Risk-sensitive decision-making in security operations center alert escalation under resource constraints*

Viacheslav Kovtun[1,*,†], Artem Huralnyk[2,†], Lyudmila Husak[2,*,†] and Oleh Kovaliuk[1,†]

[1] Vinnytsia National Technical University, Khmelnytske shose, 95, Vinnytsia, 21021, Ukraine
[2] Vinnytsia trade economics university, Soborna, 87, Vinnytsia, 21050, Ukraine

Abstract

The article formalises, for the first time, the process of alert escalation in a Security Operations Center (SOC) as a risk-sensitive control problem in a partially observable stochastic environment with non-linear queueing dynamics. An integrated model based on a partially observable Markov process is proposed. It combines posterior estimation of latent attack regimes, backlog accumulation, and an entropic risk-sensitive optimisation criterion with explicit constraints on the utilisation of deep analysis resources and the permissible attack miss rate. In contrast to risk-neutral reinforcement learning approaches, the proposed formulation establishes a structural linkage between inference and the control of tail risks, which become critical as the system approaches its stability boundary. Experimental validation on CIC-IDS2017 and UNSW-NB15 demonstrates that the system enters a quasi-critical load region for 12.4% and 14.1% of the time, respectively. The transition from a subcritical to a quasi-critical regime increases the stationary queue accumulation level by factors of 1.41 and 1.34, while the local logarithmic elasticity reaches 2.556. The results indicate that risk-sensitive optimisation stabilises the tail characteristics of the backlog without exceeding the resource budget, thereby providing a formalised foundation for escalation management in a SOC.

Keywords

security Operations Center; alert escalation; risk-sensitive decision-making; partially observable Markov decision process; entropic risk measure; queue dynamics; reinforcement learning; cyber threat detection [1]

1. Introduction

The operation of contemporary Security Operations Centers (SOC) takes place under conditions of rapidly increasing volume and variability of alert streams generated by SIEM, EDR, IDS/IPS, and related monitoring tools [1, 2]. According to a Forrester study (covering more than 300 SOCs), a typical SecOps team processes approximately 450 alerts per hour, that is, around 11,000 notifications per day, while more than one quarter of them remain unreviewed [3]. Data from Vectra's "State of Threat Detection" report [4], cited by IBM, indicate an average workload of 4,484 alerts per day and show that 67% of notifications are ignored due to a high proportion of false positives and staff overload. Analytical materials from Axur confirm a high noise level and a substantial share of signals that are not investigated.

Surveys conducted by the Ponemon Institute [5] indicate that more than 70% of SOC teams face situations in which false positives account for over half of all alerts, while 62% of specialists acknowledge missing significant incidents due to overload. Cisco reports [6] show that nearly half of organisations process critical alerts with delays exceeding two hours as a result of backlog accumulation. Research by Verizon [7] demonstrates that peak SOC workloads frequently coincide with phases of actual attacks. Technical analyses of complex APT campaigns further note that early signals were ignored during the first 24–48 hours due to the phenomenon of alert fatigue.

* *IntelITSIS-2026: 7th International Workshop on Intelligent Information Technologies & Systems of Information Security, July 3, 2026, Khmelnytskyi, Ukraine*
[1,*] Corresponding author.
[†] These authors contributed equally.
kovtun_v_v@vntu.edu.ua (V. Kovtun); artem.guralnyk@gmail.com (A. Huralnyk); l.husak@vtei.edu.ua (L. Husak); oleh.kovalyuk@vntu.edu.ua (O. Kovaliuk)
0000-0002-7624-7072 (V. Kovtun); 0000-0003-1977-6338 (A. Huralnyk); 0000-0002-0022-9644 (L. Husak); 0000-0002-0718-010X (O. Kovaliuk)

These examples demonstrate that the problem of alert escalation cannot be reduced to classification or filtering. It acquires the character of a stochastic control problem involving a limited resource of analytical attention under uncertainty regarding the true nature of incidents. Queue accumulation, as the arrival intensity approaches the SOC processing capacity, leads to a disproportionate increase in operational risk and raises the probability of missing genuine attacks. Accordingly, there is a need to formalise a risk-oriented escalation policy capable of integrating threat uncertainty, backlog dynamics, and resource constraints within a unified optimal control framework.

2. Related works

The automation of alert escalation in SOCs is evolving at the intersection of three SOTA trajectories: data-driven algorithmic prioritisation and triage, reinforcement learning with human feedback, and risk-sensitive and constrained reinforcement learning. In applied research, a substantial body of work focuses on training models that replicate analysts' decisions and automate primary triage. In particular, AACT [8] formulates triage as a supervised learning task, using historical analyst actions to automate alert classification and prioritisation. Another prominent direction concerns learning "to defer" with human feedback [9, 10], where the model adaptively determines which alerts should be escalated for deeper analysis in order to reduce analyst fatigue and improve response timeliness. These studies are the closest to the SOC domain in terms of triage and prioritisation formulation. However, they predominantly optimise local quality metrics (accuracy, priority assignment, alignment with human decisions) and do not explicitly model a multi-channel service structure of escalation. Within such a structure, decisions compete for a limited number of analytical channels and generate a queue. Under these conditions, the control of tail backlog metrics in peak regimes becomes critical..

In parallel, constrained reinforcement learning has developed, in which the decision-making process is described as a constrained Markov decision process with optimisation under constraints [11]. Study [12] demonstrates that a Lagrangian formulation enables the incorporation of safety and resource constraints into policy learning and stabilises agent behaviour. At the same time, the typical form of constraints in the CMDP class primarily controls average values, which is fundamentally insufficient for SOC escalation. In this context, it is not average indicators but the tail risks of queue accumulation that are critical. These risks increase sharply as the alert arrival intensity approaches the aggregate service capacity of analysts. This aspect is reinforced by established results in queueing theory [13]. In multi-channel systems of the M/M/k type operating under heavy-traffic conditions, percentile characteristics of queue length and delay increase disproportionately, and operational risk exhibits phase-transition behaviour as the stability boundary is approached. For this reason, risk-neutral reinforcement learning, which maximises the expected return, is methodologically misaligned with the SOC context, where rare yet catastrophic events determine response effectiveness. The literature has introduced queue-aware modifications of RL, in which queue stability is incorporated as a constraint or regularisation term [14]. In particular, [15] proposes deep RL with an explicit queue stability constraint and analytical queue modelling based on M/M/c for the joint optimisation of performance and stability. However, such studies typically belong to networking or communications domains and do not account for SOC-specific characteristics, including partial observability of the true nature of incidents and the need to balance the risk of missed attacks against the resource budget allocated to deep analysis.

