrections, and eventual recovery, with neither trigger nor recovery explained (Anthropic, 2025). These accounts motivate the delegation problem but do not establish self-excitation or intervention effectiveness; they show that earlier incidents are potential warning signals that could prompt a reassessment of delegation, the decision our framework formalizes. Permissions and environment separation remain the first line of defense; adaptive supervision addresses the residual decisions they leave open.

Existing oversight protocols can condition on history but do not model incident clustering when deciding how much to delegate (Greenblatt et al., 2024; Bhatt et al., 2025), and the sequential delegation formulations reviewed in Section 2 below do not represent explicit incident-history excitation, or its effort-dependent suppression, jointly with the delegation decision. Modeling both matters because supervision changes future risk by preventing incidents and their downstream effects, so incident history bears on how long to supervise and when to delegate again; this is the mechanism the framework studies. We formulate the problem as continuous-time stochastic control with two regimes: a more reliable but costly human-assisted regime, in which the human acts on the project and chooses a verification effort that weakens propagation, and autonomous execution by a frozen black-box AI policy that is not optimized by the manager and ignores the turbulence it generates. Delegation and takeover are relaxed switching decisions in which the human controls transition rates rather than transition times, with switching costs that induce hysteresis.

Contributions. (i) A supervisory control formulation in which delegation, takeover, project action, and verification effort are chosen jointly, under incident dynamics that the effort itself shapes,

[1] See https://www.saastr.com/replits-new-release-address-most-of-the-challenges-we-hit-vibe-coding-but-is-prosumer-vibe-coding-really-ready-for-commercial-apps-yet/

alongside a frozen autonomous controller with different information and objectives, and under a criterion that penalizes realized quadratic variation together with operating and switching costs. (ii) Hawkes-PPO, an adaptation of PPO in which a bank of exponential filters of the observed incident times provides a finite-dimensional approximation that makes policy-gradient learning implementable (Appendix A); the policy is learned from states, rewards, and incident history, without being supplied the kernel or the model parameters. (iii) A numerical comparison, in a synthetic environment with model-generated incidents, of the objective, project-quality, and volatility tradeoffs achieved by learned switching against fixed regimes and an approximate full-information oracle.

From matching starts, our Hawkes-PPO learned policy improves on both fixed regimes by point estimates, closing significantly the pure-AI-to-oracle gap from an AI start and the pure-human-to-oracle gap from a human start as a balance between those two extreme modes (Appendix B.6, Table 5). The pure-human regime attains the lowest volatility and the highest terminal quality at the highest operating cost, the pure-AI regime the reverse, and the learned policy lies between them. We provide a comparative-statics study shows how the learned switching policy, and the share of the horizon spent under human control, respond to excitation strength, human operating cost and the switching fee.

Section 2 positions the work, Section 3 introduces the model, Section 4 the numerical study, and Section 5 concludes.

2 Related Work

Failure propagation in language-model agents. Beyond the snowballing and self-conditioning experiments cited above (Zhang et al., 2024; Sinha et al., 2026), studies of self-repair report that a model's own feedback yields modest and variable gains in code repair and that stronger external feedback, including human-written feedback, yields larger ones (Olausson et al., 2024), while intrinsic self-correction can fail to improve or even degrade reasoning accuracy (Huang et al., 2024). Agent–computer-interface work documents cascading edit errors that simple guardrails interrupt (Yang et al., 2024). At the system level, failure taxonomies for multi-agent systems (Cemri et al., 2025) and step-level attribution of agentic hallucinations (Liu et al., 2026) classify where incidents originate and how they transmit, while Jamshidi et al. (2026) find that hallucinations attenuate across the multi-agent cascades they evaluate, at the price of factual-information loss. Compounding errors along a trajectory have a theoretical antecedent in imitation learning (Ross et al., 2011). These studies establish or observe propagation, and some evaluate interventions such as feedback or guardrails; none jointly optimizes sustained delegation and costly mitigation effort under explicit incident-history excitation.

Agent safety evaluations and control protocols. Benchmarks show agents behaving unsafely with operational tools under benign instructions (Vijayvargiya et al., 2026) and executing harmful tasks when jailbroken (Andriushchenko et al., 2025); the latter is an adversarial threat model outside our scope. Control protocols combine trusted monitoring with auditing (Greenblatt et al., 2024) or screen actions by resampling (Bhatt et al., 2025), and scaling analyses ask how oversight fares as the overseen system grows more capable (Engels et al., 2025). These protocols can condition on history, but the level of oversight is a protocol parameter rather than a quantity optimized against observed incident dynamics.

Human–AI delegation. Learning to defer decides, per instance, whether a model or a human should act (Madras et al., 2018; Mozannar & Sontag, 2020). Closer to our setting, Fuchs et al. (2024) train a manager that allocates control of a sequential task between human and autonomous agents, and Lykouris & Weng (2024) study deferral when human review decisions interact through congestion and delayed feedback. In these formulations errors and interventions can change the state, but none represents incident-history excitation explicitly or lets supervision effort suppress it; our problem adds both and lets the human choose how much effort to spend while holding control.

Cyber risk management. Risk induced by AI systems is increasingly examined with the concepts of established risk analysis (Thekdi et al., 2023), and its structure is particularly close to that of cyber risk management. The foundational article of Gordon & Loeb (2002) posed the problem of optimal resource allocation under cyber threats, later extended to stochastic control (Callegaro et al., 2026;

Mastrolia & Yan, 2025). At the enterprise level, quantification frameworks make this cost–benefit trade-off operational by coupling attack likelihood with the propagation of impact through functional dependencies to the services and business processes that rely on compromised assets (Tatar et al., 2020). Contagion between cyber incidents is empirically documented (Baldwin et al., 2017), and Hawkes and contagion models have entered the cyber-insurance literature relatively recently (Bessy-Roland et al., 2021; Hernández-Santibáñez et al., 2026; Cherkaoui et al., 2026), with Callegaro et al. (2026) bringing clustered attacks into the control problem. Note that these works do not consider turbulent event modeling in the volatility of the risky project, unlike this study.

Optimal switching and Hawkes-driven control. Impulse control has been applied to cyber-risk management by Hillairet et al. (2026) without incident clustering or a learning method. Relaxed (randomized) controls appear in continuous-time reinforcement learning in Wang et al. (2020) and in principal-agent contracting under moral hazard with project delegation in Kršek & Possamaï (2023), in both cases for continuous controls rather than regime switching. Relaxed switching, in which the controller sets the intensity of regime changes rather than their times, goes back to Bouchard (2009); Elie & Kharroubi (2014) and has been developed for policy-gradient learning by Denkert et al. (2025). What we add is the supervisory setting: two operating regimes with different controls and information, a frozen autonomous controller with its own objective, a project diffusion whose volatility depends on incident intensity. We extend the continuous-time reinforcement learning for controlled Hawkes jump-diffusions, with a control-dependent excitation kernel and Markovianization by exponential filter banks, developed by Bielecki et al. (2026)to switching problem and turbulent AI-risks, see Appendix A, and a quadratic-variation criterion with switching costs. Note that the use of self-exciting activity to generate volatility has also been studied in Horst & Xu (2022) for financial models in a different framework; empirically fitted self-excitation among interacting automated agents in finance (Bacry et al., 2015; Jaisson & Rosenbaum, 2015) shows that the reduced form is estimable, not that it describes language models.

3 AI-Human Risk Model and Optimization

3.1 Project State, Information, and Control Regime

We consider a finite time horizon $[0, T]$, with $T > 0$, and a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, with a filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ satisfying the usual conditions. The filtration $\mathbb{F}$ represents all information available to the project manager, including the current project state, past actions, AI outputs, and previously generated turbulent incidents. The state process X is assumed to take values in $\mathbb{R}$ and denotes the state, or quality, of a continuously evolving project at time $t \in [0, T]$. A larger value of X_t is interpreted as a more desirable project state. At any time, execution is assigned either to a human decision maker or to an AI system. We introduce the regime process $I_t \in \{0, 1\}$, where $I_t = 0$ denotes human AI-assisted control and $I_t = 1$ denotes full AI delegation control. We assume that I is an adapted càdlàg process. When $I_t = 0$, the project is under human control and the human chooses admissible actions $(a_t, e_t) \in \mathcal{A} \times \mathcal{E}$, where $\mathcal{A}$ is the admissible decision space valued in a compact set $A \subset \mathbb{R}$ to work on the project X and $\mathcal{E}$ is the admissible effort valued in the compact $E \subset \mathbb{R}$ space to monitor the assisted AI work.

The AI action is generated according to $\widehat{a}_t$ incorporating standardized information supplied to the AI, prompt template, decoding rule, available tools, retrieval procedure, and context-management mechanism recommended by the AI. The design is explained in Section 4.2. It is not optimized by the human decision maker.

The effective action/effort applied to the project is therefore

$$(a_t, e_t), \text{ if } I_t = 0, \text{ or } (\widehat{a}_t, \varepsilon), \text{ if } I_t = 1,$$

where $1 \gg \varepsilon \geq 0$ represents some automatic safeguards under full AI delegation, possibly 0. The human decision maker may also choose when to delegate or reclaim control. A switching strategy is represented by a sequence $\xi = \big((\tau_n, \iota_n)\big)_{n\geq 1}$, where τ_n are increasing $\mathbb{F}$-stopping times and $\iota_n \in \{0, 1\}$ denotes the regime entered at time τ_n. We restrict attention to admissible switching policies for which the number of switches is almost surely finite on $[0, T]$. Hence the human control problem will eventually consist of two decisions: continuous control under human supervision and switching decisions determining when work is delegated to, or reclaimed from, the AI.

3.2 CASCADE AI-TURBULENCE MODELING

A turbulent incident, *e.g.* hallucination, is defined relative to the information and evidence that are admissible for the project at the time at which the AI produces an output. At the level of generated text, this definition is closely related to the atomic-fact perspective of FActScore (Min et al., 2023), which decomposes a generation into factual claims and evaluates whether they are supported by a reliable information source. Similarly, RAGAS evaluates the faithfulness of generated claims relative to the context provided to the model (Es et al., 2024). In an AI-assisted project, the AI does not merely generate text. It may retrieve project information, infer the current state, reason about dependencies, construct plans, call tools, communicate with humans, and execute actions. Incidents may therefore arise at intermediate stages of an agent trajectory and propagate to subsequent steps. This broader perspective is consistent with recent work on agentic hallucination attribution (Liu et al., 2026).

Definition 3.1 (State and action turbulence)**.** *A state turbulence occurs when the AI forms, reports, retrieves, or propagates an incorrect representation of the project state or its environment. It is represented by a counting process* $N^S = (N_t^S)_{t\geq 0}$.

An action turbulence occurs when the AI proposes, invokes, or reports an action that is inconsistent with its actual capabilities or with the admissible action space. It is represented by a counting process $N^A = (N_t^A)_{t\geq 0}$.

The jumps of N^S and N^A represent the occurrence times of turbulence events rather than necessarily their detection times. We assume that N^S and N^A are simple, adapted counting processes. A central feature of AI-assisted projects is that turbulence are not independent isolated errors. An erroneous output can become part of the context used for subsequent decisions, thereby increasing the likelihood of additional errors. This generates the cascade from state incident to action incident to state contamination to further incidents. We refer to this endogenous propagation mechanism as a cascade of turbulent incidents and modeled by two-dimensional counting process $\mathbf{N} = (N^S, N^A)^\top$. Let $\mathbf{\Phi}$ be a measurable matrix-valued kernel with nonnegative entries. We assume that each ϕ_{ij} is locally integrable. We use the convention that the first index denotes the type of the future turbulence and the second index denotes the type of the triggering turbulence. Thus, ϕ_{ij} measures the effect of a type-j turbulence on the future occurrence rate of type-i turbulence, for example ϕ_{SS} and ϕ_{AA} model self-excitation for states and actions, whereas ϕ_{AS} and ϕ_{SA} model cross-excitation: state turbulence to future action turbulence and action turbulence to future state turbulence, respectively. We denote by $\boldsymbol{\lambda}_t = (\lambda_t^S, \lambda_t^A)^\top$ the intensity of $\mathbf{N}$ defined by

$$\boldsymbol{\lambda}_t = \boldsymbol{\mu}_t(\bar{e}_t) + \int_{(0,t)} \mathbf{\Phi}(t-s, \bar{e}_t)\, d\mathbf{N}_s,$$

where $\boldsymbol{\mu}_t = (\mu_t^S, \mu_t^A)^\top$ is the nonnegative $\mathbb{F}$-predictable baseline intensity process of $\boldsymbol{\lambda}_t$ and $\bar{e}_t := \bar{e}(i,\pi) = (1 - I_{t-})e_t + I_{t-}\varepsilon$.

Remark 3.2. *The effort of the human is to reduce the turbulences by monitoring the AI assisted work. Therefore,* μ^S, μ^A, ϕ *are nonincreasing with respect to the effort* e.

Example 3.3 (Exponential kernels)**.** *Let* $\beta_S, \beta_A > 0$ *be state and action incidents decays rate. Consider the kernel*

$$\mathbf{\Phi}(r,e) = \begin{pmatrix} Q_{SS}(e)e^{-\beta_S r} & Q_{SA}(e)e^{-\beta_A r} \\ Q_{AS}(e)e^{-\beta_S r} & Q_{AA}(e)e^{-\beta_A r} \end{pmatrix}, \quad r \geq 0, e \in E,$$

where $Q_{ij} : \mathcal{E} \to \mathbb{R}_+$. *Define* $Z_t^S := \int_{(0,t)} e^{-\beta_S(t-s)}\, dN_s^S$, *and* $Z_t^A := \int_{(0,t)} e^{-\beta_A(t-s)}\, dN_s^A$. *Then,*

$$\lambda_t^S = \mu_t^S(\bar{e}_t) + Q_{SS}(e_t)Z_{t-}^S + Q_{SA}(\bar{e}_t)Z_{t-}^A, \quad \lambda_t^A = \mu_t^A(\bar{e}_t) + Q_{AS}(\bar{e}_t)Z_{t-}^S + Q_{AA}(\bar{e}_t)Z_{t-}^A.$$

This example is purely informative at this stage, in our model we assume that the kernel $\mathbf{\Phi}$ *is unknown to use Hawkes-PPO algorithm described below, compared with a fully non-Markovian power law kernel and Oracle model that is approached with a mixture of exponential kernels following the Markovianization method proposed in Khabou & Talbi (2025); Bielecki et al. (2026)*

3.3 RELAXED FORMULATION AND PROJECT DYNAMICS

The classical switching formulation involves a controller directly choosing a sequence of switching times and paying the lump-sum cost χ_{ij} when the operating regime changes from i to j. We instead

introduce a relaxed formulation in which the controller does not directly choose the switching times, but controls their instantaneous arrival rates (Bouchard, 2009; Denkert et al., 2025). Let $J = (J_t)_{t\in[0,T]}$ be a simple counting process whose predictable intensity ν is controlled by the decision maker, so that $\widetilde{J}_t := J_t - \int_0^t \nu_s\, ds$ is an $\mathbb{F}$-local martingale. Since $I_t \in \{0,1\}$, every jump of J reverses the current regime $dI_t = (1-2I_{t-})\, dJ_t$, and $\nu_t = (1-I_{t-})\nu_t^{01} + I_{t-}\nu_t^{10}$, where ν^{01} (resp. ν^{10}) is the intensity of switching from regime 0 to 1 (resp. 1 to 0). The set of admissible relaxed switching controls is

$$\mathcal{V} := \big\{\nu = (\nu^{01}, \nu^{10}) : \nu^{01}, \nu^{10} \text{ are } \mathbb{F}\text{-predictable and } 0 \le \nu_t^{ij} \le \overline{\nu} < +\infty\big\}.$$

Definition 3.4 (Relaxed control)**.** *An admissible relaxed control is a triple $\pi = (a, e, \nu) \in \mathcal{A}\times\mathcal{E}\times\mathcal{V}$ such that $\mathbb{E}\big[\int_0^T \nu_t^2\, dt\big] < +\infty$. We denote this space by $\mathfrak{U}$.*

Let W° and W^{AI} be independent Brownian motions. The project dynamics are

$$\begin{aligned}\mathrm{d}X_t = &\Big[(1-I_t)b(t,X_t,a_t) + I_t b(t,X_t,\widehat{a}_t)\Big]\,\mathrm{d}t + \Big[(1-I_t)\sigma(t,X_t,a_t) + I_t\sigma(t,X_t,\widehat{a}_t)\Big]\,\mathrm{d}W_t^\circ\\ &+ \Big[(1-I_t)\gamma(t,X_t,e_t,\boldsymbol{\lambda}_t) + I_t\gamma(t,X_t,\varepsilon,\boldsymbol{\lambda}_t)\Big]\,\mathrm{d}W_t^{\mathrm{AI}} - M_t^S dN_t^S,\end{aligned}$$

where b, σ, γ satisfy standard Lipschitz and growth conditions and M^S is a random variable independent of $W^\circ, W^{\mathrm{AI}}, N^S, N^A$ representing the size of the incident reducing the project value. The turbulence-induced volatility γ is nondecreasing in $\boldsymbol{\lambda}$ (higher turbulence activity increases project uncertainty) and nonincreasing in e (greater human monitoring reduces uncertainty). In the human-assisted regime the AI still produces outputs that enter the project, so incidents still occur and carry losses; effort lowers their frequency, not their severity.