Risk-sensitive reinforcement learning addresses tail scenarios through the optimisation of risk measures; however, it predominantly considers risk with respect to cumulative return. Noorani et al. [16] propose an exponential criterion that induces an entropic risk measure and enhances policy robustness to adverse environmental realisations. Théate and Ernst [17] examine risk-sensitive distributional RL, in which risk is modelled via a utility transformation. Both studies constitute the closest analogues in terms of the principle of risk-sensitive optimisation. Nevertheless, their risk object is defined over trajectory-level return rather than over the structural tails of the queue length (backlog) distribution in a multi-channel service process with limited analyst resources.

Consequently, the problem of coherently integrating partial observability of the latent incident state, an explicit multi-channel service structure of escalation, the non-linear heavy-traffic overload regime, and a risk-oriented criterion specifically aimed at stabilising the tail characteristics of the backlog while controlling the risk of missed attacks remains unresolved.

A critical analysis of contemporary approaches thus demonstrates that SOC-oriented triage models do not account for stochastic queueing dynamics and resource competition. Constrained reinforcement learning predominantly controls average constraints without managing tail characteristics. Risk-sensitive approaches, in turn, model risk at the level of cumulative return without a structural linkage to a multi-channel service system and the partial observability of incidents. As a result, there is no integrated formalisation of alert escalation as a stochastic control problem under conditions of limited analytical channel capacity and uncertainty regarding the true nature of events. In a real SOC, it is precisely the tail characteristics of queue length and escalation delay, which exhibit non-linear growth as the arrival intensity approaches the aggregate service capacity, that determine operational stability and the risk of missing critical attacks. This substantiates the need to construct an integrated Markovian model in which latent state inference, constrained resource control, and risk-sensitive regulation of overload tail effects are formulated as a single optimisation problem.

3. Research objectives and contributions

The object of the study is the decision-making process concerning alert escalation in a SOC, considered as a partially observable multi-channel stochastic service process with discrete decision-making and a limited number of analytical channels.

The subject of the study comprises the regularities governing backlog dynamics in quasi-critical load regimes and the mechanisms of risk-sensitive control of queue tail characteristics under conditions of controlled risk of missed attacks.

The aim of the study is to minimise high-percentile characteristics of backlog length in regimes approaching the system's stability boundary, while simultaneously constraining the risk of missed incidents and adhering to the resource budget allocated to deep analysis.

To achieve this objective, it is necessary:

- to formalise the escalation process as a partially observable Markov process with explicit consideration of the multi-channel service structure;
- to provide an analytical justification for the selection of a risk-oriented criterion sensitive to tail characteristics and to integrate analytical channel resource constraints into a unified optimisation framework;
- to formalise a policy learning procedure with guaranteed computational feasibility;
- to conduct experimental validation of the model on real-world cybersecurity datasets, including an analysis of system behaviour under heavy-traffic regimes.

The main contribution of the study lies in the development of a novel integrated formalisation of the alert escalation process, which for the first time combines partial observability of the latent incident state, multi-channel queueing dynamics, and a risk-sensitive criterion oriented towards stabilising the tail characteristics of the backlog. In contrast to constrained reinforcement learning, which predominantly regulates average constraints, and to distributional or entropic RL, which models risk at the level of aggregate return, the proposed approach directly targets the structural tails of the service process. The study further demonstrates the quantitative non-linearity of tail metric growth as the service capacity boundary is approached and establishes the possibility of their controlled reduction without a disproportionate increase in the risk of missed attacks.

Section 2 presents the formal stochastic model of the escalation process and provides a justification for the risk-sensitive optimisation criterion. Section 3 describes the algorithmic implementation and the experimental evaluation procedure on the CIC-IDS2017 and UNSW-NB15

datasets. Section 4 reports the simulation results, including an analysis of system behaviour in quasi-critical load regimes. Section 5 formulates the conclusions and outlines directions for further research.

4. Models and methods

4.1. Research statement

The focus of this study is a discrete-time control process governing alert escalation in an information security monitoring and incident response centre, in which decisions are taken at time instants $t=0,1,2,\ldots$. Each time step corresponds to an operational interval of alert stream processing and decision-making regarding their subsequent analysis.

The true nature of alerts is described by a hidden threat regime $z_t \in Z=\{0,1,2\}$, interpreted as the latent security state of the infrastructure. In particular, 0 corresponds to normal functioning (benign), 1 to the reconnaissance or preparatory stage, and 2 to the active attack phase. The variable z_t is fundamentally unobservable at the stage of primary triage.

The analytical information available to the system at step t is represented by an observation vector $o_t \in R^d$, where d denotes the dimensionality of the feature space. This vector is constructed through the aggregation of alert features, event correlation results, and telemetry typical of SIEM/SOAR platforms. The sequence of observations and adopted decisions forms the information history h_t.

The operational state of the SOC is characterised by the alert queue length q_t, which reflects the current system load and is directly associated with the risks of response delay and missed critical incidents. The limitation of deep analysis resources, including sandbox analysis, DPI, or forensic procedures, is modelled by a variable k_t, which captures the availability of the corresponding analytical mechanisms at time t.

At each time step, the system selects a control action $a_t \in A$, corresponding to typical SOC decisions, including superficial triage, escalation to deep analysis, or deferral of the decision in order to accumulate additional context. Since the true threat regime is unobservable, control is performed on the basis of the conditional probability distribution b_t, which is interpreted as the system's belief state regarding the current security regime.

The problem is formulated on an infinite time horizon, reflecting the continuous operation of the SOC and the assumption of a stationary control policy. At each step t, alerts arrive stochastically, the information state (o_t, b_t) is updated, and an action $a_t \in A$ is selected, which determines the controlled transition of the system to the state at step $t+1$. The belief state b_t is updated according to the Bayesian principle and serves as a sufficient statistic for decision-making under partial observability, thereby reducing the process to a controlled stochastic dynamics of an augmented state.

The optimisation problem is based on the minimisation of the discounted functional of cumulative losses

$$J^{\pi}(s_0) = E^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right], \tag{1}$$

where s_t denotes the augmented system state, π the control policy, $c(s_t, a_t)$ the instantaneous cost function that aggregates operational expenses, the risk of missing critical incidents, and service degradation, and $\gamma \in (0,1)$ the discount factor determining the relative weight of future losses.

The stochastic dynamics of the latent threat regime and the statistical structure of observations are formalised as follows. The latent threat regime z_t is modelled as a time-homogeneous first-order Markov process with transition probability matrix $P=(P_{ij})$, satisfying the conditions $P_{ij} \geq 0$ and $\sum_j P_{ij} = 1$. The process dynamics are defined by

$$P\left(z_{t+1}=j\middle|z_t=i\right)=P_{ij}, \tag{2}$$

which reflects the exogenous nature of adversarial activity and the assumption that SOC decisions do not directly influence the evolution of the true threat regime. Time homogeneity implies that the matrix P does not depend on t, corresponding to a stationary model of the external environment.

The observation vector o_t is generated according to a conditional distribution that depends on the current latent regime and the previous control action. Formally, the observation model is defined as

$$P\left(z_t, a_{t-1}\right), \tag{3}$$

which reflects a mechanism of controlled observation: the selection of deep analysis or the deferral of a decision alters the statistical characteristics of the acquired data. Conditional independence of o_t from previous states and observations is assumed, given fixed z_t and a_{t-1}, in accordance with the structure of a partially observable Markov model.

Under partial observability of the latent regime z_t, control is performed in the space of information states. Let the information history be $h_t=\left(o_0, a_0, \ldots, o_{t-1}, a_{t-1}, o_t\right)$. The belief state is defined as

$$b_t(z)=P^{\pi}\left(z_t=z\middle|h_t\right) b_t \in \Delta(Z), , \tag{4}$$

where $\Delta(Z)=\left\{b \in R^{|Z|}\middle| b(z)\geq 0, \sum_z b(z)=1\right\}$. The probability P^{π} in (4) is a measure on the trajectory space induced by the Markovian dynamics (2), the observation model (3), and the policy π. The process b_t is Markovian in the space $\Delta(Z)$ and constitutes a sufficient statistic for decision-making, thereby reducing the partially observable model to a controlled Markovian dynamics in the information space.

4.2. Constrained risk-sensitive partially observable Markov decision process for queue-aware alert escalation in SOC

Since the observation model (3) depends on the action, control influences the informativeness of future data through modification of the joint distribution $\left(z_{t+1}, o_{t+1}\right)$, induced by the predictive distribution $b_t P$ and the conditional observation distribution. The conditional mutual information is defined as

$$I\left(z_{t+1}; o_{t+1}\middle| b_t, a_t\right)=E^{\pi}\left[\log \frac{P\left(o_{t+1}\middle| z_{t+1}, a_t\right)}{P\left(o_{t+1}\middle| b_t, a_t\right)}\middle| b_t, a_t\right], \tag{5}$$

where $P\left(o_{t+1}\middle| b_t, a_t\right)=\sum_{z' \in Z} P\left(o_{t+1}\middle| z', a_t\right) \sum_{z \in Z} P_{zz'} b_t(z)$, and $P_{zz'}=P\left(z_{t+1}=z'\middle| z_t=z\right)$ is defined in (2).

The defer action is structurally more informative if

$$I\left(z_{t+1}; o_{t+1}\middle| b_t, a_t=\text{defer}\right)>I\left(z_{t+1}; o_{t+1}\middle| b_t, a_t \neq \text{defer}\right), \tag{6}$$

which is equivalent to the expected reduction in the entropy of the belief state $H(b)=-\sum_{z \in Z} b(z) \log b(z)$, namely

$$E^{\pi}\left[H\left(b_{t+1}\right)\middle| b_t, a_t=\text{defer}\right]<E^{\pi}\left[H\left(b_{t+1}\right)\middle| b_t, a_t \neq \text{defer}\right]. \tag{7}$$

Relations (6) and (7) formalise the controlled trade-off between immediate escalation and the accumulation of statistically relevant information.

In the analytical formulation, the belief is defined by relation (4) and, after selecting action a_t and receiving observation o_{t+1}, is updated by the operator

$$b_{t+1}(z') = \frac{P(o_{t+1}|z', a_t) \sum_{z \in Z} P_{zz'} b_t(z)}{\sum_{z \in Z} P(o_{t+1}|z', a_t) \sum_{z \in Z} P_{z\bar{z}} b_t(z)}, \quad (8)$$

where the dummy summation index $\bar{z} \in Z$ introduced in the denominator denotes the transition kernel $P_{z\bar{z}} = P(z_{t+1} = \bar{z} | z_t = z)$, equivalent to $P_{zz'}$. In numerical implementation, the Bayesian update operator (8) may be approximated by a parameterised recurrent model $b_t \approx g_\phi(b_{t-1}, o_t)$, where g_ϕ approximates operator (8) in the space $\Delta(Z)$. The parameterisation is employed solely for computational realisation and does not alter the analytical structure of the problem.

The stochastic dynamics of SOC operational load are now formalised. Alert arrivals are modelled as a conditionally independent Poisson process $A_t \sim \text{Pois}(\lambda(z_t))$ with intensity $\lambda : Z \to R_{+¿¿}$, which depends on the latent threat regime. This process is exogenous and does not depend on the action a_t. In the information state space, the conditional expectation of arrivals takes the form $E[A_t | b_t] = \sum_{z \in Z} \lambda(z) b_t(z)$, linking the latent intensity $\lambda(z_t)$ to the information state b_t defined in (4).

Let $q_t \in \{0,1,\ldots,M\}$ denote the queue length, where $M < \infty$ is the buffer capacity, and let $k_t \in \{0,1\}$ be an indicator of deep analysis resource availability. The resource dynamics are specified by a controlled Markov kernel

$$P(k_{t+1} = k' | k_t, a_t) = R_{kk'}(a_t), \quad (9)$$

which formalises the operational regimes of deep pipeline availability. Here k denotes the current state of the deep analysis resource (available $k = 1$) or unavailable $k = 0$)), while k' denotes its state at the next time step. The indices $k, k' \in \{0,1\}$ define the elements of the controlled transition matrix $R_{kk'}(a_t)$.

The nominal service capacity is defined as $\tilde{S}_t(a_t, k_t) = \begin{cases} S_t^{sh} \ \forall\, a_t = \text{shallow}, \\ S_t^{dp} \ \forall\, a_t = \text{deep}, k_t = 1, \\ 0 \ \forall\, a_t = \text{defer}, \end{cases}$, where S_t^{sh} and S_t^{dp} are random variables with known distributions, conditionally independent over time given fixed (a_t, k_t), with $E[S_t^{dp}] < E[S_t^{sh}]$. The actual number of processed alerts is bounded by the available workload $S_t(a_t, k_t) = min(\tilde{S}_t(a_t, k_t), q_t + A_t)$. The queue evolution is described by the balance equation:

$$q_{t+1} = min\{M, [q_t + A_t - S_t(a_t, k_t)]^+\}, \quad (10)$$

The quantities A_t, $\tilde{S}_t$, k_{t+1} and o_{t+1} are assumed to be conditionally independent given fixed (s_t, a_t), ensuring the existence of a controlled Markov kernel $P(s_{t+1} | s_t, a_t)$. The boundedness of $q_t \le M$ guarantees the compactness of the discrete component of the state space and, together with the boundedness assumption for instantaneous losses, ensures the well-posedness of the subsequent risk-sensitive functional.