3.4 Objective

The performance criterion associated with $\pi \in \mathfrak{U}$ starting in state $i \in \{0,1\}$ is

$$V(i) = \sup_{\pi\in\mathfrak{U}} \mathbb{E}^\pi\Big[g(X_T) + \int_0^T \Big(f(s,X_s) - C_s - \frac{\kappa(I_{s-})}{2}\nu_s^2\Big)ds - \frac{\eta}{2}[X]_T - \int_0^T \chi_{I_{s-},1-I_{s-}}\, dJ_s\Big], \quad (1)$$

where g is the terminal reward, f the running reward, C_s the regime-dependent operating cost, $\eta > 0$ a risk-sensitivity parameter, and $\chi_{i,1-i} > 0$ the cost of switching from regime i to regime $1-i$. $[X]_T$ denotes the quadratic variation of the process X. Unlike the variance which model the deviation from the mean, this incorporates the risk related to high variations on X, which is more relevant risk aversion cost in the context of turbulent risk. In practice the controller faces a trade-off between the relatively high cost of human effort ($c_0 > c_1$ typically) and the lower operating cost of AI delegation. AI delegation introduces turbulence risk penalized through η, while positive switching costs rule out costless high-frequency switching and generate a hysteresis effect. The solution to (1) relies on a Markovian procedure initially introduced in Khabou & Talbi (2025); Bielecki et al. (2026) extended to our turbulent volatility model. The main idea is to approach the unknown kernel Φ with a mixture of exponential kernels as in Example 3.3 so that the value function associated with this approximation converges to the primal value V. This reduction enables us to reduce the problem to an integro-partial HJB equation. The details are provided in Appendix A.

4 Numerical illustration: learning when to delegate to AI

4.1 Model and objective specification

We consider time-homogeneous project dynamics with linear mean reversion, action-dependent baseline volatility, and a saturating AI-induced volatility coefficient:

$$b(t,x,a) = a - \delta x, \quad \sigma(t,x,a) = \sigma_0 + \sigma_1 a, \quad \gamma(t,x,e,\ell) = \gamma_0 + \gamma_1 \frac{\ell^S + \ell^A}{1 + \ell^S + \ell^A},$$

where $\ell = (\ell^S, \ell^A)$ denotes the turbulence intensities. The bivariate Hawkes process has constant baseline intensities and power-law excitation kernels, memory-dependent (non-Markovian),

approached with a mixture of exponential similar to Bielecki et al. (2026), see Appendix A and represented in the simulator by a 20-term exponential mixture fitted on the filter grid. Its amplitudes decrease linearly with monitoring effort $\mu_t^i(e) = \mu_i$, $\Phi_{ij}(r,e) = (1-e)A_{ij}(1+\rho_{ij}r)^{-\beta_{ij}}$, $i,j \in \{S,A\}$. We use quadratic operating costs and symmetric switching costs, $c_0(a,e) = k_0 + k_a a^2 + k_e e^2$, $c_1(a,e) = k_{\text{AI}} + a^2$, and $\chi_{01} = \chi_{10} = \chi$. In this specification effort acts only through the excitation amplitude, which vanishes at $e = 1$; μ, γ are effort-independent. Parameter values and the specifications of f and g are collected in Appendix B.1, Table 2.

4.2 Black-box AI modeling

We propose a design of action $\hat{a}$ for the AI recommendation. When optimizing, the AI is not considering the turbulence its action is generating, and only focus on the pure project value only. To model the black-box by the autonomous AI mode, we consider the project-only oriented dynamics

$$dY_s = (b_s - \delta Y_s)ds + \sigma(b_s)dW_s^\circ,$$

where b is the action to be optimize by the AI. The AI to thus selects the black-box strategy as follow:

$$\sup_{b\in[0,1]} \mathbb{E}\Big[g(Y_T) + \int_0^T \{f(Y_s) - k_{AI} - b_s^2\}ds - \frac{\eta}{2}[Y]_T\Big].$$

The solution of this problem is details in Appendix B.2. The optimizer a^* is given as a function of time t and evaluated at the realized state X_t. This pretraining excludes turbulent risk, monitoring and switching. Fixing a time discretization $(t_k)_{k\geq 0}$ the AI autonomous action is

$$\widehat{a}_t^{\text{raw}} = a^*(t, X_t) + R_t, dR_t = -\theta_0 R_t\, dt + \theta_1 s(L_t)\, dW_t^R - M^A dN_t^A,$$

with $L_t = \lambda_t^A + \lambda_t^S$ and $R_0, M^A, s(\cdot)$ specified in Table 2. The error R_t evolves in both regimes. Action-corruption events push R downward and lower the recommendation; the recovery drift subsequently returns it towards zero, while Hawkes risk raises its Brownian uncertainty.

4.3 Reinforcement learning, Hawkes-PPO and value comparison

The Hawkes-PPO algorithm used is presented in Appendix B.3, Algorithm 1 adapted from Bielecki et al. (2026). The Hawkes filters are computed solely from observed risk-event times; the true intensities and excitation coefficients are not supplied to the policy. Standard PPO losses, network architectures, training settings, and deterministic deployment are specified in the Appendix B.3. We first compare the values of Hawkes-PPO with a non-switching policy and a know-everything oracle (see Appendix B.4), a pure AI mode (Appendix B.2) and a pure human mode (Appendix B.5). We use, 4,096 fresh common-noise Monte Carlo simulations. Figure 1 reports the value (1), the mean pathwise volatility $\mathbb{E}[\sqrt{[X]_T/T}]$, and the mean terminal project state $\mathbb{E}[X_T]$ with 90% Monte Carlo confidence interval. AI-start and Human-start switching results are kept separate. We observe that the Oracle achieves obviously the best objective, followed very closely to the Hawkes-PPO, far from a pure human or AI regime, as a sanity check of our reinforcement learning method together with optimality. Hawkes-PPO reduces significantly the volatility of the project staying close to the Oracle, far from a pure AI-regime and still very close to a pure human mode. Our algorithm also over performed a pure AI mode for the terminal value of the project, without considering cost of effort, action for the pure human mode. The switching gain is a tradeoff among risk, implementation cost, and project benefits. Among the eight RL variants of Appendix B.6, only Hawkes-PPO improves on the fixed-Human benchmark from a Human start, with the best value compared with an oracle, and the filter bank helps PPO but not SAC or DDPG; the switching gain is therefore algorithm-dependent under our budget.

4.4 Switching between autonomous AI and AI-assisted Human

Figure 2 illustrates the results obtained from the Hawkes-PPO path with four switches at $1.46, 3.88, 4.94, 7.25$. During AI spells, state-risk events raise Hawkes intensity and quadratic variation, lower the pathwise score, and precede increased switching intensity towards Human mode. During the two event-free Human spells, mitigation falls and the intensity of returning to AI rises as event history decays. The pattern is consistent with the learned balance between persistent risk and Human cost. A stochastic intensity does not force a switch at its maximum or at a fixed threshold.

4.5 How incident dynamics and costs shape supervision

We next examine how the supervision policy responds to incident dynamics and intervention costs. Using Hawkes-PPO with the same public incident-history filters, we vary excitation strength A_{ij} (0.5 or 1.5 $\times$ baseline), Human fixed operating cost k_0 (0.75 or 1.25 $\times$ baseline), and the per-switch fee χ (0.5 or 2 $\times$ baseline). Appendix B.7 specifies the design, protocol, and robustness checks. The results of this sensitivity analysis are shown in Table 1 for $T = 8$. We first observe a strong increase in human involvement when the excitation amplitude rises. Low excitation (half the baseline) leading to less human involvement while high excitation ($3/2$ of a baseline) involved significantly more the Human supervision with less effort, fewer switches, fewer bursts and higher terminal quality, at a lower objective. Increasing the Human fixed-cost multiplier from 0.75 to 1.25 reduces Human time (76.0% to 13.7%) for AI starts and from 99.6% to 19.1% for Human starts. Effort during Human operation rises, while incident counts and burst probability increase, and terminal quality decreases at baseline excitation. The crossed checks preserve the Human-time and incident directions, but effort and quality responses have exceptions: at high excitation, higher Human cost increases terminal quality despite more incidents. Comparing per-switch fee multipliers 0.5 and 2, switching decreases in both initial regimes, while Human time falls for AI starts and rises for Human starts. Appendix B.7 gives supporting figures and numerical analysis.

5 Conclusion

We introduced a framework for turbulent AI risk in which AI-generated incidents can trigger further incidents and amplify uncertainty in a project's evolution. Within this framework, we formulated an optimal switching problem between full delegation to a black-box AI system and human–AI assisted decision-making. The analysis captures the trade-off between the lower operating costs of AI delegation and the cost of human intervention to mitigate cascading risks. It highlights the importance of adapting human involvement to the evolving risk state, accounting for both immediate operating costs and the consequences of incident accumulation. By modeling cascading incidents and adaptive human intervention, this work contributes to the broader study of AI risk management Ziosi et al. (2026), with a specific focus on the dynamic accumulation and amplification of AI-generated risks.

Our framework provides a stylized description of these interactions, and its application to deployed systems would require empirical calibration and validation of the incident dynamics. A natural extension is to consider interconnected AI agents cooperating under human oversight. Whether incidents attenuate or amplify across agents is empirically open. (Jamshidi et al., 2026) find attenuation but incidents could propagate across agents and create feedback loops, turning local failures into systemic risks. This extension would raise a further control problem: how to allocate human oversight across a network to contain cascading incidents while retaining the benefits of AI delegation.

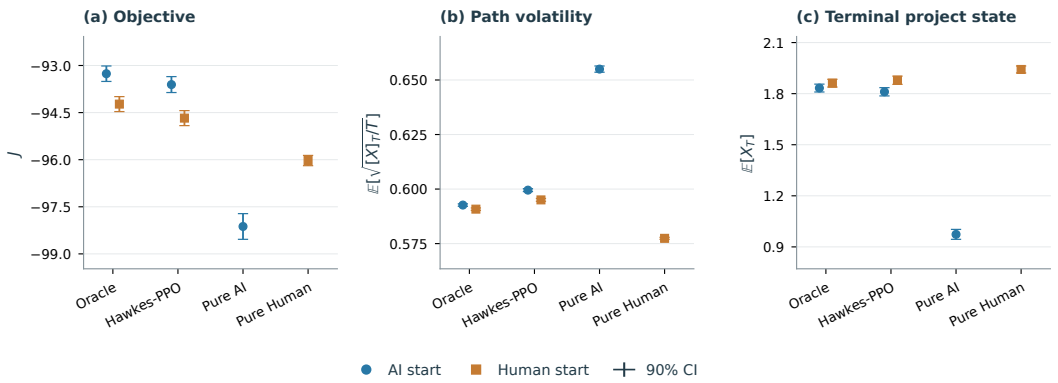


Figure 1: Four-method comparison. Error bars are 90% confidence intervals. Higher objective and terminal state are better; lower realized volatility indicates less quadratic variation.

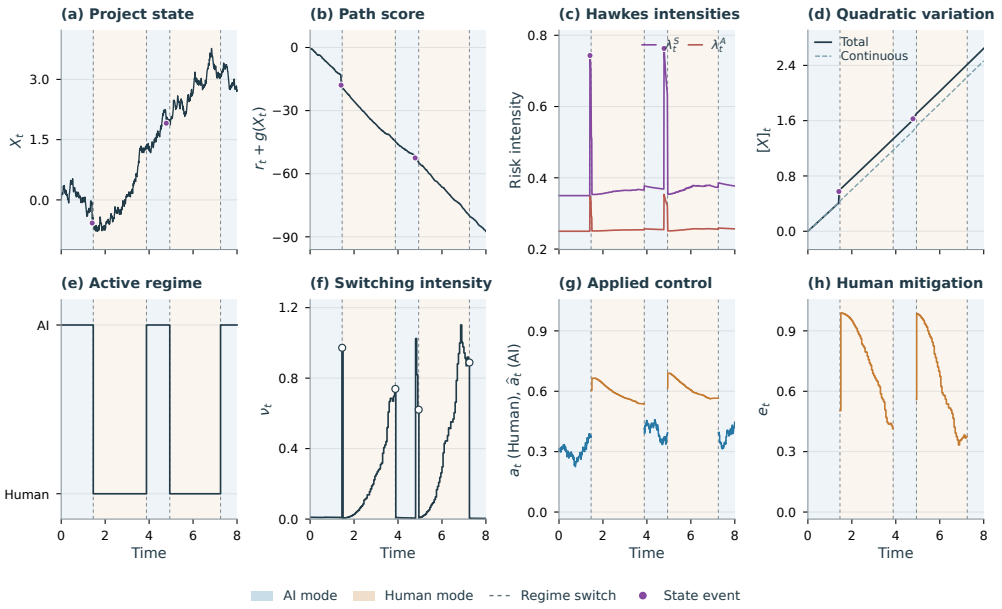


Figure 2: Vertical dashed lines mark regime switches. From left to right, top to bottom: paths of X, net running reward and terminal reward, intensity of state and action turbulence (Hawkes intensities), quadratic variation of the project, regime activation, optimal switching intensity, action of the agent human or AI, human effort.

Table 1: Supervision and project outcomes across model configurations.

Configuration	Human time (%)	Effort (Human)	Switches	Incidents	Burst (%)	Terminal quality	Objective
Panel A. AI initial regime							
Baseline	**46.4**	**0.896**	**1.46**	**5.40**	**45.9**	**1.81**	**-93.61**
Excitation 0.5	0.9	0.893	0.27	5.59	45.9	1.17	-85.96
Excitation 1.5	73.6	0.842	1.22	5.13	38.6	1.88	-95.85
Human cost 0.75	76.0	0.731	1.15	5.05	36.7	1.82	-89.68
Human cost 1.25	13.7	0.973	1.01	6.20	57.4	1.30	-96.33
Switch fee $\times 0.5$	45.4	0.880	1.34	5.44	47.0	1.90	-93.72
Switch fee $\times 2$	43.0	0.853	0.91	5.53	48.6	2.11	-94.36
A 0.5; cost 0.75	2.4	0.792	0.23	5.58	45.8	1.16	-85.84
A 0.5; cost 1.25	0.5	0.753	0.23	5.61	46.0	1.15	-86.00
A 1.5; cost 0.75	86.5	0.783	1.24	4.99	35.5	1.77	-90.86
A 1.5; cost 1.25	64.0	0.914	1.68	5.26	42.6	1.91	-100.14
Panel B. Human initial regime							
Baseline	**58.4**	**0.809**	**2.01**	**5.28**	**43.4**	**1.88**	**-94.68**
Excitation 0.5	5.3	0.516	1.42	5.58	45.7	1.19	-87.33
Excitation 1.5	98.1	0.696	0.22	4.92	33.8	2.02	-96.85
Human cost 0.75	99.6	0.623	0.07	4.93	33.5	1.93	-89.75
Human cost 1.25	19.1	0.862	2.09	6.13	56.5	1.33	-97.73
Switch fee $\times 0.5$	58.1	0.793	1.92	5.31	44.0	1.96	-94.65
Switch fee $\times 2$	82.5	0.693	0.66	5.09	38.3	2.51	-96.17
A 0.5; cost 0.75	13.6	0.607	1.28	5.51	44.6	1.19	-86.75
A 0.5; cost 1.25	4.2	0.417	1.41	5.60	45.8	1.15	-87.65
A 1.5; cost 0.75	99.6	0.708	0.13	4.90	33.5	1.83	-90.45
A 1.5; cost 1.25	74.4	0.826	2.01	5.16	40.2	1.98	-101.52

4,096 evaluation paths per initial regime. Entries are path means, except effort, which is total Human effort divided by total Human time. Burst denotes at least three combined state and action incidents in a window of one time unit. Excitation, Human cost, and per-switch fee are baseline multipliers.