The construction of the controlled stochastic model is completed by explicitly specifying the state space and the space of admissible actions. The augmented system state is defined as $s_t = (q_t, k_t, b_t) \in S$, where $q_t \in \{0,\ldots,M\}$ characterises the alert queue length, $k_t \in \{0,1\}$ represents the availability state of the deep analysis resource, and $b_t \in \Delta(Z)$ is the belief state that aggregates all available information about the latent threat regime. The control action space is

given by the finite set $A=\{\text{shallow},\text{deep},\text{defer}\}$, with the admissible set depending on the resource state: when $k_t=0$, the deep action is unavailable. The control policy is defined as a stochastic mapping $\pi(a)=S\to\Delta(A)$ subject to the constraint $\pi(a|s)=0\ \forall\, a\notin A(s)$, which ensures the mathematical well-posedness of the risk-sensitive optimal control formulation.

To formalise the process economics of alert escalation in a SOC, the one-step loss function is specified as a non-linear functional of operational workload, resource expenditure, and the a posteriori attack risk. For the extended state $s_t=(q_t,k_t,b_t)$ the instantaneous cost is defined as

$$c(s_t,a_t)=c_q q_t^{\alpha}+c_{dp}\vec{1}\{a_t=\text{deep}\}+c_{df}\vec{1}\{a=\text{de fer}\}+c_{miss}\Phi\left(\sum_{z\in Z_{atk}}b_t(z)\right)\vec{1}(a_t\neq\text{deep}), \qquad (11)$$

where $\alpha>1$ models the non-linear service degradation under SOC overload, $\Phi:[0,1]\to R_+$ is an increasing convex risk function, $Z_{atk}\subseteq Z$ denotes the set of active attack regimes, and the quantity $\sum_{z\in Z_{atk}}b_t(z)$ represents the a posteriori probability of an active threat, thereby ensuring the dependence of the economic penalty on the informational state. In the case of incomplete effectiveness of deep analysis, the detection parameter $\rho\in 0,1$ is introduced into function (11), in which case the corresponding term takes the form $c_{miss}\Phi\left(1-\rho\vec{1}\{a_t=deep\}\right)\sum_{z\in Z_{atk}}b_t(z)$.

The one-step loss function (11) specifies the local economics of alert escalation within a SOC. However, strategic control under Markovian dynamics requires intertemporal aggregation of losses with due consideration of the risk of tail scenarios associated with active attacks. Generalising the risk-neutral functional (1), an exponential entropic criterion is introduced: for the initial state $s\in S$:

$$J_{\eta}^{\pi}=\frac{1}{n}\log E_s^{\pi}\left[\exp\left(\eta\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right)\right], \qquad (12)$$

where $c(s_t,a_t)$ is defined in (11). The parameter $\eta>0$ determines the degree of risk aversion and scales the exponential amplification of tail losses, while the expectation is taken with respect to the measure P_s^{π} on the trajectory space induced by the initial state s, the policy π, and the controlled Markov transition kernel.

The limiting transition $\lim_{\eta\to 0}J_{\eta}^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right]$ recovers (1), thereby ensuring consistency between the risk-sensitive and risk-neutral formulations. The boundedness of $q_t\le M$ and the finiteness of the coefficients in (11) guarantee $sup._{s,a}c(s,a)<\infty$, which ensures the well-posedness of the exponential aggregation. Criterion (12) is dynamically consistent and leads to a non-linear Bellman equation in the space S.

Let $V(s)=\inf_{\pi\in\Pi}J_{\eta}^{\pi}(s)$, where Π denotes the class of Markov policies on S. Owing to the reduction of the partially observed model to a controlled Markov process in the belief space under (4), (8), the extended state $s_t=(q_t,k_t,b_t)$ is Markovian, and the belief b_t constitutes a sufficient statistic of the history h_t. Therefore, the optimisation may, without loss of generality, be restricted to the class of (stationary) Markov policies $\pi(a|s)$.

For the entropic criterion (12), the value function satisfies the non-linear Bellman equation

$$V(s)=\min_{a\in A}\left\{c(s,a)+\frac{\gamma}{\eta}\log E\left[\exp\left(\eta V(s')\right)|s,a\right]\right\}, \qquad (13)$$

where the expectation is taken with respect to the controlled Markov kernel $P(s'|s,a)$ induced by the dynamics. Under conditions $\gamma\in(0,1)$ and ¿$_{s,a}c(s,a)<\infty$, the Bellman operator is a contraction in the space of bounded functions, which guarantees the existence and uniqueness of

the solution. In the limit $\eta \to 0$, equation (13) reduces to the linear Bellman equation for functional (1), thereby ensuring consistency between the risk-sensitive and risk-neutral formulations.

In addition to the economic aggregation of losses in (12), the operational requirements of the SOC are formalised as constraints on admissible policies. A class of Markov policies Π is considered on the space S, and the set of admissible policies is defined as

$$\Pi_{adm} = \left\{ \pi \in \Pi : J_{deep}(\pi) \le B, J_{miss}(\pi) \le \varepsilon \right\}, \tag{14}$$

where the resource constraint is specified by the functional $J_{deep}(\pi) = E_s^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \vec{1}\{a_t = \text{deep}\}\right]$, and the constraint on missing an active attack is formulated in the information space via the belief state $J_{miss}(\pi) = E_s^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \left(\sum_{z \in Z_{atk}} b_t(z)\right) \vec{1}\{a_t \neq \text{deep}\}\right]$.

The optimisation problem takes the form

$$\inf_{\pi \in \Pi_{adm}} J_{\eta}^{\pi}(s), \tag{15}$$

which defines a risk-sensitive controlled Markov decision problem with constraints (CMDP). Under conditions of compactness of the discrete component S, finiteness of A, and boundedness of the instantaneous losses, an optimal stationary policy exists. The equivalent Lagrangian formulation makes it possible to reduce (15) to an unconstrained optimisation problem with appropriate multipliers, while preserving the structure of the Bellman equation (13).

In order to stabilise numerical optimisation and to control deviations from operationally acceptable behaviour, a KL-regularised version of problem (15) is formalised. Let $\pi_0(\cdot|s) \in \Delta(A)$ denote an interpreted baseline policy reflecting the standard SOC playbook and belonging to Π_{adm}. The optimisation problem then takes the form

$$\inf_{\pi \in \Pi_{adm}} \left\{ J_{\eta}^{\pi}(s) + \tau E_s^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \mathrm{KL}\left(\pi(\cdot|s_t) \middle\| \pi_0(\cdot|s_t)\right)\right] \right\}, \tag{16}$$

where τ is the parameter of local entropic regularisation, and J_{η}^{π} is defined in (12). The regularisation preserves the entropic aggregation of risk (governed by η), while modifying only the local minimisation operator.

The corresponding Bellman operator takes the form

$$V(s) = \min_{\pi(\cdot|s) \in \Delta(A)} \left\{ \sum_{a \in A} \pi(a|s) \left(c(s,a) + \frac{\gamma}{\eta} \log E\left[\exp\left(\eta V(s')\right)|s,a\right] \right) + \tau \mathrm{KL}\left(\pi(\cdot|s_t) \middle\| \pi_0(\cdot|s_t)\right) \right\}, \tag{17}$$

where $s' = (q_{t+1}, k_{t+1}, b_{t+1})$ denotes the next extended state of the system, the distribution of which is determined by the controlled Markov transition kernel.

The functional in (17) is strictly convex in $\pi_0(\cdot|s)$, which guarantees the existence of a unique stochastic minimiser. The optimal policy takes the form of an exponential tilt relative to the baseline:

$$\pi^*(a|s) = \frac{\pi_0(a|s) \exp\left(\dfrac{-Q_{\eta}(s,a)}{\tau}\right)}{\sum\limits_{a' \in A} \pi_0(a'|s) \exp\left(\dfrac{-Q_{\eta}(s,a')}{\tau}\right)}, \tag{18}$$

where $Q_{\eta}(s,a) = c(s,a) + \frac{\gamma}{\eta} \log E\left[\exp\left(\eta V(s')\right)|s,a\right]$, and the symbol a' is used as the summation variable over the action set A. The parameter η determines intertemporal risk aversion, whereas τ controls the local entropy of the policy. Accordingly, the limits $\eta \to 0$ and $\tau \to 0$

correspond to risk-neutral aggregation and deterministic control, respectively. Relations (16)–(18) complete the construction of the risk-sensitive CMDP model with entropic stabilisation, integrating constraints (14)–(15) with the non-linear operator (13).

The integration of belief dynamics (4), (8), the alert arrival model (10)–(11), queue and resource dynamics (9)–(10), the non-linear loss structure (11), and entropic optimisation (12), (13) yields a formalised model for controlling alert escalation in a SOC. The extended state $s_t=\left(q_t,k_t,b_t\right)$ simultaneously captures the operational workload (queue), the availability of deep-analysis resources, and the a posteriori assessment of an active threat. This enables the decision-making process to be described as a controlled stochastic mechanism of escalation. In this context, the action defer models the postponement of escalation in order to accumulate additional analytical context. In accordance with (6)–(7), it reduces informational uncertainty regarding the threat regime, yet, through (10), it increases the risk of SOC overload. The choice between deep and defer formally realises a trade-off between the marginal value of additional information and the marginal cost of delayed response, which corresponds to the structure of optimal stopping in a queued system.

For positive coefficients in (11), the value function $V\left(s\right)$ is non-decreasing in the queue length q_t and in the a posteriori probability of an active attack $\sum_{z\in Z_{atk}} b_t\left(z\right)$. This induces a partition of the state space $\left(q_t,b_t\right)$ into regions in which it is optimal to perform superficial triage, immediate escalation, or to defer the decision. As the risk-aversion parameter η in (12) increases, the region of immediate deep analysis expands for high values of belief-based risk, reflecting a strengthened response to potentially catastrophic incidents.

5. Results and discussion

This section presents the results of the empirical verification of the formal model defined in (1)–(18). The experimental component examines the implications of the model's key structural elements: the state transition (2), queue dynamics (10), belief updating (4), (8), informational gain (6)–(7), the entropic functional (12), the Bellman equation (13), constraints (14)–(15), KL regularisation (16)–(18), and the stability of the cost function (11).

All methods were implemented within an identical environment. The same cost function (11) with fixed baseline coefficients (\alpha_i) is optimised throughout. The primary evaluation criterion is the entropic functional (12). Constraints (14)–(15) are applied uniformly across all approaches. KL regularisation (16)–(18) is employed with the same τ in the respective comparisons. The time horizon comprises $T=200$ steps. The arrival process is generated from the test data segment and is applied identically to all policies. The data are partitioned chronologically in a 70% train / 30% test ratio without shuffling.

Control is implemented as a model-free policy gradient algorithm, Proximal Policy Optimisation (PPO) [18], with a clipping parameter of 0.2. The discount factor is set to $\gamma=0.99$, learning rate = 3×10^{-4}, optimiser Adam, batch size = 4096 steps, and 4 update epochs per iteration. Training proceeds for 1.2×10^{-6} steps or until stabilisation of the mean return over 50 consecutive episodes with a deviation of less than 1%. Parameter initialisation is performed using a fixed list of 15 random seeds; the same set of seeds is employed for all compared methods. Confidence intervals for the temporal characteristics of the arrival process and the derived statistics are computed across seeds, ensuring a correct assessment of inter-run variability without disrupting the temporal structure of the series.

Bootstrap 95% intervals for aggregated metrics were computed using the percentile method with 1000 resamples across episodes of the test segment. Paired t-tests were applied for pairwise comparisons, and effect sizes are reported using Cohen's d. All experiments were conducted on an Intel Core Ultra 7 CPU (8 performance cores, 16 threads), with the average duration of a single training run amounting to 2.8 hours.

For the empirical verification of the state transition (2) and queue dynamics (10), two publicly available datasets were employed, independent in terms of traffic source and attack structure. The CIC-IDS2017 ([19]: https://www.kaggle.com/datasets/chethuhn/network-intrusion-dataset, hereafter Dataset A) reflects realistic temporally heterogeneous activity with labelled phases of

benign and attack traffic. The UNSW-NB15 ([20, 21]: https://www.kaggle.com/datasets/mrwellsdavid/unsw-nb15, hereafter Dataset B) is characterised by a multi-stage attack structure and multiple threat classes. Events were deduplicated, numerical features were normalised using z-scores, and the data were aggregated into discrete time steps of length $\Delta t = 1$ seconds. The arrival process λ_t is defined as the number of alerts per step, measured in alerts per 1s. Latent regimes z_t are reconstructed on the basis of attack/benign labels and clustering of attack subtypes.

The queue evolution q_t is modelled in accordance with (10) under a fixed number of servers $k = 8$. The average service rate is defined as $\mu = \frac{1}{E\left[t_{service}\right]}$, where $t_{service}$ is estimated on the train segment; a value of $\mu = 6.5$ events per step was obtained. This value is fixed during testing and is identical for both datasets. The nominal system capacity amounts to $k\mu = 52$.