AI use statement. In this work, we used generative AI tools for the aesthetic of the paper, correction of the english language and grammar, assist in the writing of proofs, advice on relevant existing bibliography and its structure, provide feedback on research methodology, implement methods. We have not used generative AI tools for ideas and creativity related to the model interpretation of the results, proofs ideas and assumptions. The experiments were independently conceived and designed by the authors. AI tools were used to assist with refining the experimental setup, implementing the experiments, and conducting numerical simulations. We have reviewed all AI-assisted work. We checked LLM-generated phrasing, proposed references, simulations details. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement. This research adheres to the ICLR Code of Ethics. All experiments were conducted using generative and synthetic data. This study does not involve human subjects or personally identifiable information. No new sensitive data collection was performed. The contributions of this work are strictly methodological and aimed at advancing machine learning foundations for AI risk. We perceive no direct, high-risk societal harms, conflicts of interest, or negative ethical implications resulting from this foundational research.

Reproducibility statement. The two appendix ensures the reproducibility of our model. Appendix A proves convergence of the kernel-approximated full-information value (Theorem A.6) and bounds the loss from current-observation feedback (Theorem A.11); the residual due to learning is left uncertified (A.5). Appendix B specifies parameters, protocols, seeds and algorithms. Code will be released upon acceptance.

REFERENCES

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.09024. arXiv:2410.09024.

Anthropic. Project vend: Can Claude run a small shop? (And why does that matter for the future of AI?). https://www.anthropic.com/research/project-vend-1, June 2025.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(1):1550005, 2015. arXiv:1502.04592.

Adrian Baldwin, Iffat Gheyas, Christos Ioannidis, David Pym, and Julian Williams. Contagion in cyber security attacks. *Journal of the Operational Research Society*, 68(7):780–791, 2017. doi: 10.1057/jors.2016.37.

Yannick Bessy-Roland, Alexandre Boumezoued, and Caroline Hillairet. Multivariate Hawkes process for cyber insurance. *Annals of Actuarial Science*, 15(1):14–39, 2021.

Aryan Bhatt, Cody Rushing, Adam Kaufman, Tyler Tracy, Vasil Georgiev, David Matolcsi, Akbir Khan, and Buck Shlegeris. Ctrl-Z: Controlling AI agents via resampling. *arXiv preprint arXiv:2504.10374*, 2025. URL https://arxiv.org/abs/2504.10374.

Tomasz R. Bielecki, Thibaut Mastrolia, and Haoze Yan. Continuous-time reinforcement learning for controlled hawkes jump-diffusions, 2026. URL https://arxiv.org/abs/2608.19151.

Bruno Bouchard. A stochastic target formulation for optimal switching problems in finite horizon. *Stochastics: An International Journal of Probability and Stochastics Processes*, 81(2):171–197, 2009.

Pierre Brémaud and Laurent Massoulié. Stability of nonlinear hawkes processes. *The Annals of Probability*, 24(3):1563–1588, 1996. ISSN 00911798, 2168894X. URL http://www.jstor.org/stable/2244985.

Giorgia Callegaro, Claudio Fontana, Caroline Hillairet, and Beatrice Ongarato. A stochastic gordon–loeb model for optimal cybersecurity investment under clustered attacks. *Annals of Actuarial Science*, pp. 1–29, 2026.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.13657. arXiv:2503.13657.

Ziheng Cheng, Xin Guo, and Yufei Zhang. Deterministic policy gradient for reinforcement learning with continuous time and state, 2026 (v2). URL https://arxiv.org/abs/2509.23711.

Yousra Cherkaoui, Caroline Hillairet, Thomas Peyrat, and Anthony Réveillac. Stress scenarios of cyber loss processes with dependencies. Working paper hal-05558990, HAL, 2026. URL https://hal.science/hal-05558990v1.

Robert Denkert, Huyên Pham, and Xavier Warin. Control randomisation approach for policy gradient and application to reinforcement learning in optimal switching. *Applied Mathematics & Optimization*, 91(1):9, 2025.

Romuald Elie and Idris Kharroubi. BSDE representations for optimal switching problems with controlled volatility. *Stochastics and Dynamics*, 14(03):1450003, 2014.

Joshua Engels, David Baek, Subhash Kantamneni, and Max Tegmark. Scaling laws for scalable oversight. *Advances in Neural Information Processing Systems*, 38:97327–97366, 2025.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pp. 150–158, 2024.

Andrew Fuchs, Andrea Passarella, and Marco Conti. Optimizing delegation in collaborative human-ai hybrid teams. *ACM Transactions on Autonomous and Adaptive Systems*, 19(4):1–33, 2024.

Lawrence A Gordon and Martin P Loeb. The economics of information security investment. *ACM Transactions on Information and System Security (TISSEC)*, 5(4):438–457, 2002.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *International Conference on Machine Learning*, number 650, pp. 16295 – 16336, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Guozhi Hao, Jun Wu, Qianqian Pan, and Rosario Morello. Quantifying the uncertainty of LLM hallucination spreading in complex adaptive social networks. *Scientific reports*, 14(1):16375, 2024.

A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58 (1):83–90, 1971. doi: 10.1093/biomet/58.1.83.

Nicolás Hernández-Santibáñez, Nabil Kazi-Tani, and Mariano Vazquez-Gaete. Cyber risk prevention under risk averse spectral criteria. *Applied Mathematics & Optimization*, 94(1):14, 2026.

Caroline Hillairet, Thibaut Mastrolia, and Wissal Sabbagh. Optimal impulse control for cyber risk management. *Applied Mathematics & Optimization*, 94(1):18, 2026.

Ulrich Horst and Wei Xu. The microstructure of stochastic volatility models with self-exciting jump dynamics. *The Annals of Applied Probability*, 32(6):4568–4610, 2022. doi: 10.1214/22-AAP1796.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Thibault Jaisson and Mathieu Rosenbaum. Limit theorems for nearly unstable Hawkes processes. *The Annals of Applied Probability*, 25(2), 2015.

Saeid Jamshidi, Arghavan Moradi Dakhel, Kawser Wazed Nafi, and Foutse Khomh. Hallucination cascade: Analyzing error propagation in multi-agent llm systems. *arXiv preprint arXiv:2606.07937*, 2026.

David W Kammler. Approximation with sums of exponentials in lp[0, ∞). *Journal of Approximation Theory*, 16(4):384–408, 1976. ISSN 0021-9045. doi: https://doi.org/10.1016/0021-9045(76)90071-X. URL https://www.sciencedirect.com/science/article/pii/0021904576900 71X.

Mahmoud Khabou and Mehdi Talbi. Markov approximation for controlled Hawkes jump-diffusions with general kernels. *arXiv preprint arXiv:2507.11294*, 2025.

Daniel Kršek and Dylan Possamaï. Randomisation with moral hazard: a path to existence of optimal contracts. *arXiv preprint arXiv:2311.13278*, 2023.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Xuannan Liu, Xiao Yang, Zekun Li, Peipei Li, and Ran He. Agenthallu: Benchmarking automated hallucination attribution of LLM-based agents, 2026. URL https://arxiv.org/abs/2601.06818.

Thodoris Lykouris and Wentao Weng. Learning to defer in congested systems: the AI-human interplay. *arXiv preprint arXiv:2402.12237*, 2024.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, 2018.

Thibaut Mastrolia and Haoze Yan. Agency problems and adversarial bilevel optimization under uncertainty and cyber threats. *arXiv preprint arXiv:2505.08989*, 2025.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/2006.01862.

Theo X. Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. Is self-repair a silver bullet for code generation? In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.09896. arXiv:2306.09896.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011. 15:627-635.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Akshit Sinha, Arvindh Arun, Shashwat Goel, Steffen Staab, and Jonas Geiping. The illusion of diminishing returns: Measuring long horizon execution in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3b4e1336f775c3dba16ebbb8d2afd258-Paper-Conference.pdf. arXiv:2509.09677.

J. Sirignano and K. Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018. doi: 10.1016/j.jcp.2018.08.029.

Unal Tatar, Omer Keskin, Hayretdin Bahsi, and C. Ariel Pinto. Quantification of cyber risk for actuaries: An economic-functional approach. Research report, Society of Actuaries, May 2020. URL https://www.soa.org/resources/research-reports/2020/quantification-cyber-risk/.

Shital Thekdi, Unal Tatar, Joost Santos, and Samrat Chatterjee. Disaster risk and artificial intelligence: A framework to characterize conceptual synergies and future opportunities. *Risk Analysis*, 43(8): 1641–1656, 2023. doi: 10.1111/risa.14038.

Sanidhya Vijayvargiya, Aditya Bharat Soni, Xuhui Zhou, Zora Zhiruo Wang, Nouha Dziri, Graham Neubig, and Maarten Sap. OpenAgentSafety: A comprehensive framework for evaluating real-world AI agent safety. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.06134. arXiv:2507.06134.

Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34, 2020.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.15793. arXiv:2405.15793.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59670–59684, 2024.

Marta Ziosi, Miro Plueckebaum, Stephen Casper, Henry Papadatos, Ze Shen Chin, Peter Slattery, James Gealy, Tim GJ Rudner, Brian Tse, Ariel Gil, et al. Open problems in frontier ai risk management. *arXiv preprint arXiv:2604.25982*, 2026.

A MARKOVIAN APPROXIMATION AND OBSERVABLE-POLICY STABILITY

We use the approximation framework of Khabou & Talbi (2025); Bielecki et al. (2026), with mixture of exponential approximation background from Kammler (1976). We extend their framework to intensity-dependent diffusion, especially with the dependency λ in the volatility process γ. The main difficulty is in the observation of $\hat{a}$. We will distinguish a full observation mode in which the error process R is fully observed and only the recommendation G is observed.

We first state the main assumptions on the coefficient in our model, see Section A.1. Then, we introduce the definition of memory filters and the associated convergences of the approached intensity, derived from the mixture of exponential to the primal one, see Section A.2. We then turn to the convergence of the approached problem to the value function of the primal optimization (1) when the error R induced by the AI is observed, see Section A.3. Finally, in Section A.4, when the recommendation G is observed in AI mode but not the error R, we derive information-loss bounds between full-information and observable-history control for a fixed recommendation map G, under additional verification, sensitivity, and policy-existence assumptions.

A.1 MODEL AND ASSUMPTIONS

We use the same framework as in Bielecki et al. (2026) with the similar Poisson embedding construction in Brémaud & Massoulié (1996). Write $\pi = (a, e, \nu) \in \mathfrak{U}$. We recall

$$\mathbf{S}_t = (X_t, R_t), \qquad \widehat{a}_t = \mathrm{Proj}_A\, G(t, \mathbf{S}_{t-}),$$
$$\alpha_t := \alpha(t, \mathbf{s}, i, \pi) = (1 - I_{t-})a_t + I_{t-}\widehat{a}_t, \qquad \bar{e}_t := \bar{e}(i, \pi) = (1 - I_{t-})e_t + I_{t-}\varepsilon.$$

Use Euclidean vector norms and the Frobenius norm for diffusion matrices. For intensity and count vectors, write $\|v\|_1 := \sum_i |v_i|$. For kernel matrices $M \in \mathbb{R}^{2\times 2}$, use the induced matrix norm

$$\|M\| := \|M\|_{1\to 1} := \sup_{v\in\mathbb{R}^2\setminus\{0\}} \frac{\|Mv\|_1}{\|v\|_1} = \max_{j=1,2} \sum_{i=1}^{2} |M_{ij}|.$$

For each integer $K \geq 1$, we define $\Phi^K(r, e) := \sum_{\ell=1}^{K} Q_{K,\ell}(e)e^{-\beta\ell r}$, which represents a K-term exponential approximation of the true kernel Φ and $Q_{K,\ell}(\cdot) \in E \to \mathbb{R}^{2\times 2}$ denotes the effort effect on the kernel. We set the following assumption for a fixed kernel Φ.

Let $U := A \times E \times [0, \overline{\nu}]$, where A and E are compact, and assume $\varepsilon \in E$. The original and approximating systems have the same initial state and regime, with empty incident history.

Assumption A.1 (Kernel approximation)**.** *For a fixed $\beta > 0$, the selected approximants satisfy*

$$\epsilon_K := \int_0^T d_K(r)\, dr \underset{K\to\infty}{\longrightarrow} 0, \quad d_K(r) := \sup_{e\in E} \|\Phi^K(r, e) - \Phi(r, e)\|.$$

We refer to Remark 3.2 in Bielecki et al. (2026) for the sufficient condition satisfying the above assumption. In equations shared by both models, use $q \in \{0, K\}$, with $\Phi^0 := \Phi$ and the convention

$$(\mathbf{S}^0, N^0, \lambda^0, J^0, I^0, \pi^0) = (\mathbf{S}, N, \lambda, J, I, \pi).$$

The same convention applies to components and derived quantities. Use a common Brownian motion $B = (W^0, W^{AI}, W^R)$, independent marked Poisson random measures Π^j of intensity $dt\, d\theta\, \rho_j(dm)$, $j \in \{S, A\}$, and an independent switching measure Π^{sw} of intensity $dt\, d\theta$. Here each ρ_j is a probability law. Write $\Pi = (\Pi^S, \Pi^A, \Pi^{\mathrm{sw}})$ for this common Poisson embedding; it is shared by the

original and approximating systems. We define

$$\boldsymbol{\lambda}_t^q = \boldsymbol{\mu}(t, \bar{e}_t^q) + \int_{(0,t)} \boldsymbol{\Phi}^q(t-s, \bar{e}_t^q)\, d\boldsymbol{N}_s^q, \tag{2}$$

$$M^{q,j}(dt, dm) = \int_{\mathbb{R}_+} \mathbf{1}_{\{\theta \le \lambda_t^{q,j}\}} \Pi^j(dt, d\theta, dm), \qquad N_t^{q,j} = M^{q,j}((0,t] \times \mathcal{M}_j),$$

$$dJ_t^q = \int_{\mathbb{R}_+} \mathbf{1}_{\{\theta \le \nu_t^q\}} \Pi^{\mathrm{sw}}(dt, d\theta), \qquad dI_t^q = (1 - 2I_{t-}^q)\, dJ_t^q,$$

$$d\mathbf{S}_t^q = b_{\mathbf{S}}(t, \mathbf{S}_{t-}^q, I_{t-}^q, \pi_t^q, \boldsymbol{\lambda}_t^q)dt + \Sigma_{\mathbf{S}}(t, \mathbf{S}_{t-}^q, I_{t-}^q, \pi_t^q, \boldsymbol{\lambda}_t^q)dB_t + \sum_j \int_{\mathcal{M}_j} m M^{q,j}(dt, dm) \tag{3}$$

where

$$b_{\mathbf{S}}(t, \mathbf{s}, i, \pi, \ell) := \begin{pmatrix} (1-i)b(t, x, a) + ib(t, x, \widehat{a}_t) \\ -\theta_0 R_t \end{pmatrix}$$

$$\Sigma_{\mathbf{S}}(t, \mathbf{s}, i, \pi, \ell) := \begin{pmatrix} (1-i)\sigma(t, x, a) + i\sigma(t, x, \widehat{a}) & (1-i)\gamma(t, x, e, \ell) + i\gamma(t, x, \varepsilon, \ell) & 0 \\ 0 & 0 & \theta_1 s(L_t) \end{pmatrix}.$$

Write $\mathbf{s} = (x, r)$ for a generic joint state.