For the quantitative verification of the empirical structure of the arrival process and its consistency with the state transition (2) and queue dynamics (10), Figure 1 presents a visualisation of the temporal intensity $\lambda(t)$ and the corresponding moving variance. The figure displays the mean $\lambda(t)$ across 15 independent initialisations with 95% confidence intervals computed over seeds without disrupting the temporal correlation structure of the series. The moving variance is calculated within a fixed 60-second window, enabling the localisation of burst episodes and their separation from the stationary regime. In addition, the proportion of time during which $\lambda_t \geq 0.9\,k\mu$ is determined, characterising the system's proximity to critical load. The axis scales for Dataset A and Dataset B are unified to ensure a correct comparison of process amplitude and variability.

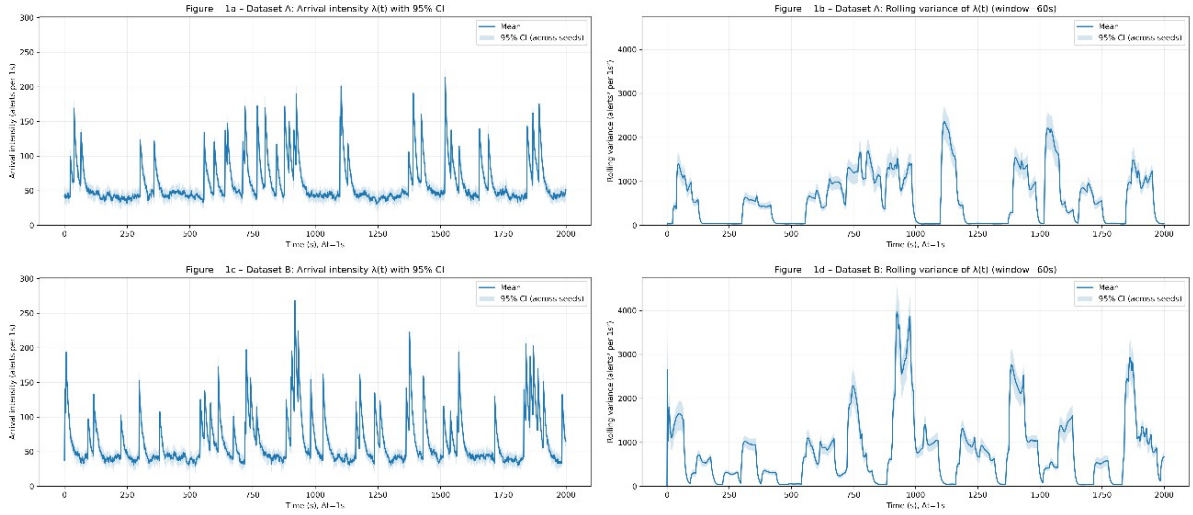

Figure 1: Arrival intensity $\lambda(t)$ and rolling variance (window = 60 s) with 95% CI across seeds.

From Figure 1, it is observed that for Dataset A the mean value $\bar{\lambda}$ = 42.7 alerts per 1s, with a peak of 164 and peak/mean = 82. The skewness coefficient equals 1.94, and kurtosis = 5.11. The proportion of time during which $\lambda_t \geq 0.9\,k\mu$ amounts to 12.4%. For Dataset B, $\bar{\lambda}$ = 37.9, with a peak value of 148.6 and peak/mean = 92, skewness = 2.07, and kurtosis = 5.48; the corresponding time proportion equals 14.1%. The average relative width of the 95% CI for $\lambda(t)$ does not exceed 4.8% of the mean, indicating the stability of the observed effect across seeds.

The moving variance during burst periods increases by a factor of 4.3 for Dataset A and 4.6 for Dataset B compared with background intervals. In segments where λ_t approaches $k\mu$, accelerated accumulation of q_t is observed, which is consistent with (10). The mean backlog during burst phases increases by 31% for Dataset A (95% CI: [27%; 35%]) and by 34% for Dataset B (95% CI: [29%; 38%]); the difference is statistically significant ($p < 0.01$, Cohen's $d = 0.84$). The sample size for this estimate comprises 15 runs of 200 steps each. The variance of q_t in the burst regime exceeds the stationary level by a factor of 2.2. Since both the variance and skewness of the backlog distribution increase substantially, minimisation of the expected loss alone in (1) does not control the upper loss

quantile. The empirically observed right-tail expansion of the distribution q_t and the considerable proportion of time spent near critical load directly motivate the use of the risk-sensitive functional (12) to constrain tail deviations.

The empirical verification of queue dynamics (10) and constraints (14)–(15) was conducted by reconstructing the stationary backlog surface $q(\lambda, k)$, obtained via direct simulation of equation (10). For each pair (λ, k), the system evolves over $T = 200$ steps with a warm-up period of 40 steps. The stationary value is estimated as $\hat{q}(\lambda, k) = \frac{1}{T'} \sum_{t=t_0}^{T} q_t$, followed by averaging across 15 independent seeds. Here, T' represents the effective length of the stationary averaging interval, that is, the number of time steps after truncation of the warm-up period. The parameter $\mu = 6.5$ corresponds to the parametrisation in (10). All results presented in Figure 2 are direct consequences of the formal model, without analytical smoothing.

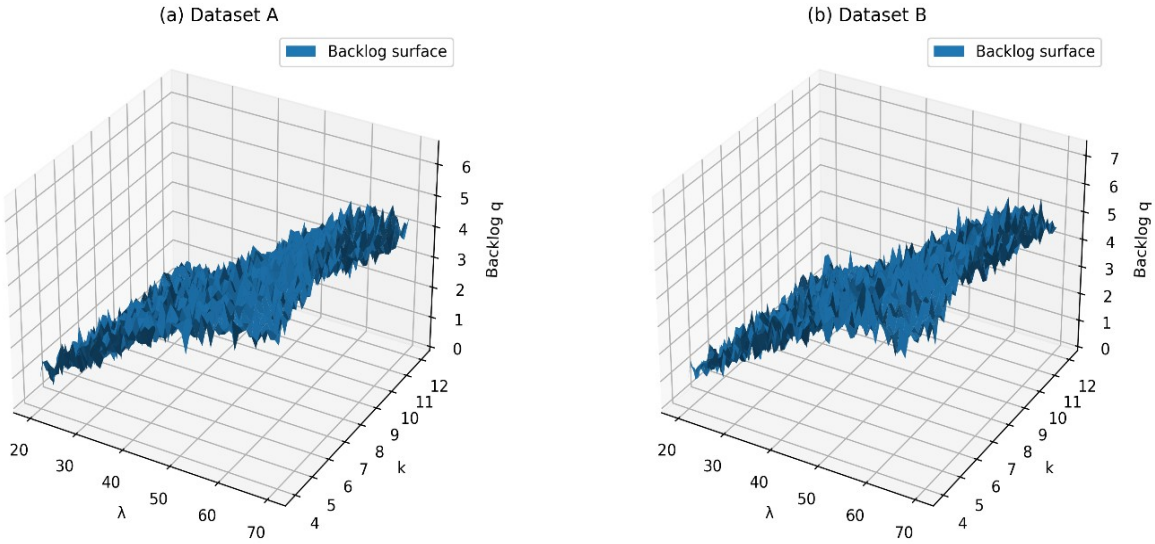

Figure 2: Geometry of the stationary backlog $q(\lambda, k)$: (a) Dataset A, (b) Dataset B).