Assumption A.2 (Kernel regularity)**.** *The baseline $\mu(t, e)$ is deterministic, nonnegative, bounded, and continuous. The kernels are measurable and entrywise nonnegative on $[0, T] \times E$. There exist finite constants $H, L > 0$, independent of K and the controller, such that, for all $K \ge 1$, $q \in \{0, K\}$, $t, r \in [0, T]$, and $e, e' \in E$,*

$$\sup_{r,e} \|\Phi^q(r, e)\| \le H,$$
$$\|\mu(t, e) - \mu(t, e')\|_1 + \|\Phi^q(r, e) - \Phi^q(r, e')\| \le L|e - e'|.$$

Assumption A.3 (Frozen controller and dynamics)**.** *All coefficient maps are measurable. The map G is uniformly Lipschitz in $\mathbf{s}$, with $\sup_t |G(t, 0)| < \infty$. For each regime i, uniformly in t,*

$$\begin{aligned} &|b_{\mathbf{S}}(t, \mathbf{s}, i, \pi, \ell) - b_{\mathbf{S}}(t, \mathbf{s}', i, \pi', \ell')| + \|\Sigma_{\mathbf{S}}(t, \mathbf{s}, i, \pi, \ell) - \Sigma_{\mathbf{S}}(t, \mathbf{s}', i, \pi', \ell')\|_F \\ &\qquad \le L\big(|\mathbf{s} - \mathbf{s}'| + |\pi - \pi'| + \|\ell - \ell'\|_1\big), \\ &|b_{\mathbf{S}}(t, 0, i, \pi, 0)| + \|\Sigma_{\mathbf{S}}(t, 0, i, \pi, 0)\|_F \le L. \end{aligned} \tag{4}$$

The jump amplitudes are state- and control-independent, with

$$\sum_j \int |m|^4 \rho_j(dm) < \infty. \tag{5}$$

The open-loop and feedback systems considered below admit unique nonexplosive strong solutions and satisfy

$$\sup_{K \ge 1} \sup_{q \in \{0, K\}, \pi} \mathbb{E}\left[\sup_{t \le T} |\mathbf{S}_t^{q,\pi}|^4 + \|N_T^{q,\pi}\|_1^4 + \int_0^T \|\lambda_t^{q,\pi}\|_1^4\, dt\right] < \infty. \tag{6}$$

Remark (Sufficient conditions)*. Suppose Assumption A.2 and* (4)–(5) *hold, $\mathbb{E}[|\mathbf{S}_0|^4] < \infty$, and controls take values in a compact set with $0 \le \nu \le \bar{\nu} < \infty$. For feedback controls, assume that after substituting the policy into both the coefficients and the intensities, the resulting drift and diffusion are locally Lipschitz in (X, R) between jumps, for each fixed regime and past incident history, uniformly in time on bounded sets. These conditions suffice for unique nonexplosive strong solutions and* (6)*, uniformly in K and the policy. This is a bounded-kernel variant of the sufficient conditions in (Bielecki et al., 2026, Proposition 3.6), allowing intensity-dependent diffusion and state- and control-independent marks with finite fourth moments. No subcriticality condition is needed on the finite horizon $[0, T]$.*

Assumption A.4 (Rewards)**.** *Let $r_0(t, \mathbf{s}, i, \pi) = f(t, x) - c(t, \mathbf{s}, i, \pi)$, where the operating cost includes the effective AI action. For every regime i,*

$$|r_0(t, \mathbf{s}, i, \pi) - r_0(t, \mathbf{s}', i, \pi')| + |g(x) - g(x')| \le L(1 + |\mathbf{s}| + |\mathbf{s}'|)(|\mathbf{s} - \mathbf{s}'| + |\pi - \pi'|),$$
$$|r_0(t, 0, i, \pi)| + |g(0)| \le L.$$

Equivalent expected objective. Put $m_{2,j}^X = \int |m_X|^2 \rho_j(dm)$ and define

$$\mathfrak{r}(t,\mathbf{s},i,\pi,\ell) = r_0(t,\mathbf{s},i,\pi) - \frac{\kappa(i)}{2}\nu^2 - \chi_{i,1-i}\nu - \frac{\eta}{2}\left(\|\Sigma_X(t,\mathbf{s},i,\pi,\ell)\|^2 + \sum_j m_{2,j}^X \ell_j\right). \quad (7)$$

Compensation of the incident and switching measures gives exactly

$$J^{q,\mathrm{rel}}(\pi) = \mathbb{E}\left[g(X_T^q) + \int_0^T \mathfrak{r}(t,\mathbf{S}_{t-}^q, I_{t-}^q, \pi_t^q, \lambda_t^q)\,dt\right]. \quad (8)$$

As a consequence of Assumption A.4 for any fixed $i \in \{0,1\}$ we have

$$\begin{aligned}&|\mathfrak{r}(t,\mathbf{s},i,\pi,\ell) - \mathfrak{r}(t,\mathbf{s}',i,\pi',\ell')| \\ &\quad \le C(1+|\mathbf{s}|+|\mathbf{s}'|+\|\ell\|_1+\|\ell'\|_1)(|\mathbf{s}-\mathbf{s}'|+|\pi-\pi'|+\|\ell-\ell'\|_1).\end{aligned}$$

A.2 Filter representation and inherited Hawkes estimates

For any incident history N^j, $j \in \{S,A\}$ we define

$$Z_t^{K,\ell,j} = \int_{(0,t]} e^{-\beta\ell(t-s)}\,dN_s^j, \qquad \ell = 1,\dots,K.$$

As in (Bielecki et al., 2026, Section 3.2),

$$dZ_t^{K,\ell,j} = -\beta\ell Z_t^{K,\ell,j}\,dt + dN_t^j, \qquad \boldsymbol{\lambda}_t^K = \mu(t,\bar{e}_t^K) + \sum_{\ell=1}^K Q_{K,\ell}(\bar{e}_t^K) Z_{t-}^{K,\ell}.$$

For later use, set $\mathfrak{b}_K(t) := \int_{(0,t)} d_K(t-s)\,d\|N_s\|_1$. Since $d_K \le 2H$, pathwise

$$\int_0^T \mathfrak{b}_K(t)\,dt \le \epsilon_K \|N_T\|_1, \qquad \sup_{t\le T}\mathfrak{b}_K(t) \le 2H\|N_T\|_1.$$

Consequently, uniformly over the controls considered,

$$\mathbb{E}\left[\int_0^T \left(\mathfrak{b}_K(t) + \mathfrak{b}_K(t)^2\right)dt\right] \le C\epsilon_K. \quad (9)$$

In particular, evaluating Φ^K on the original incident history, with the original effort, approximates its intensity in $L^1(\mathbb{P}\otimes dt)$ with error at most $C\epsilon_K$.

Lemma A.5 (Common-control intensity estimate)**.** *If $\pi^K = \pi$ is the same predictable control process, then*

$$A_K := \sup_\pi \mathbb{E}\left[\int_0^T \|\lambda_t^{K,\pi} - \lambda_t^\pi\|_1\,dt\right] \le C\epsilon_K. \quad (10)$$

Proof. Fix a predictable control $\pi^K = \pi$. Note that $J^K = J$ and $I^K = I$, hence $\bar{e}^K = \bar{e}$. The intensity comparison in (Bielecki et al., 2026, proof of Theorem 3.10, Step 2) therefore applies with the current control replaced by $\bar{e}$ and no baseline state-error term, since $\mu(t,e)$ is independent of $\mathbf{S}$. Together with (9) and the uniform resolvent bound in (Bielecki et al., 2026, Lemma 3.1), this gives (10). The constants are independent of K and π, so the estimate is uniform over the admissible controls. □

A.3 Full-information approximation

Let $\mathbb{F}^{\mathrm{full}}$ be the usual augmentation of the filtration generated by the common initial variables $(\mathbf{S}_0, I_0)$ and the common driving noises. Let $\mathcal{A}^{\mathrm{full}}$ contain all U-valued $\mathbb{F}^{\mathrm{full}}$-predictable controls. Define

$$V^{q,\mathrm{full}} := \sup_{\pi\in\mathcal{A}^{\mathrm{full}}} J^{q,\mathrm{rel}}(\pi).$$

These are values of the common strong full-information formulation, not of the smaller observable policy class below. The following theorem extends (Bielecki et al., 2026, Theorem 3.10) to turbulence risk adding the intensity process into the volatility of the controlled process X.

Theorem A.6 (Full-information convergence)**.** *Under Assumptions A.1–A.4, as $K \to \infty$,*

$$\sup_{\pi \in \mathcal{A}^{\mathrm{full}}} \mathbb{E}\left[\sup_{t \le T} |\mathbf{S}_t^{K,\pi} - \mathbf{S}_t^{\pi}|^2 + \int_0^T \|\lambda_t^{K,\pi} - \lambda_t^{\pi}\|_1^2 \, dt\right] \longrightarrow 0, \tag{11}$$
$$\Delta_K := \sup_{\pi \in \mathcal{A}^{\mathrm{full}}} |J^{K,\mathrm{rel}}(\pi) - J^{\mathrm{rel}}(\pi)| \longrightarrow 0.$$

Consequently $|V^{K,\mathrm{full}} - V^{\mathrm{full}}| \le \Delta_K \to 0$.

Proof. All suprema over π below are over $\mathcal{A}^{\mathrm{full}}$. Fix a predictable control π. In this proof, we use the notation $C > 0$ for a generic constant which can varies lines to lines and independent of the choice of π and K. The switching construction gives $I^K = I$. For the sake of simplicity we remove the superscript π in the notation and write $\Delta\mathbf{S} = \mathbf{S}^K - \mathbf{S}$ and $\Delta\lambda = \lambda^K - \lambda$. By Lemma A.5, $A_K \le C\epsilon_K \to 0$. Equation (6) gives $\sup_{K,\pi} \mathbb{E}[\int_0^T \|\Delta\lambda_t\|_1^4 \, dt] < \infty$. Hölder's inequality therefore yields

$$B_K := \sup_{\pi} \mathbb{E}\left[\int_0^T \|\Delta\lambda_t\|_1^2 \, dt\right] \le C A_K^{2/3} \le C\epsilon_K^{2/3} \longrightarrow 0. \tag{12}$$

We define for $j \in \{S, A\}$

$$\Delta b_s := b_{\mathbf{S}}(s, \mathbf{S}_{s-}^K, I_{s-}, \pi_s, \lambda_s^K) - b_{\mathbf{S}}(s, \mathbf{S}_{s-}, I_{s-}, \pi_s, \lambda_s),$$
$$\Delta\Sigma_s := \Sigma_{\mathbf{S}}(s, \mathbf{S}_{s-}^K, I_{s-}, \pi_s, \lambda_s^K) - \Sigma_{\mathbf{S}}(s, \mathbf{S}_{s-}, I_{s-}, \pi_s, \lambda_s),$$
$$\delta_j^K(s,\theta) := \mathbf{1}_{\{\theta \le \lambda_s^{K,j}\}} - \mathbf{1}_{\{\theta \le \lambda_s^{j}\}}, \qquad \widetilde{\Pi}^j(ds, d\theta, dm) := \Pi^j(ds, d\theta, dm) - ds\, d\theta\, \rho_j(dm).$$

From (3) we get

$$\Delta\mathbf{S}_t = \mathcal{D}_t^K + \mathcal{B}_t^K + \mathcal{M}_t^K + \mathcal{C}_t^K,$$

where

$$\mathcal{D}_t^K := \int_0^t \Delta b_s \, ds, \; \mathcal{B}_t^K := \int_0^t \Delta\Sigma_s \, dB_s, \; \mathcal{M}_t^K := \sum_j \int_0^t \int_{\mathbb{R}_+ \times \mathcal{M}_j} m\delta_j^K(s,\theta)\, \widetilde{\Pi}^j(ds, d\theta, dm),$$
$$\mathcal{C}_t^K := \sum_j \bar{m}_j \int_0^t \Delta\lambda_s^j \, ds, \quad \bar{m}_j := \int_{\mathcal{M}_j} m\rho_j(dm).$$

Set $F_K^{\pi}(t) := \mathbb{E}[\sup_{r \le t} |\Delta\mathbf{S}_r|^2]$, hence

$$F_K^{\pi}(t) \le 4(\mathbb{E}[\sup_{r \le t} |\mathcal{D}_r^K|^2] + \mathbb{E}[\sup_{r \le t} |\mathcal{B}_r^K|^2] + \mathbb{E}[\sup_{r \le t} |\mathcal{M}_r^K|^2] + \mathbb{E}[\sup_{r \le t} |\mathcal{C}_r^K|^2]).$$

Note that $b_{\mathbf{S}}$ does not depend on the intensity. Similarly of (Bielecki et al., 2026, proof of Lemma 3.8, Equation (17)) we get $|\Delta b_s| \le L|\Delta\mathbf{S}_{s-}|$. From Cauchy-Schwarz Inequality we have

$$\mathbb{E}\left[\sup_{r \le t} |\mathcal{D}_r^K|^2\right] \le t\, \mathbb{E}\left[\int_0^t |\Delta b_s|^2 \, ds\right]$$
$$\le TL^2 \int_0^t F_K^{\pi}(s) \, ds.$$

Then, by adapting the martingale estimate in (Bielecki et al., 2026, proof of Lemma 3.8, equations (18)–(19)) in L^2 together with Doob's inequality, then Itô's isometry, we get

$$\mathbb{E}\left[\sup_{r \le t} |\mathcal{B}_r^K|^2\right] \le 4\, \mathbb{E}\left[\int_0^t \|\Delta\Sigma_s\|_F^2 \, ds\right]$$
$$\le 8L^2 \int_0^t F_K^{\pi}(s) \, ds + 8L^2 B_K.$$

Note that the additional B_K term accounts for the intensity-dependent diffusion specific to our problem compared with Bielecki et al. (2026).

The thinning identity underlying (Bielecki et al., 2026, proof of Lemma 3.8, equation (14)) gives

$$\int_{\mathbb{R}_+} |\delta_j^K(s,\theta)|^2 \, d\theta = |\Delta\lambda_s^j|,$$

since $|\delta_j^K|^2 = |\delta_j^K|$. Doob's inequality and the marked-Poisson isometry together with (5) therefore give

$$\begin{aligned}\mathbb{E}\left[\sup_{r\le t}|\mathcal{M}_r^K|^2\right] &\le 4\sum_j \left(\int_{\mathcal{M}_j}|m|^2\rho_j(dm)\right)\mathbb{E}\left[\int_0^t |\Delta\lambda_s^j|\,ds\right]\\ &\le CA_K.\end{aligned}$$

Then, using $\int_{\mathbb{R}_+}\delta_j^K(s,\theta)\,d\theta = \Delta\lambda_s^j$ and Cauchy-Schwarz Inequality, we get

$$\begin{aligned}\mathbb{E}\left[\sup_{r\le t}|\mathcal{C}_r^K|^2\right] &\le t\left(\sum_j |\bar{m}_j|^2\right)\mathbb{E}\left[\int_0^t \|\Delta\lambda_s\|_1^2\,ds\right]\\ &\le CB_K.\end{aligned}$$

Combining the four estimates with $|v_1+v_2+v_3+v_4|^2 \le 4\sum_{k=1}^4 |v_k|^2$ gives

$$F_K^\pi(t) \le C\int_0^t F_K^\pi(s)\,ds + C(A_K+B_K).$$

Gronwall's inequality therefore implies

$$\sup_\pi F_K^\pi(T) \le C(A_K+B_K) \longrightarrow 0.$$

Together with (12), this proves (11). Finally, adapting (Bielecki et al., 2026, proof of Theorem 3.10, Step 5), with the effective reward (7), Assumption A.4, (6), and Cauchy-Schwarz' Inequality, we get

$$|J^{K,\mathrm{rel}}(\pi) - J^{\mathrm{rel}}(\pi)| \le C\big(F_K^\pi(T)+B_K\big)^{1/2}.$$

Thus $\Delta_K \le C(\sup_\pi F_K^\pi(T)+B_K)^{1/2} \to 0$, therefore $|V^{K,\mathrm{full}} - V^{\mathrm{full}}| \le \Delta_K \to 0$. □

A.4 Information loss under current-observation feedback

Fix $K \ge 1$ and a deterministic frozen recommendation (black-box feedback from the AI decision) G defined by $G(t,\mathbf{S}_t) := \widehat{a}_t^{\mathrm{raw}} = a^*(t,X_t) + R_t$, and impose Assumptions A.1–A.4. We use the same notations that the one introduced in Sections A.1–A.2. Let O_t^K be the current observation vector, *i.e.* $O_t = (t, X_t, I_t, Z_t, N_t^S, N_t^A, \mathbf{1}_{\{I_t=1\}}\widehat{a}^{\mathrm{raw}})$. The raw recommendation $I_t^K G(t,\mathbf{S}_t^K)$ is observed, but not R_t^K separately. Assume this raw output has a càdlàg version whose left limit is $I_{t-}^K G(t,\mathbf{S}_{t-}^K)$, $dt\otimes\mathbb{P}^\pi$-almost everywhere.