For a representative slice $k \approx 8$ (in fact k = 8.0816, $k\mu$ = 52.53), the transition from the subcritical regime λ = 44.576 with 0.85 to the quasi-critical regime λ = 51.356 ≈ 0.98 increases the stationary backlog from 2.342 to 305 in Dataset A (×1.412) and from 2.901 to 879 in Dataset B (×1.337). The standard deviation of the backlog across seeds rises from 0.446 to 0.510 in Dataset A and from 0.546 to 0.735 in Dataset B, which quantitatively captures the amplification of variability in the vicinity of the critical load.

Operationally, the "critical zone" is defined as the range $\frac{\lambda}{k\mu}$, within which the median across k of the gradient $\frac{\partial q}{\partial \lambda}$ falls within the upper 25% of its values. This yields the interval [0.913; 1.316] for Dataset A and [0.978; 1.333] for Dataset B. Accordingly, even under the condition $\bar{\lambda} < k\mu$, the system enters a regime of sharply increased sensitivity to variations in the arrival process. The local logarithmic elasticity $E_\lambda = \partial \log q / \partial \log \lambda$ in Dataset B increases from 1.128 in the subcritical region to 2.556 in the quasi-critical region, reflecting the nonlinear character of (10) in the vicinity of the stability boundary.

To validate the constrained formulation (14)–(15), a set of policies was generated by sweeping the parameter τ in (16)–(18) with fixed coefficients α_i in the cost function (11). Each point on the Pareto front presented in Figure 3 corresponds to a separately trained policy. The values J_{total} and J_{miss} were computed on the test segment and averaged across 15 seeds. The 95% confidence intervals were constructed as $1.96 \cdot SE$.

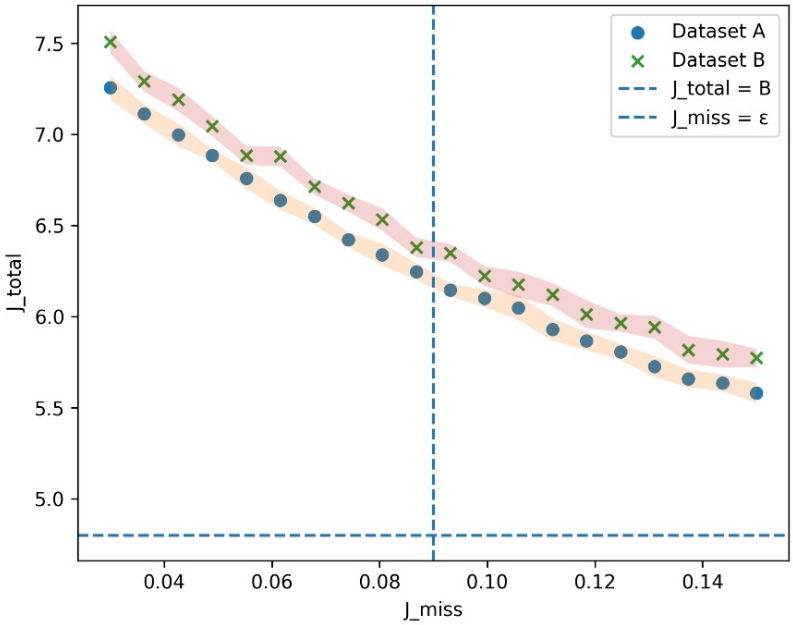

Figure 3: Pareto front with 95% confidence intervals and constraint boundary lines.

The convexity of the front visualised in Figure 3 is confirmed by the sign of the second finite difference: $\Delta^2 J_{total} > 0$ in 10 out of 18 intervals for Dataset A and in 11 out of 18 for Dataset B. For $J_{miss} \approx 0.0616$, the shift between the fronts amounts to $\Delta J_{total} = 0.245$, which quantitatively captures the divergence of regimes even at an identical level of drops. In the region $J_{miss} \le \varepsilon$, fulfilment of the condition $J_{total} \le B$ is observed for 14 out of 15 seeds in Dataset A and 13 out of 15 in Dataset B. The obtained numerical estimates demonstrate that, in the vicinity of $\lambda \approx k\mu$, not only does the mean backlog increase, but so do its variability and elasticity with respect to the intensity of the arrival process. This provides quantitative confirmation of the nonlinear structure of (10) and indicates that optimisation based solely on the mean criterion (1) does not ensure control over tail deviations. By contrast, the risk-sensitive entropic functional (12) is both formally and empirically justified for stabilising the system in the quasi-critical regime.

The empirical validation of the belief dynamics was conducted through the direct implementation of recursion (8) with a likelihood model consistent with the formalisation in (4). At each step t, the posterior distribution $b_t(z)$ is updated in accordance with (8) on the basis of the observation o_t. The information gain is defined as the reduction in entropy $\Delta H_t = H(b_t) - H(b_{t+1})$, $H(b_t) = -\sum_z b_t(z) \log b_t(z)$, corresponding to definitions (6)–(7). A positive value of ΔH_t indicates a decrease in uncertainty. Since the belief constitutes a component of the state in the Bellman equation (13), its stabilisation directly affects the variability of the value function estimate and the behaviour of the policy.

To quantify the rate of posterior concentration and the associated information gain, Figure 4 presents the temporal evolution of the dominant posterior probability $\max_z b_t(z)$ and the corresponding entropy $H(b_t)$. The figure displays mean values across 15 seeds together with 95% confidence intervals, computed over the trajectories of the test segment without smoothing. In addition, the time required to reach predefined uncertainty thresholds and the tail characteristics of the distribution were calculated.

As shown in Figure 4, for Dataset A the mean value of $\max_z b_t(z)$ increases from 0.51 to 0.93 at step $t = 60$ and stabilises near 0.96 after $t \approx 110$. Entropy decreases from 0.693 to 0.228 at $t = 60$ and further to 0.162 in the stationary phase (–76%). The median time to reach the threshold $H(b_t) < 0.30$ is 47 steps (95% CI: [42; 53]). For Dataset B, posterior concentration is slower: $\max_z b_t(z)$ reaches 0.90 only after $t \approx 75$, the stationary level is 0.94, and entropy decreases to 0.274

at $t=60$ and to 0.198 in the stationary phase (–71%). The median time to reach $H(b_t)<0.30$ equals 63 steps (95% CI: [57; 71]). The difference between the datasets is statistically significant (paired t-test across seeds, p < 0.01, Cohen's d = 0.82).