Let $\mathcal{A}^{K,\mathrm{obs}}$ consist of deterministic Borel feedbacks

$$\pi_t = f(t, O_{t-}^K) \in U$$

whose closed-loop systems are admissible. Define

$$V^{K,\mathrm{obs}} := \sup_{\pi\in\mathcal{A}^{K,\mathrm{obs}}} J^{K,\mathrm{rel}}(\pi).$$

Note that $\mathcal{A}^{K,\mathrm{obs}} \subseteq \mathcal{A}^{\mathrm{full}}$. In particular, both classes use (8), the same initial law, and the same continuous-time decision convention but $\mathcal{A}^{K,\mathrm{obs}}$ does not obsevred R directly.

Write $\lambda = \lambda_i^K(t,z,\pi)$ for the intensity in Section A.2, evaluated at (i,z,π). For a smooth pair $\varphi = (\varphi_0,\varphi_1)$, define the generator by

$$\begin{aligned}(\mathcal{L}^{K,\pi}\varphi)_i(t,\mathbf{s},z) := b_{\mathbf{S}}\cdot D_{\mathbf{s}}\varphi_i + \frac{1}{2}\operatorname{tr}\big(\Sigma_{\mathbf{S}}\Sigma_{\mathbf{S}}^\top D_{\mathbf{s}}^2\varphi_i\big) - \sum_{\ell=1}^K \sum_{j\in\{S,A\}} \beta\ell z^{\ell,j}\partial_{z^{\ell,j}}\varphi_i\\ + \sum_{j\in\{S,A\}}\lambda^j\int_{\mathcal{M}_j}\left[\varphi_i(t,\mathbf{s}+m,z+\delta_j)-\varphi_i(t,\mathbf{s},z)\right]\rho_j(dm)\\ + \nu[\varphi_{1-i}(t,\mathbf{s},z)-\varphi_i(t,\mathbf{s},z)] + \mathfrak{r}(t,\mathbf{s},i,\pi,\lambda_i^K(t,z,\pi)),\end{aligned}$$

with $\mathfrak{r}$ from (7).

Assumption A.7. *We assume that there exists an estimator h for which is a deterministic Borel-measurable map $h(t, o) \in \mathbb{R}$, with associated estimate*

$$\widehat{R}_t^{K,h} := h(t, O_{t-}^K).,$$

such that for any policy $\pi \in \mathcal{A}^{K,obs}$

$$\mathbb{E}^\pi \int_0^T |R_{t-}^K - \widehat{R}_t^{K,h}|^2 \, dt < \infty.$$

Assumption A.8 (Verification, sensitivity, and feedback selection)**.** *For the fixed K, the following hold.*

(i) There exists a unique pair of solutions (v_0^K, v_1^K) which are continuously differentiable with respect to (t, z) and twice continuously differentiable with respect to (x, r), to the coupled integro-partial HJB equation for $i \in \{0, 1\}$

$$\partial_t v_i^K + \sup_{u \in U} \mathcal{L}^{K,u} v_i^K(\cdot; \pi) = 0, \qquad v_i^K(T, \mathbf{s}, z) = g(x),$$

such that

$$V^{K,\text{full}} = v_{I_0}^K(0, \mathbf{S}_0, \mathcal{Z}_0^K).$$

Moreover, v_i^K is continuous up to the terminal time, and there exists a finite constant C_K such that

$$|v_i^K(t, s, z)| \le C_K\big(1 + |s|^2 + \|z\|^2\big), \qquad (t, s, z) \in [0, T] \times \mathbb{R}^2 \times \mathcal{Z}_K, \quad i \in \{0, 1\}.$$

(ii) The Hamiltonian is Borel measurable and continuous in the control; its control suprema are measurable. For some $L_K < \infty$,

$$|\mathcal{L}^{K,\pi} v_i^K(t, x, z, r; u) - \mathcal{L}^{K,\pi} v_i^K(t, x, z, r'; u)| \le L_K |r - r'|, \qquad u \in U.$$

(iii) There exists a deterministic Borel feedback $\hat{\pi}_t^{\text{obs}}$ such that $\hat{\pi}_t^{\text{obs}} = \hat{\pi}^{*,\text{obs}}(t, O_{t-}^K)$ belongs to $\mathcal{A}^{K,\text{obs}}$ and satisfies

$$\hat{\pi}^{*,\text{obs}}(t, o) \in \arg\max_{u \in U} \mathcal{L}^{K,\pi} v_i^K(t, x, z, h(t, o), u).$$

Remark A.9. *Assumption A.7 is structural to our problem and depends on the capacity of the supervisor to monitor AI error estimation. Assumption A.8 is a classical-solution (verification) hypothesis: when a classical solution exists, the argument of Bielecki et al. (2026) identifies it with $V^{K,full}$. Existence for the coupled system is not addressed here.*

We now set

$$B_h^K(t) := \mathbb{E}^{\hat{\pi}^{*,\text{obs}}} |R_{t-}^K - \widehat{R}_t^{K,h}|^2.$$

Remark A.10. *Conditioning on the current vector O_{t-}^K, rather than its history, gives the orthogonal decomposition*

$$B_h^K(t) = \mathbb{E}^{\hat{\pi}^{*,\text{obs}}} \operatorname{Var}^{\hat{\pi}^{*,\text{obs}}}(R_{t-}^K \mid O_{t-}^K) + \mathbb{E}^{\hat{\pi}^{*,\text{obs}}} \left| \mathbb{E}^{\hat{\pi}^{*,\text{obs}}}[R_{t-}^K \mid O_{t-}^K] - h(t, O_{t-}^K) \right|^2.$$

Thus $B_h^K(t)$ equals the expected conditional variance precisely when $h(t, O_{t-}^K)$ is the conditional mean under the comparison policy.

Theorem A.11 (Current-feedback value bound for fixed G)**.** *Under the assumptions above, there exists a function $L_K > 0$ such that*

$$0 \le V^{K,\text{full}} - V^{K,\text{obs}} \le 2L_K \int_0^T \sqrt{B_h^K(t)} \, dt.$$

Proof. We set

$$\widehat{\mathbf{S}}_t^{K,h} := (X_{t-}^K, \widehat{R}_t^{K,h}).$$

We define the Hamiltonian

$$\mathcal{H}_i^K(t,\mathbf{s},z) := \sup_{u\in U}\big(\mathcal{L}^{K,u}v^K\big)_i(t,\mathbf{s},z),$$

and

$$\mathcal{R}_t^{K,h} := \mathcal{H}^K_{I^K_{t-}}(t,\mathbf{S}^K_{t-},Z^K_{t-}) - \big(\mathcal{L}^{K,\pi^{*,\mathrm{obs}}_t}v^K\big)_{I^K_{t-}}(t,\mathbf{S}^K_{t-},Z^K_{t-}).$$

Using the maximizing property at $\widehat{\mathbf{S}}^{K,h}_t$ gives

$$\begin{aligned}\mathcal{R}_t^{K,h} &= \mathcal{H}^K_{I^K_{t-}}(t,\mathbf{S}^K_{t-},Z^K_{t-}) - \mathcal{H}^K_{I^K_{t-}}(t,\widehat{\mathbf{S}}^K_{t-},Z^K_{t-})\\ &\quad + \big(\mathcal{L}^{K,\pi^{*,\mathrm{obs}}_t}v^K\big)_{I^K_{t-}}(t,\widehat{\mathbf{S}}^{K,h}_t,Z^K_{t-}) - \big(\mathcal{L}^{K,\pi^{*,\mathrm{obs}}_t}v^K\big)_{I^K_{t-}}(t,\mathbf{S}^K_{t-},Z^K_{t-}).\end{aligned}$$

Recalling that $|\sup_{u\in U}a_u - \sup_{u\in U}b_u| \le \sup_{u\in U}|a_u-b_u|$, together with Assumption A.8(ii) implies

$$0 \le \mathcal{R}_t^{K,h} \le 2L_K\left|R^K_{t-} - \widehat{R}^{K,h}_t\right|.$$

In particular, Assumption A.7 yields

$$\mathbb{E}^{\pi^{*,\mathrm{obs}}}\Big[\Big]\int_0^T \mathcal{R}_t^{K,h}\,dt\Big] < \infty.$$

For the sake of simplicity, we denote the effective running reward by

$$r_t := r\big(t,S^K_{t-},I^K_{t-},\pi^{*,\mathrm{obs}}_t,\lambda^K_t\big).$$

Applying the jump-diffusion Itô formula to $v^K_{I^K_t}(t,\mathbf{S}^K_t,Z^K_t)$, we obtain from the integro-partial HJB equation

$$v^K_{I^K_t}(t,\mathbf{S}^K_t,Z^K_t) + \int_0^t r_s\,ds = v^K_{I_0}(0,\mathbf{S}_0,Z^K_0) - \int_0^t \mathcal{R}_s^{K,h}\,ds + M_t,$$

where M is a local martingale starting at zero.

From Assumption A.8(i), Assumption A.3, and $Z^{K,\ell,j}_t \le N^{K,j}_t$ we get

$$\mathbb{E}^{\pi^{*,\mathrm{obs}}}[\sup_{t\le T}|v^K_{I^K_t}(t,\mathbf{S}^K_t,Z^K_t)|] + \mathbb{E}^{\pi^{*,\mathrm{obs}}}[\int_0^T |r_t|\,dt] < \infty.$$

Using a localization approach, the martingale term has zero expectation. Using the terminal condition and the verification identity in Assumption A.8(i), we conclude that

$$V^{K,\mathrm{full}} - J^{K,\mathrm{rel}}(\pi^{*,\mathrm{obs}}) = \mathbb{E}^{\pi^{*,\mathrm{obs}}}[\int_0^T \mathcal{R}_t^{K,h}\,dt].$$

Finally, since $\pi^{*,\mathrm{obs}} \in \mathcal{A}^{K,\mathrm{obs}} \subseteq \mathcal{A}^{\mathrm{full}}$,

$$\begin{aligned}0 &\le V^{K,\mathrm{full}} - V^{K,\mathrm{obs}}\\ &\le V^{K,\mathrm{full}} - J^{K,\mathrm{rel}}(\pi^{*,\mathrm{obs}})\\ &\le 2L_K\int_0^T \mathbb{E}^{\pi^{*,\mathrm{obs}}}\left|R^K_{t-} - \widehat{R}^{K,h}_t\right|\,dt\\ &\le 2L_K\int_0^T \sqrt{B^K_h(t)}\,dt,\end{aligned}$$

where the last inequality follows from Cauchy–Schwarz. The right-hand side is finite because Assumption A.7 gives

$$\int_0^T \sqrt{B^K_h(t)}\,dt \le \left(T\,\mathbb{E}^{\pi^{*,\mathrm{obs}}}\int_0^T \left|R^K_{t-} - \widehat{R}^{K,h}_t\right|^2 dt\right)^{1/2} < \infty.$$

□

Remark A.12. *The upper bound on the error made between the value function with full observation, including AI error observation and the value with estimated AI error depends on the variance accuracy of this estimator. Variance reductiton technics can be applied but are out of the scope of this study and we let this point for future research.*

A.5 Error decomposition for the learned policy

Fix $K \geq 1$ and the initial law used in the preceding sections. Let $\widehat{\pi}_K \in \mathcal{A}^{K,\mathrm{obs}}$ be an admissible learned current-observation feedback, evaluated in the K-term approximating model under the same continuous-time decision convention. Define its remaining suboptimality by

$$\varepsilon^K_{\mathrm{learn}} := V^{K,\mathrm{obs}} - J^{K,\mathrm{rel}}(\widehat{\pi}_K) \geq 0.$$

The discrepancy from the original full-information optimum admits the exact decomposition

$$V^{\mathrm{full}} - J^{K,\mathrm{rel}}(\widehat{\pi}_K) = \big(V^{\mathrm{full}} - V^{K,\mathrm{full}}\big) + \big(V^{K,\mathrm{full}} - V^{K,\mathrm{obs}}\big) + \varepsilon^K_{\mathrm{learn}}.$$

The first difference need not be nonnegative, but its absolute value is bounded by Δ_K from Theorem A.6. Consequently, under the assumptions of Theorem A.11,

$$V^{\mathrm{full}} - J^{K,\mathrm{rel}}(\widehat{\pi}_K) \leq \underbrace{\Delta_K}_{\text{kernel approximation}} + \underbrace{2L_K \int_0^T \sqrt{B^K_h(t)}\, dt}_{\text{current-observation restriction}} + \underbrace{\varepsilon^K_{\mathrm{learn}}}_{\text{neural approximation and training}}. \tag{13}$$

This decomposition separates the errors addressed by the theoretical analysis from those associated with the numerical policy solver. Section A.3 controls the kernel-approximation term, with $\Delta_K \to 0$, while Section A.4 bounds the loss from restricting control to the current observation features. The latter bound need not vanish as K increases.

The residual $\varepsilon^K_{\mathrm{learn}}$ includes both the approximation error of the neural feedback class and the suboptimality remaining after finite training. The present analysis retains this residual explicitly and does not establish a certified optimality bound for the trained PPO policy.

In (13), $\widehat{\pi}_K$ denotes the continuously evaluated feedback associated with the learned network. Its sampled deployment and numerical simulation, as well as evaluation under the original kernel when applicable, introduce additional discrepancies that are separate from the three terms displayed above.

A.6 Markovianization procedure and Exponential-mixture fit

The first step of the Hawkes-PPO algorithm requires to approach the non-Markovian kernel Φ with a mixture of exponential. In this context, since the value of the problem with non-Markovian kernel is approached with a mixture of Markovian exponential kernels, we refer to Algorithm 1 in Bielecki et al. (2026) for this fitting preliminary procedure, mathematically justified by the results of this Appendix. We use the same procedure for updating the memory process Z in Algorithm 1.

B Numerical Implementation and Benchmarks

This appendix specifies the common numerical environment, the construction of the AI recommender, the observation-based RL methods, and the full-information benchmarks. All reported outcomes use the same physical model of the AI black box. The full-information value functions below define the oracle benchmarks; they do not replace the observation-based control problem in the main text.

B.1 Choice of parameters and evaluation protocol

Table 2: Physical, AI and simulation parameters. Kernel rows/columns follow (S, A). Learning settings are in the appendix.

Parameter	Numerical choice	Parameter	Numerical choice
(T, Δ)	$(8, 0.05)$	(X_0, R_0, Z_0)	$(0, 0, 0)$
(a, e, ν)	$[0, 1]^2 \times [0, 4]$	ε	0.01
δ	0.18	(σ_0, σ_1)	$(0.16, 0.04)$
$f(x)$	$0.1x + 0.05\tanh(x - 1)$	$g(x)$	$1.2x + 0.6\tanh(x - 1)$
(k_0, k_a, k_e)	$(3.25, 0.75, 1.25)$	k_{AI}	1.5
(η, κ, χ)	$(50, 0.30, 0.20)$	μ	$(0.35, 0.25)$
M^S	$\mathrm{Unif}[0.10, 0.45]$	M^A	$\mathrm{Unif}[0.35, 0.85]$
(γ_0, γ_1)	$(0.1, 1.1)$	$s(L)$	$0.0125 + \frac{0.0775L}{0.75+L}$
(θ_0, θ_1)	$(2, 2)$	A	$\begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{pmatrix}$
(ρ_{ij}, β_{ij})	$(2, 3/2)$	Filter rates	$\beta_k = k/4, \quad k = 1, ..., 20$

Controls $(a, e, \nu) \in [0, 1]^2 \times [0, 4]$ are selected every $\Delta = 0.05$; each decision interval comprises five microsteps of $\delta t = 0.01$. Coefficients are frozen within each microstep. Linear state, error, and filter decay use exponential integration, including event-age decay for jump effects. We simulate risk events and regime switches by Poisson thinning, using shared candidate streams across policies to reduce noise in comparisons. The thinning Poisson rate is 32 per channel and the switching rate is 4; a capacity of 10 events per channel per microstep is checked, with a run invalidated if its bound or capacity is exceeded. Switches take effect at microstep endpoints; a fresh supervisory decision occurs at the next decision epoch. The committed Human action and mitigation can therefore become active after an AI-to-Human switch within the interval. The current recommendation $\widehat{a}_t$ is recomputed at every microstep. All risk events remain observable in both regimes. The public RL observation is

$$O_t = (t, X_t, I_t, Z_t, N_t^S, N_t^A, \mathbf{1}_{\{I_t=1\}}\widehat{a}^{\mathrm{raw}}). \tag{14}$$

The full-information oracle additionally observes R_t in both regimes and knows all model coefficients and the nominal recommendation map.

All displayed objectives are undiscounted. The realized quadratic variation is accumulated as

$$[X]_T = \int_0^T \{\sigma(\alpha_t)^2 + \gamma(L_t)^2\}\, dt + \sum_{\tau_n^S \le T} (M_n^S)^2.$$

B.2 Project-Only AI Recommender: black-box design

The nominal recommendation is computed before supervisory training and is then frozen for all methods. The AI solves a project-only problem on (t, x), with control $b \in [0, 1]$, dynamics

$$dY_s = (b_s - 0.18Y_s)\, ds + \sigma(b_s)\, dW_s, \qquad \sigma(b) = 0.16 + 0.04b,$$

and value

$$v(t, x) = \sup_b \mathbb{E}_{t,x}\left[g(Y_T) + \int_t^T \left\{ f(Y_s) - 1.5 - b_s^2 - \frac{\eta}{2}\sigma(b_s)^2 \right\} ds \right], \qquad \eta = 50.$$

Here $T = 8$, and f, g are the reward same as in Table 2. The quadratic variations term in this nominal objective is intrinsic project variation: $[Y]_T - [Y]_t = \int_t^T \sigma(b_s)^2 ds$. The AI ignores Hawkes risk, added risk diffusion, its own recommendation errors, mitigation, or switching in this optimization.

The project-only HJB is

$$\begin{aligned} 0 &= v_t + \sup_{0 \le b \le 1} \left\{ (b - 0.18x)v_x + \tfrac{1}{2}\sigma(b)^2 v_{xx} + f(x) - 1.5 - b^2 - 25\sigma(b)^2 \right\}, \\ v(T, x) &= g(x). \end{aligned} \tag{15}$$

We use a DGM network, introduced in Sirignano & Spiliopoulos (2018), approximates v, using automatic differentiation of the continuous HJB residual and a value parameterization that imposes the terminal condition exactly.

For given network derivatives $p = v_x$ and $q = v_{xx}$, the control-dependent Hamiltonian is

$$H(b; p, q) = bp - b^2 + \left(\frac{q}{2} - 25\right)(0.16 + 0.04b)^2.$$

Its bounded maximum is evaluated by comparing the endpoints $b = 0, 1$ and, when $2.08 - 0.0016q > 0$, the candidate

$$b_{\text{int}} = \text{clip}\left(\frac{\text{p} + 0.0064\text{q} - 0.32}{2.08 - 0.0016\text{q}}, 0, 1\right).$$

Outside the exported state grid, we approximate the reward by its affine asymptote and use the resulting analytical optimal action, retaining the project quadratic-variation penalty. The extrapolation action is the clipped affine-tail optimum

$$\text{clip}\left(\frac{0.1/0.18 + (1.2 - 0.1/0.18)\text{e}^{-0.18(8-\text{t})} - 0.32}{2.08}, 0, 1\right).$$

For a nonconcave Hamiltonian, only the endpoints are required. The maximizing action defines $a^*(t, x)$.

Table 3: Project-only recommender training and deployment.

Setting	Choice
Network and optimizer	Width 64, one gated DGM block, Adam, gradient norm cap 10.
Training	6,000 updates; batch 2,048; learning rates 10^{-3} for updates 1–3,000, $3 \cdot 10^{-4}$ for 3,001–4,800, and 10^{-4} for 4,801–6,000.
Collocation and selection	t uniform on $[0, 8]$; x sampled 65% from $[-4, 6]$ and 35% from $[-16, 16]$. Residual validation on 8,192 points every 250 updates; selected update 6,000.
Exported policy	Grid $[0, 8] \times [-12, 12]$, spacings $\Delta t = \Delta x = 0.01$, bilinear interpolation; analytic affine-tail action outside the x-range.
Independent check	HJB residual RMSE $4.596 \cdot 10^{-4}$; off-grid action interpolation RMSE $6.176 \cdot 10^{-7}$. Initial recommendation $a^*(0, 0) = 0.30080232$.

The deployed black box combines this frozen nominal controller with the persistent error process:

$$\begin{aligned} \widehat{a}_t &= \text{clip}_{[0,1]}(\widehat{a}_t^{\text{raw}}), \quad \widehat{a}_t^{\text{raw}} = a^*(t, X_t) + R_t \\ dR_t &= -2R_t\, dt + 2s(L_t)\, dW_t^R - M^A dN_t^A, \\ s(L) &= 0.0125 + \frac{0.0775L}{0.75 + L}, \qquad M^A \sim U[0.35, 0.85], \qquad R_0 = 0. \end{aligned}$$

RL observes $\widehat{a}_t^{\text{raw}}$ only in AI mode, while the project implements the clipped action $\widehat{a}_t$. Neither $a^\star(t, X_t)$ nor R_t is supplied separately. The pure autonomous-AI benchmark fixes $I \equiv 1$ and $\nu \equiv 0$ and uses this same corrupted recommendation, AI mitigation 0.01, and the full supervisory objective. Thus it is evaluated with all Hawkes risks, the full project QV penalty, and AI cost $1.5 + \alpha_t^2$, despite the nominal recommender solving only (15).

B.3 Hawkes-PPO algorithm and foundations

Hawkes-PPO is PPO (Schulman et al., 2017) applied to the public observation specified in (14): time, project state, current regime, event counts, the common Hawkes filter bank, and the current raw recommendation only in AI mode. In Human mode the recommendation input is masked. Neither the nominal controller a^* nor the error state R is supplied to the actor or critic.

There are two independent actor networks and two independent value networks, indexed by the current regime. They share no trainable parameters across regimes. A single trained policy is used for both initial regimes; the network selected along a trajectory changes when the regime changes.

Each network has two width-128 hidden layers with SiLU activations. For observation o in regime i, the actor produces three means $\mu_{\theta_i}(o)$ and three log standard deviations $\ell_{\theta_i}(o)$, clipped to $[-5, 1]$, and samples

$$y \sim \mathcal{N}\Big(\mu_{\theta_i}(o), \operatorname{diag}(\mathrm{e}^{2\ell_{\theta_i}(o)})\Big), \quad (a, e, \nu) = \left(\frac{1+\tanh y_1}{2}, \frac{1+\tanh y_2}{2}, 2(1+\tanh y_3)\right). \tag{16}$$

Let r_k be the complete interval reward: running benefit less operating cost, $\kappa\nu^2/2$ with $\kappa = 0.30$, realized switch fees, and $\eta/2 = 25$ times the QV increment. Terminal benefit is included once, on termination. Training uses $\widetilde{r}_k = r_k/50$. For a rollout, old value predictions are kept fixed when forming the generalized advantage estimates:

$$\begin{aligned} \delta_k &= \widetilde{r}_k + (1-d_k)V_{\phi^{\mathrm{old}}_{I_{k+1}}}(o_{k+1}) - V_{\phi^{\mathrm{old}}_{I_k}}(o_k), \\ A_k &= \delta_k + 0.95(1-d_k)A_{k+1}, \qquad R_k = A_k + V_{\phi^{\mathrm{old}}_{I_k}}(o_k). \end{aligned}$$

Here d_k is the terminal indicator, the discount factor is one, and the advantage recursion starts from zero beyond the rollout boundary. Crucially, the bootstrap uses the next observation's regime value after a switch, and is suppressed on terminal transitions. The advantages are centered and standardized over the entire rollout, giving $\widehat{A}_k$.

Writing $q_k(\theta) = \pi_\theta(u_k \mid o_k)/\pi_{\theta^{\mathrm{old}}}(u_k \mid o_k)$, the minimized actor and value losses are

$$\begin{aligned} L_\pi &= -\,\mathbb{E}_{\mathrm{batch}}\Big[\min\{q_k(\theta)\widehat{A}_k, \operatorname{clip}(\mathrm{q_k}(\theta), 0.8, 1.2)\widehat{\mathrm{A}}_\mathrm{k}\}\Big] - 0.003\,\mathbb{E}_{\mathrm{batch}}[\mathcal{H}_G(o_k)], \\ L_V &= \tfrac{1}{2}\mathbb{E}_{\mathrm{batch}}\Big[(V_{\phi_{I_k}}(o_k) - R_k)^2\Big]. \end{aligned}$$

The likelihood calculation includes the $\tanh$ Jacobian; the fixed affine action-scale factor cancels in the ratio. The entropy term is specifically that of the latent Gaussian, $\mathcal{H}_G = \sum_{j=1}^{3}[\ell_j + \frac{1}{2}\log(2\pi e)]$, not the entropy of the bounded physical controls. It is a training regularizer and is excluded from reported objective values.

Table 4: Hawkes-PPO training configuration.

Setting	Choice
Regime networks	Actor $46-128-128-6$; value $46-128-128-1$, independently for each regime.
Initialization	Orthogonal weights, hidden gain $\sqrt{2}$; actor output gain 0.01, value output gain 1; log-standard-deviation bias -0.7. Switching-mean bias $\operatorname{atanh}(-0.95)$; initial $a, e \approx 0.5$, $\nu \approx 0.1$.
Parallel sampling	1,024 environments, exactly 512 AI starts and 512 Human starts. Each reset restores the assigned initial regime.
Rollout and update	32 decisions per environment; four shuffled epochs; minibatch 4,096; Adam; actor/value gradient norm caps 1 and 5.
Budget	480 rollouts, 15,728,640 transitions, 15,360 actor and 15,360 value updates. No replay buffer or target-network soft updates.
Initial learning rates	Actor $3 \cdot 10^{-4}$, value $5 \cdot 10^{-4}$; common schedule below.
Selection	Every 786,432 transitions, 1,024 development paths per initial regime; maximize the equally weighted mean of both start values. Selected checkpoint at 15,728,640 transitions.
Seeds	Fit index 1; initialization 760001000; training environment 760000001; development 761000001.

Both learning rates are multiplied by the same schedule $c(n)$, where n counts collected environment transitions:

$$c(n) = \begin{cases} 1, & n \le n_0, \\ 0.1 + 0.45\left[1 + \cos\left(\pi\dfrac{n-n_0}{n_1-n_0}\right)\right], & n_0 < n < n_1, \\ 0.1, & n \ge n_1, \end{cases} \quad n_0 = 3{,}932{,}160, \quad n_1 = 11{,}796{,}480.$$

The small switching initialization changes only the two switching-mean output biases; it does not remove Gaussian exploration. It therefore initializes the deterministic action near $\nu = 0.1$, rather than forcing all sampled switching intensities to equal 0.1.

At development and final evaluation, the Gaussian mean replaces the sampled latent action in (16). Thus controls are deterministic functions of current observations, but switches remain stochastic at the selected intensity. The same frozen checkpoint is evaluated from both starts. Training uses fresh on-policy rollouts, without oracle labels, saved teacher paths, arbitrary-state reset data, or additional improvement rounds outside the stated budget.

Algorithm 1 Hawkes-PPO with balanced starts and regime-specific networks

Input: Simulator and public observation map; $M = 1024$ environments, $L = 32$ steps, $K = 480$ rollouts, $K_{\mathrm{ep}} = 4$ epochs, minibatch size $B = 4096$.
Input: Policy/value architectures, loss coefficients, and learning-rate schedule in Table 4.
Output: Selected parameters $\theta^\star = (\theta_0^\star, \theta_1^\star)$ for deterministic deployment.

Initialize (θ_0, θ_1) and (ϕ_0, ϕ_1) independently; set switching mean biases to $\mathrm{atanh}(-0.95)$.
Assign $M/2$ environments to each initial regime; reset $X = R = Z = N^S = N^A = 0$.
$J^\star \leftarrow -\infty$; $\theta^\star \leftarrow \theta$.
for $n = 1, \ldots, K$ **do**
    $(\theta_{\mathrm{old}}, \phi_{\mathrm{old}}) \leftarrow (\theta, \phi)$; initialize an empty rollout buffer $\mathcal{D}$.
    for $k = 0, \ldots, L-1$ **in all** M **environments in parallel do**
        Observe $o_k = (t_k, X_k, I_k, Z_k, N_k^S, N_k^A, \mathbf{1}_{\{I_k=1\}}\widehat{a}^{\mathrm{raw}})$.
        Sample z_k from the Gaussian actor $\theta_{\mathrm{old},I_k}$; set $u_k = \mathcal{T}(z_k)$.
        Hold u_k for one decision interval; advance five simulator microsteps.
        Observe (r_k, o_{k+1}, d_k); set $\widetilde{r}_k = r_k/50$.
        Store transition, z_k, old log probability and old value predictions in $\mathcal{D}$.
        Reset terminated environments to their assigned initial regimes.
    end for
    Compute δ_k, A_k and R_k by terminal-masked GAE with $\gamma = 1$, $\lambda_{\mathrm{GAE}} = 0.95$.
    Standardize A_k over $\mathcal{D}$ to obtain $\widehat{A}_k$; freeze R_k and old predictions.
    Set η_π, η_V using the transition-count schedule at nML.
    for $j = 1, \ldots, K_{\mathrm{ep}}$ **do**
        for each shuffled minibatch $\mathcal{B} \subset \mathcal{D}$ of size B **do**
            Route each sample through its current-regime actor and value network.
            $\theta \leftarrow \mathrm{AdamStep}(\theta, \nabla_\theta L_{\mathrm{actor}}(\mathcal{B}), \eta_\pi;\ \text{norm cap } 1)$.
            $\phi \leftarrow \mathrm{AdamStep}(\phi, \nabla_\phi L_{\mathrm{value}}(\mathcal{B}), \eta_V;\ \text{norm cap } 5)$.
        end for
    end for
    if nML is a multiple of 786432 **then**
        Evaluate $u = \mathcal{T}(\mu_{\theta_I}(o))$ on 1024 development paths per initial regime.
        $\widehat{J} \leftarrow \frac{1}{2}(\widehat{J}_{\mathrm{AI}} + \widehat{J}_{\mathrm{Human}})$.
        if $\widehat{J} > J^\star$ **then**
            $(J^\star, \theta^\star) \leftarrow (\widehat{J}, \mathrm{copy}(\theta))$.
        end if
    end if
end for
return $\theta^\star$; report separate-start outcomes on 4096 fresh common-noise paths per start.

B.4 FULL-INFORMATION SWITCHING ORACLE

The oracle knows the model coefficients, the frozen map $a^*(t, X_t)$ and the current recommendation error in either regime. Write its value as $V_i(t, x, z, \zeta)$, where ζ represents the process R. The additional coordinate describes oracle information and is not an extra observation supplied to RL. For $j \in \{S, A\}$, let $\mathbf{e}^j$ denote the filter increment with components $(\mathbf{e}^j)^{\ell,k} = \mathbf{1}_{\{k=j\}}$. With $\alpha_0 = a$, $\alpha_1 = \mathrm{clip}(a^*(t, x) + \zeta, 0, 1)$, the controlled generator, excluding regime switching, of a test function $F(t, x, z, \zeta)$ is

$$\begin{aligned}\mathcal{L}_i^{a,e} F(t,x,z,\zeta) = (\alpha_i - 0.18x)F_x - 2\zeta F_\zeta - \sum_{\ell,j} \beta_\ell z^{\ell,j} F_{z^{\ell,j}} \\ + \frac{1}{2}\big[\sigma(\alpha_i)^2 + \gamma(L_i)^2\big] F_{xx} + 2s(L_i)^2 F_{\zeta\zeta} \\ + \lambda_i^S \, \mathbb{E}_{M^S}\big[F(t, x - M^S, z + \mathbf{e}^S, \zeta) - F(t,x,z,\zeta)\big] \\ + \lambda_i^A \, \mathbb{E}_{M^A}\big[F(t, x, z + \mathbf{e}^A, \zeta - M^A) - F(t,x,z,\zeta)\big] .\end{aligned}$$

Let $\overline{m}_{S,2} = \mathbb{E}[(M^S)^2] = 0.0858333333333$ and $q_i = \sigma(\alpha_i)^2 + \gamma(L_i)^2 + \lambda_i^S \overline{m}_{S,2}$. The coupled HJB equations are

$$\begin{aligned}0 = \partial_t V_i + \sup_{a,e \in [0,1],\, \nu \in [0,4]} \Big\{ \mathcal{L}_i^{a,e} V_i + f(x) - c_i - 25 q_i + \nu(V_{1-i} - V_i - 0.20) - 0.15\nu^2 \Big\}, \\ V_i(8, x, z, \zeta) = g(x), \qquad i \in \{0, 1\}.\end{aligned}$$

The $\lambda_i^S \overline{m}_{S,2}$ term is the compensator of the jump-QV penalty; jump losses in V_i are separately represented by the generator. Switching changes the regime at the same (x, z, ζ), with no recommendation redraw or error reset. Optimization over a, e is vacuous in AI mode in the continuous problem. The switching maximizer is

$$\nu_i^* = \mathrm{clip}\left(\frac{V_{1-i} - V_i - 0.20}{0.30}, 0, 4\right).$$

For Human project control, maximize $aV_{0,x} + (\frac{1}{2}V_{0,xx} - 25)(0.16 + 0.04a)^2 - 0.75a^2$ on $[0, 1]$. The implementation compares endpoints and, when the quadratic is concave, the clipped stationary candidate

$$a_{\mathrm{stat}} = \frac{V_{0,x} + 0.0064 V_{0,xx} - 0.32}{1.58 - 0.0016 V_{0,xx}}.$$

Mitigation e is maximized over the grid $0, 0.1, \ldots, 1$, followed by local candidate refinements of size 0.05, 0.025, and 0.0125. This finite search approximates the continuous mitigation maximum.

DGM implementation. Two independently parameterized gated DGM networks Sirignano & Spiliopoulos (2018) represent the two regimes, each with width 64 and one gated block. Exactly zero-weight filters are omitted, leaving ten active filter coordinates; the input is $(t, x, \zeta, z_{\mathrm{active}}) \in \mathbb{R}^{13}$. The ansatz

$$V_i^\omega(t, x, z, \zeta) = g(x) + (8 - t)\{-8 + F_i^\omega(t, x, z, \zeta)\}$$

enforces the terminal condition exactly. First and second derivatives are obtained by automatic differentiation. Uniform-mark expectations use three-point Gauss–Legendre quadrature in training and five points in residual audits. The loss is the collocation average of $\frac{1}{2}\sum_{i=0}^{1}(\mathcal{R}_i/10)^2$, where $\mathcal{R}_i$ is the HJB residual with greedily selected controls. No RL labels or temporal Bellman targets are used for DGM training.

Training uses 6000 Adam updates with fresh batches of 512 collocation states, gradient-norm cap 10, and learning rate $10^{-3}[0.1 + 0.9\{1 + \cos(\pi n/6000)\}/2]$. A time curriculum samples t uniformly from $[8 - H_n, 8]$, with 12% of samples placed at the left endpoint and $H_n = \min\{8, 0.5 + 7.5n/2100\}$. Synthetic event histories generate physically interpretable filter states: conditional on rates sampled uniformly in $[0.15, 1.25]$, per-channel counts are Poisson with mean rate times t and are capped at 32 for collocation. Ages are uniform in $[0, t]$, with 15% of histories shifted toward recent events by the map $r \mapsto r^2/\max(t, 0.01)$; 15% of filter samples are set to zero. The project sample is $0.6(1 - e^{-0.18t}) + 0.7\sqrt{\max(t, 0.02)}\xi$, $\xi \sim \mathcal{N}(0, 1)$, with 20% replaced by $\mathcal{U}[-5, 6]$. Error samples

are $-0.6\sum_n e^{-2r_n^A} + 0.07\xi_E$, with 15% replaced by $\mathcal{U}[-2, 1.3]$. These are collocation distributions, not restrictions on physical state support or data supplied to RL.

Every 1000 updates, a candidate policy is evaluated on 512 development paths per initial regime, with equal-weight mean return used for selection. Initialization seed is 730000001 and development seed is 731000000. The selected checkpoint is update 3000, although all 6000 updates were completed. Its held-out collocation residual RMSEs are 0.1530 in Human and 0.5321 in AI; these describe the selected weights rather than a different residual-selected checkpoint. They do not provide an optimality certificate.

The oracle controls are deployed in the common fixed-grid simulator from Section B.1. Consequently, its continuous HJB value and its simulated return need not coincide exactly: controls are committed for a decision interval, and coefficients are frozen on microsteps. Reported objectives are Monte Carlo policy returns, not network predictions. We use "oracle" to mean a known-model, full-information numerical benchmark, not a certified exact optimum.

B.5 Optimal Fixed-Human Benchmark: pure human mode design

This benchmark solves the control problem with $I_t \equiv 0$ and $\nu_t \equiv 0$, while optimizing both a_t and e_t. It tests the benefit of switching against an optimized Human policy, rather than against constant Human actions. The numerical solution is an approximation to this optimal fixed-regime problem.

The recommendation error does not affect the Human project's coefficients or intensities, and there is no future AI delegation. Thus the sufficient state is (t, x, z), with value $W(t, x, z)$. Action-corruption events still matter through their excitation of future events.Using the filter increments $\mathbf{e}^S$ and $\mathbf{e}^A$ defined above, the fixed-Human generator is

$$\begin{aligned}\mathcal{L}_H^{a,e}W(t,x,z) = (a - 0.18x)W_x - \sum_{\ell,j}\beta_\ell z^{\ell,j}W_{z^{\ell,j}} \\ + \frac{1}{2}\left[\sigma(a)^2 + \gamma(L_0)^2\right]W_{xx} \\ + \lambda_0^S\,\mathbb{E}_{M^S}\left[W(t, x - M^S, z + \mathbf{e}^S) - W(t,x,z)\right] \\ + \lambda_0^A\left[W(t,x,z+\mathbf{e}^A) - W(t,x,z)\right].\end{aligned}$$

Here $\lambda_0^j = \lambda_0^j(z, e)$, $L_0 = \lambda_0^S + \lambda_0^A$, and all derivatives are evaluated at (t, x, z). The fixed-Human HJB is

$$\begin{aligned}0 = W_t + \sup_{a,e\in[0,1]}\Big\{\mathcal{L}_H^{a,e}W + f(x) - 3.25 - 0.75a^2 - 1.25e^2 \\ - 25[\sigma(a)^2 + \gamma(L_0)^2 + \lambda_0^S\overline{m}_{S,2}]\Big\}, \qquad W(8,x,z) = g(x).\end{aligned}$$

The action-event jump term remains even though no recommendation is implemented.

A single width-64, one-block gated DGM network uses the 12 inputs $(t, x, z_{\text{active}})$ and the hard-terminal ansatz $W^\omega = g(x) + (8 - t)(-12 + F_H^\omega)$. Action maximization, mitigation search, mark quadrature, Adam learning-rate schedule, gradient cap and time/filter collocation scheme are the same as for the switching oracle. The project collocation distribution is $2.5(1 - e^{-0.18t}) + 0.6\sqrt{\max(t, 0.02)}\xi$, with 20% replaced by $\mathcal{U}[-5, 6]$; no error-coordinate samples are needed. The loss is $\mathbb{E}_{\text{colloc}}[(\mathcal{R}_H/10)^2]$. Training completes 6000 updates with batch size 512 and initialization seed 730000017. Every 1000 updates, selection uses 512 Human-start development paths with seed 731000000. The selected update is 2000, with held-out residual RMSE 0.02234 and on-development-path residual RMSE 0.03286. The final independent Human-start objective is -96.027 with 90% interval $[-96.191, -95.864]$. As with the switching oracle, this is the return of an approximate HJB-derived policy deployed in the common simulator, rather than a certified value bound.

B.6 Additional RL Baselines

We train PPO, SAC Haarnoja et al. (2018), DDPG Lillicrap et al. (2015) and CT-DDPG Cheng et al. (2026 (v2); Bielecki et al. (2026) both with and without the same 40 Hawkes filters. The prefix "Hawkes-" denotes the filtered version. All receive $(t, X, I, N^S, N^A, \mathbf{1}_{\{I=1\}}\widehat{a}^{\text{raw}})$; only the filtered

versions receive Z. The unfiltered implementations set the 40 filter coordinates to zero, retaining the same network dimensions. Neither version receives R, the nominal recommendation a^*, the current recommendation in Human mode, or oracle labels. Every method uses separate networks for the current regime, rather than separate policies for the initial regime. All MLPs have two hidden layers of width 128 with SiLU activations. Actors have 46 inputs; SAC/DDPG action-value critics and CT-DDPG rate networks have 49 inputs. SAC has two action-value critics per regime, DDPG one, and CT-DDPG one value and one rate network per regime. Native PPO/SAC/DDPG heads are independently initialized; the two CT-DDPG regime heads start as identical copies with subsequently separate parameters. CT-DDPG initially sets input weights on filter coordinates to zero, and learns these weights during training.

All fits use 1,024 parallel environments, half assigned to each standard initial regime, with $\Delta = 0.05$, simulation step 0.01, undiscounted returns and reward divisor 50. Initial deterministic switching centers are approximately 0.1 in both regimes. Initial Human action/effort centers are $(0.5, 0.5)$ for PPO/SAC/DDPG and $(0.35, 0.45)$ for CT-DDPG. Exploration is still active: a small deterministic center does not mean that executed switching rates remain close to 0.1. PPO without filters otherwise uses exactly the Hawkes-PPO procedure, including its pre-tanh Gaussian entropy coefficient 0.003 and 480 rollouts of 32 decisions, each with four update epochs and minibatch size 4,096.

SAC and DDPG: Let $v \in [-1, 1]^3$ denote normalized actions, mapped to physical controls by $u = ((v_1 + 1)/2, (v_2 + 1)/2, 2(v_3 + 1))$. Write $\widetilde{r} = r/50$ and let d indicate episode termination. SAC samples a tanh-squashed diagonal Gaussian, uses twin critics, and minimizes

$$y_{\mathrm{SAC}} = \widetilde{r} + (1 - d)\left[\min_{j=1,2} \bar{Q}_j(o', v') - \alpha \log \pi_\theta(v' \mid o')\right], v' \sim \pi_\theta(\cdot \mid o'),$$

$$L_Q = \sum_{j=1}^{2} \mathbb{E}[(Q_j(o, v) - y_{\mathrm{SAC}})^2], L_\pi = \mathbb{E}[\alpha \log \pi_\theta(v \mid o) - \min_j Q_j(o, v)].$$

Here the entropy is the differential entropy of the squashed normalized action distribution, including the tanh Jacobian, rather than PPO's pre-tanh Gaussian entropy. Log standard deviations are bounded by $[-5, 1]$ and initialized at -0.7. The temperature starts at $\alpha = 0.01$ and is learned with target entropy -3 via

$$L_\alpha = -\mathbb{E}\left[\log \alpha \left(\mathrm{stopgrad} \log \pi_\theta(v \mid o) - 3\right)\right].$$

Its Adam learning rate is 3×10^{-4}, without decay, and $\alpha \in [10^{-5}, 0.2]$ is enforced after each update. The selected unfiltered and filtered SAC checkpoints have temperatures 1.1844×10^{-4} and 2.0036×10^{-4}, respectively. Entropy affects training only; all reported objectives use the original control criterion, and evaluation uses the transformed Gaussian mean.

DDPG instead uses one deterministic actor and one action-value critic per regime, with targets $y_{\mathrm{DDPG}} = \widetilde{r} + (1 - d)\bar{Q}(o', \bar{\mu}(o'))$, critic loss $\mathbb{E}[(Q - y_{\mathrm{DDPG}})^2]$, and actor loss $-\mathbb{E}[Q(o, \mu(o))]$. Both off-policy methods store 262,144 transitions, collect 32 initial vector steps of uniform actions, and then perform two minibatch updates per vector step, with batch size 4,096. DDPG adds clipped Gaussian exploration in normalized coordinates, with standard deviation decreasing linearly from 0.25 to 0.05 over the first 80% of training. SAC explores by sampling its actor. Polyak updates use $\bar{\theta} \leftarrow (1 - \tau)\bar{\theta} + \tau\theta$ after each update: SAC updates its two target critics, while DDPG updates its target critic and actor.

Implemented CT-DDPG extension. Following the value/advantage-rate approach of CT-DDPG Cheng et al. (2026 (v2); Bielecki et al. (2026), this baseline uses an integrated martingale residual and an advantage-rate network, rather than a discrete-time action-value critic. Let $q_\psi(o, u)$ be its raw rate network and $\mu_\theta(o)$ its actor. The centered rate is

$$A_{\psi,\theta}(o, u) = q_\psi(o, u) - q_\psi(o, \mu_\theta(o)).$$

For a replay segment of $\ell \in \{4, \ldots, 20\}$ decisions, sampled uniformly, the joint value/rate update minimizes the squared residual

$$\mathbb{E}\left[\left\{\bar{V}(o_{k+\ell}) - V(o_k) + \sum_{j=k}^{k+\ell-1}\left(\frac{r_j^{\mathrm{run}}}{50} - \Delta A_{\psi,\theta}(o_j, u_j)\right)\right\}^2\right]. \tag{17}$$

The reference actor is held fixed during this critic update. The reward r_j^{run} includes running, QV and switching contributions, with the terminal payoff removed once. Values enforce the known terminal condition through $V(o) = g(x)/50 + (T-t)v_\phi(o)$. Actor updates maximize the uncentered rate, using loss $-\Delta\mathbb{E}[q_\psi(o, \mu_\theta(o))]$. Only the value network has a Polyak target; there is no target actor or target rate network.

Returns under the common training budget. Table 5 reports the selected low-initial-intensity fits using the common evaluation protocol. Under the common environment-interaction budget, Hawkes-PPO achieves the highest evaluated objective among the RL methods from both initial regimes, and it approaches the oracle well. The weaker performance of the alternative methods may partly reflect optimization difficulties under the available training budget. In particular, CT-DDPG selects relatively early checkpoints and subsequently deteriorates, suggesting instability in the present implementation. However, these results do not establish nonconvergence of the alternative algorithms or an inherent superiority of PPO. The comparison concerns the selected fitted policies; the reported confidence intervals quantify evaluation uncertainty, rather than variability across training seeds.

Table 5: Objective values and 90% Monte Carlo confidence intervals. Each reported entry uses 4096 paths; a dash indicates an incompatible initial regime for a fixed-regime benchmark. Higher is better.

Method	AI start	Human start
Oracle (HJB/DGM)	-93.262 [-93.508, -93.015]	-94.230 [-94.468, -93.991]
Hawkes-PPO	-93.608 [-93.861, -93.354]	-94.676 [-94.915, -94.437]
Pure AI	-98.127 [-98.537, -97.718]	—
Pure Human (HJB/DGM)	—	-96.027 [-96.191, -95.864]
Other RL methods:		
PPO	-94.954 [-95.198, -94.710]	-97.219 [-97.407, -97.031]
Hawkes-SAC	-95.750 [-96.033, -95.467]	-96.579 [-96.853, -96.305]
SAC	-94.759 [-95.029, -94.489]	-96.079 [-96.347, -95.810]
Hawkes-DDPG	-95.810 [-96.128, -95.492]	-96.816 [-97.128, -96.504]
DDPG	-95.591 [-95.876, -95.306]	-96.783 [-97.065, -96.502]
Hawkes-CT-DDPG	-98.776 [-99.155, -98.397]	-101.670 [-101.902, -101.438]
CT-DDPG	-98.984 [-99.357, -98.611]	-101.811 [-102.106, -101.516]

B.7 Comparative statics, supervision mechanisms, and numerical sensitivity

This subsection provides additional supporting evidence for Section 4.5. The results are reported in Table 6.

Configuration design. Let A^0 denote the baseline excitation matrix, k_0 the baseline Human fixed operating-cost coefficient, and χ the baseline per-switch fee. We vary one factor at a time:

$$\begin{aligned} &\text{Excitation strength:} && A = s_A A^0, && s_A \in \{0.5, 1, 1.5\}, \\ &\text{Human fixed cost:} && k_0 \mapsto s_c k_0, && s_c \in \{0.75, 1, 1.25\}, \\ &\text{Per-switch fee:} && \chi \mapsto s_K \chi, && s_K \in \{0.5, 1, 2\}. \end{aligned}$$

The Human action and effort costs remain unchanged when k_0 varies. The quadratic switching-rate penalty remains unchanged when the per-switch fee varies. The baseline fee is $\chi = 0.2$, so the three per-switch charges are 0.1, 0.2, and 0.4; the intensity-penalty coefficient stays at 0.15. The baseline is shared across the three comparisons, giving seven distinct one-factor configurations. Crossing $s_A \in \{0.5, 1.5\}, s_c \in \{0.75, 1.25\}$ adds four configurations, for eleven in total. All unlisted physical and cost parameters retain their baseline values.

The intensity specification is

$$\lambda_t^j = \mu_j + (1 - \bar{e}_t)H_t^j, \qquad \mu = (0.35, 0.25),$$

where $\bar{e}_t$ is the effective mitigation effort defined in the main model. In particular, effort does not suppress the baseline intensity.

Actor and value learning rates start at 3×10^{-4} and 5×10^{-4}, remain constant through $3,932,160$ transitions, and decay according to the original schedule to one tenth by $11,796,480$ transitions. Independent fits run in parallel without changing these per-fit settings or budgets.

Evaluation, uncertainty and Monte Carlo. Each selected policy is evaluated on $4,096$ paths per initial regime in one batch, using seed 783000000 and the original numerical grid. Common random streams are used across configurations, and AI and Human starts are reported separately. The reused baseline reproduces all twenty-three original recorded per-path measurements exactly for both initial regimes.

For ordinary path means, pointwise 90% Monte Carlo intervals use

$$\bar{Y} \pm z_{0.95} \frac{s_Y}{\sqrt{n}}, \qquad z_{0.95} \simeq 1.64485.$$

Marginal burst probabilities use Wilson intervals. Endpoint contrasts use paired path differences, with paired influence functions for conditional-effort ratios. These intervals condition on the selected fitted policies. They exclude variability across training seeds, checkpoint-selection uncertainty, and discretization error, and are not adjusted for multiple comparisons.

Reported measurements. For path p, define Human duration and Human effort exposure as

$$T_{H,p} = \int_0^T \mathbf{1}_{\{I_t=0\}}\, dt, \qquad E_{H,p} = \int_0^T \mathbf{1}_{\{I_t=0\}} e_t\, dt.$$

Human time and conditional effort are estimated by

$$\widehat{h} = \frac{1}{nT} \sum_{p=1}^{n} T_{H,p}, \qquad \widehat{e}_H = \frac{\sum_{p=1}^{n} E_{H,p}}{\sum_{p=1}^{n} T_{H,p}}.$$

The table reports $100\widehat{h}$. Total Human effort exposure is $\sum_p E_{H,p}/(nT) = \widehat{h}\,\widehat{e}_H$, which differs from effort conditional on Human operation. The latter is a pooled ratio, not the average of path-specific conditional ratios. Its standard error uses the path influences

$$\psi_p = \frac{E_{H,p} - \widehat{e}_H T_{H,p}}{n^{-1} \sum_{q=1}^{n} T_{H,q}}, \qquad \widehat{\mathrm{se}}(\widehat{e}_H) = \frac{s_\psi}{\sqrt{n}}.$$

Incidents are $N_T^S + N_T^A$; realized switches count accepted regime changes. A burst occurs when at least three combined incidents fall within any closed sliding window of one time unit. Terminal quality is X_T, not $g(X_T)$. Path volatility is $\sqrt{[X]_T/T}$ averaged across paths, rather than the square root of mean quadratic variation.

Physical outcomes and the objective. At baseline costs, raising excitation from 0.5 to 1.5 reduces mean incidents from 5.595 to 5.127 for AI starts and from 5.581 to 4.918 for Human starts, while increasing terminal quality from 1.168 to 1.881 and from 1.189 to 2.023, respectively. These outcomes combine a changed incident environment with the response of a separately fitted policy.

Raising Human fixed cost from 0.75 to 1.25 at baseline excitation increases incidents from 5.054 to 6.202 for AI starts and from 4.926 to 6.130 for Human starts. Burst probabilities increase from 36.7% to 57.4% and from 33.5% to 56.5%, respectively. Conditional effort rises, but reduced Human duration causes total Human effort exposure to fall. Increasing the per-switch fee multiplier from 0.5 to 2 reduces mean switches from 1.344 to 0.907 for AI starts and from 1.924 to 0.659 for Human starts. Incident counts and burst probabilities increase for AI starts and decrease for Human starts, while terminal quality increases for both.

Reward and cost decomposition. We separate running and terminal project benefits, Human and AI operating costs, switching-rate and realized-switch costs, and the continuous and jump components of the quadratic-variation penalty. Their signed sum reconstructs the reported objective. This decomposition distinguishes improvements obtained through cost savings from improvements in physical project outcomes. In particular, an improved objective need not coincide with fewer incidents or greater terminal quality.

Supervision around incidents and takeovers. We align baseline trajectories around incidents, candidate cluster onsets, and actual AI-to-Human takeovers. An incident may belong to either channel. A candidate cluster onset is an incident followed by at least two further incidents within one time unit; this is a retrospective event definition, not a unique partition into maximal clusters.

We retain complete windows on $[-1, 1]$ within the recorded trajectory. Within each path, the earliest eligible anchor is retained and subsequent anchors must be at least two time units apart. Recorded left-constant microstep values are sampled on a 0.05-unit lag grid. Retained windows are first averaged within each path; eligible paths then receive equal weight. Conditional effort and the AI-only takeover rate use ratios of the corresponding path-averaged numerators and exposure denominators, with path-level delta-method uncertainty.

Risk during periods selected for supervision. At baseline excitation, higher Human fixed cost is associated with less Human time and higher conditional effort. During Human operation, total unmitigated excitation can be reconstructed as

$$H_t^S + H_t^A = \frac{\lambda_t^S + \lambda_t^A - 0.6}{1 - e_t}.$$

This diagnostic excludes microsteps with $e_t \geq 0.999$. At the high-cost setting, it excludes 23.02% and 16.86% of Human microsteps for AI and Human starts, compared with 0.18% and 0.12% at baseline cost. Among included Human microsteps, the mean excitation across cost multipliers $0.75, 1, 1.25$ is $0.230, 0.286, 0.306$ for AI starts and $0.188, 0.243, 0.235$ for Human starts. The latter pattern is nonmonotonic, and both comparisons are sensitive to the effort-dependent exclusions.

Joint changes in excitation and Human cost. The four crossed configurations test whether the one-factor patterns persist when risk and oversight cost change together. At both crossed cost levels, increasing excitation from 0.5 to 1.5 increases Human time and terminal quality and reduces incident counts and burst probability for both initial regimes. At both crossed excitation levels, increasing Human cost from 0.75 to 1.25 reduces Human time and increases incidents and burst probability. These Human-time and incident directions therefore persist in the crossed checks.

Effort, switching, and terminal quality have exceptions. Higher Human cost reduces conditional effort at low excitation but increases it at high excitation, for both starts. The AI-start effort response to stronger excitation is unresolved at low Human cost, with difference -0.008 and paired 90% interval $[-0.020, 0.003]$, but becomes positive at high Human cost, unlike the negative response at baseline cost. At high excitation, higher Human cost increases terminal quality by 0.133 for AI starts and 0.147 for Human starts, despite higher incident burden; at low excitation, quality decreases. Thus, a universal quality loss from more expensive supervision is not supported.

Endpoint comparisons also do not establish monotonicity. With expensive supervision, AI-start incident counts across excitation multipliers $0.5, 1, 1.5$ are $5.606, 6.202, 5.261$. At high excitation, Human-start quality across cost multipliers $0.75, 1, 1.25$ is $1.833, 2.023, 1.980$. Switching responses likewise depend on the starting regime and cost-risk combination.

Table 6: Supervision and project outcomes under the original numerical protocol. Means appear above conditional 90% evaluation-path Monte Carlo intervals.

Configuration	Human time (%)	Conditional effort	Switches / path	Incidents / path	Burst prob. (%)	Terminal quality	Objective
Panel A: AI initial regime							
Baseline	46.436 [45.694,47.179]	0.896 [0.893,0.898]	1.457 [1.430,1.484]	5.396 [5.335,5.458]	45.874 [44.597,47.157]	1.811 [1.787,1.835]	-93.608 [-93.861,-93.354]
Excitation 0.5	0.946 [0.857,1.034]	0.893 [0.882,0.904]	0.271 [0.251,0.291]	5.595 [5.523,5.666]	45.923 [44.645,47.206]	1.168 [1.143,1.193]	-85.962 [-86.204,-85.720]
Excitation 1.5	73.607 [73.045,74.168]	0.842 [0.838,0.846]	1.222 [1.204,1.241]	5.127 [5.068,5.187]	38.647 [37.404,39.906]	1.881 [1.856,1.905]	-95.849 [-96.069,-95.628]
Human cost 0.75	75.998 [75.463,76.533]	0.731 [0.726,0.735]	1.151 [1.135,1.166]	5.054 [4.996,5.113]	36.670 [35.441,37.917]	1.817 [1.793,1.841]	-89.677 [-89.868,-89.487]
Human cost 1.25	13.716 [13.237,14.196]	0.973 [0.972,0.974]	1.013 [0.981,1.045]	6.202 [6.126,6.277]	57.446 [56.171,58.712]	1.297 [1.271,1.323]	-96.334 [-96.669,-96.000]
Switch fee $\times 0.5$	45.382 [44.635,46.130]	0.880 [0.877,0.883]	1.344 [1.319,1.370]	5.440 [5.378,5.502]	47.021 [45.741,48.306]	1.899 [1.874,1.924]	-93.720 [-93.978,-93.462]
Switch fee $\times 2$	43.028 [42.200,43.856]	0.853 [0.849,0.857]	0.907 [0.890,0.924]	5.533 [5.469,5.597]	48.633 [47.350,49.918]	2.109 [2.081,2.137]	-94.360 [-94.627,-94.092]
A 0.5; cost 0.75	2.384 [2.172,2.595]	0.792 [0.781,0.802]	0.231 [0.214,0.248]	5.580 [5.508,5.651]	45.801 [44.523,47.084]	1.156 [1.131,1.181]	-85.839 [-86.078,-85.599]
A 0.5; cost 1.25	0.450 [0.402,0.499]	0.753 [0.734,0.772]	0.230 [0.211,0.249]	5.606 [5.534,5.678]	45.996 [44.718,47.279]	1.148 [1.123,1.173]	-85.996 [-86.240,-85.753]
A 1.5; cost 0.75	86.481 [86.165,86.797]	0.783 [0.779,0.788]	1.236 [1.217,1.255]	4.985 [4.927,5.044]	35.498 [34.278,36.737]	1.773 [1.749,1.797]	-90.862 [-91.045,-90.679]
A 1.5; cost 1.25	64.031 [63.394,64.668]	0.914 [0.912,0.916]	1.679 [1.650,1.709]	5.261 [5.201,5.321]	42.578 [41.313,43.853]	1.906 [1.881,1.930]	-100.143 [-100.409,-99.877]
Panel B: Human initial regime							
Baseline	58.424 [57.654,59.194]	0.809 [0.806,0.812]	2.013 [1.981,2.044]	5.283 [5.223,5.343]	43.384 [42.115,44.661]	1.879 [1.854,1.903]	-94.676 [-94.915,-94.437]
Excitation 0.5	5.334 [5.190,5.478]	0.516 [0.507,0.525]	1.422 [1.397,1.447]	5.581 [5.509,5.652]	45.654 [44.377,46.937]	1.189 [1.164,1.214]	-87.328 [-87.570,-87.085]
Excitation 1.5	98.061 [97.855,98.266]	0.696 [0.691,0.700]	0.222 [0.204,0.240]	4.918 [4.861,4.975]	33.813 [32.609,35.040]	2.023 [1.999,2.047]	-96.849 [-97.017,-96.682]
Human cost 0.75	99.594 [99.510,99.679]	0.623 [0.619,0.628]	0.070 [0.060,0.081]	4.926 [4.868,4.983]	33.521 [32.318,34.744]	1.934 [1.911,1.958]	-89.746 [-89.910,-89.581]
Human cost 1.25	19.065 [18.554,19.577]	0.862 [0.858,0.866]	2.088 [2.053,2.122]	6.130 [6.056,6.205]	56.470 [55.192,57.739]	1.333 [1.307,1.359]	-97.731 [-98.061,-97.401]
Switch fee $\times 0.5$	58.076 [57.288,58.865]	0.793 [0.790,0.796]	1.924 [1.894,1.954]	5.310 [5.250,5.370]	44.043 [42.771,45.322]	1.965 [1.940,1.989]	-94.649 [-94.892,-94.407]
Switch fee $\times 2$	82.461 [81.736,83.186]	0.693 [0.688,0.697]	0.659 [0.636,0.683]	5.087 [5.029,5.145]	38.330 [37.089,39.587]	2.513 [2.486,2.540]	-96.174 [-96.380,-95.968]
A 0.5; cost 0.75	13.579 [13.168,13.989]	0.607 [0.601,0.614]	1.277 [1.258,1.295]	5.506 [5.437,5.576]	44.580 [43.307,45.861]	1.187 [1.162,1.212]	-86.750 [-86.983,-86.518]
A 0.5; cost 1.25	4.214 [4.109,4.319]	0.417 [0.411,0.424]	1.409 [1.384,1.433]	5.597 [5.525,5.669]	45.850 [44.572,47.133]	1.154 [1.129,1.179]	-87.649 [-87.895,-87.404]
A 1.5; cost 0.75	99.572 [99.499,99.644]	0.708 [0.704,0.713]	0.133 [0.118,0.147]	4.896 [4.839,4.953]	33.521 [32.318,34.744]	1.833 [1.809,1.856]	-90.453 [-90.617,-90.289]
A 1.5; cost 1.25	74.429 [73.792,75.066]	0.826 [0.823,0.829]	2.014 [1.978,2.050]	5.160 [5.102,5.218]	40.234 [38.981,41.501]	1.980 [1.955,2.004]	-101.517 [-101.763,-101.271]