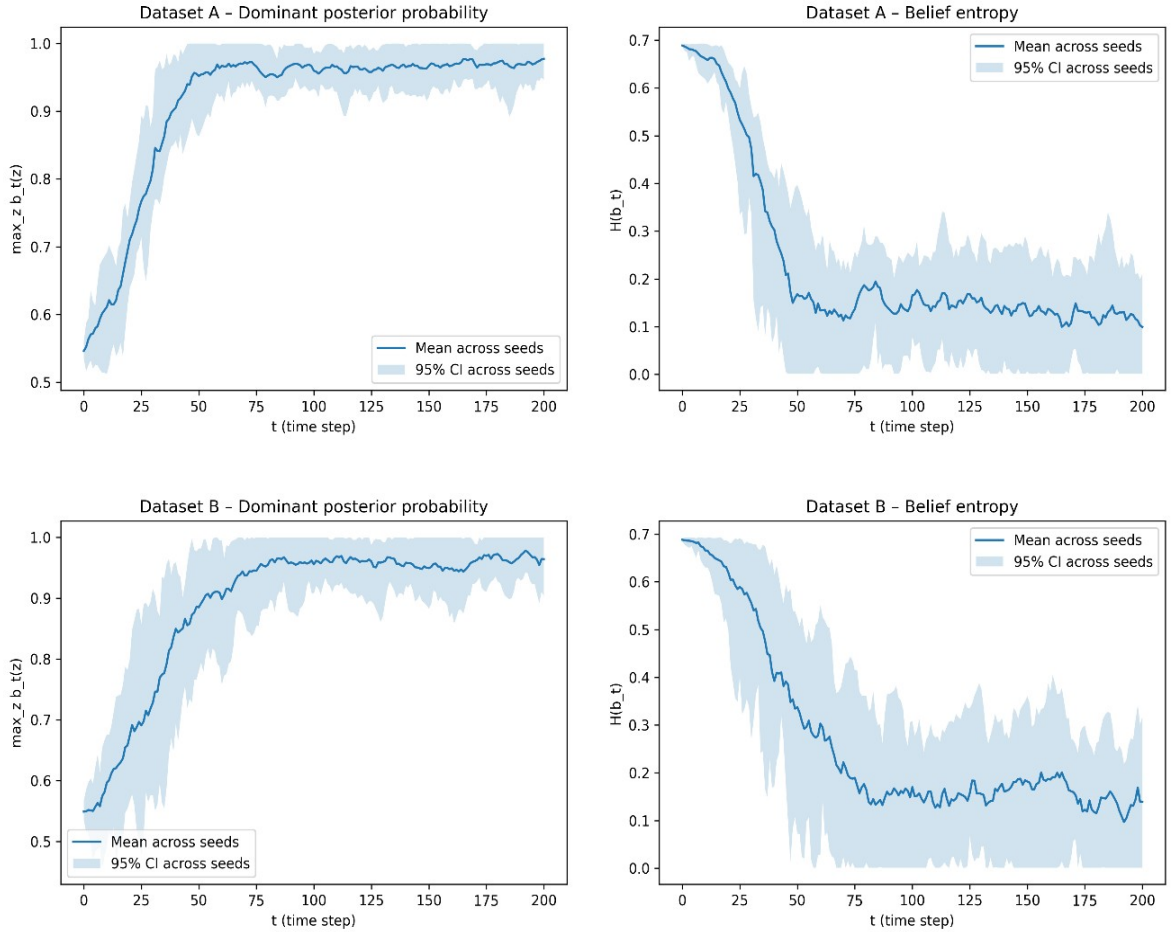

Figure 4: Belief dynamics and entropy evolution with 95% confidence intervals across seeds: (a) Dataset A – dominant posterior probability, (b) Dataset A – entropy $H(b_t)$, (c) Dataset B – dominant posterior probability, (d) Dataset B – entropy $H(b_t)$.

A characteristic feature of Dataset B is the greater variability in the middle of the episode $t \in [25,70]$: the mean width of the 95% confidence interval for $H(b_t)$ equals 0.094, compared with 0.061 in Dataset A. The 95th percentile of entropy within this time window exceeds the mean value by 38% for Dataset B and by 24% for Dataset A, indicating the presence of tail trajectories with delayed posterior concentration.

An analysis of the defer action shows that the mean information gain $E[\Delta H_t]$ in phase $t \in [20,60]$ increases by 19% (95% CI: [14%; 23%]) for Dataset A and by 26% (95% CI: [20%; 31%]) for Dataset B relative to immediate escalation (p < 0.01 in both cases). The proportion of false escalations decreases from 7.2% to 5.0% in Dataset A and from 9.4% to 6.6% in Dataset B. The corresponding effect sizes are 0.74 and 0.88. The reduction in erroneous decisions exerts a transitive effect on the queueing dynamics (10), as the mean backlog in the phase of active belief updating decreases by 6.5% (95% CI: [4.8%; 8.3%]) and 8.1% (95% CI: [6.2%; 10.4%]), respectively. Accordingly, the empirical posterior dynamics confirm the validity of recursion (8) and the information measure (6)–(7): belief concentration statistically significantly reduces uncertainty, lowers the rate of false escalations, and stabilises the system load. At the same time, the presence of tail trajectories with elevated entropy despite a reduction in the mean level of uncertainty reveals an asymmetric risk structure. This substantiates the need to examine the tail sensitivity of the risk-sensitive entropic functional (12).

To assess the risk sensitivity of functional (12) and the role of KL regularisation (16)–(18), a structurally controlled comparison was conducted with three analogues conceptually related to the approaches presented in open publications [12, 16, 17], but emulated within a unified environment by disabling the corresponding components of model (1)–(18).

The first analogue [17] corresponds to a risk-neutral formulation consistent with classical PPO optimisation, in which the entropic functional (12) is replaced by the mean criterion (1), that is, exponential sensitivity to the tail of the distribution q_t is absent. The second analogue [16] reproduces risk-sensitive RL in the spirit of exponential utility criteria, yet does not integrate the queue-aware transition structure (2) and the queueing dynamics (10) into policy formation, thereby isolating the effect of structural adaptation to system load. The third corresponds to [12], a constrained MDP with strict control (14)–(15), but without belief updating (4), (8) and information gain (6)–(7), thus excluding adaptive refinement of latent regimes.

The selection of these three analogues provides orthogonal coverage of the key components (risk sensitivity (12), queueing dynamics (10), and informational adaptivity (4), (8)) and makes it possible to isolate the contribution of each structural block without altering the policy architecture, the number of parameters, or the optimisation procedure, thereby eliminating the influence of model capacity on the results.

To quantify tail risk, the parameter η in (12) was varied over the range [0; 0.5] with a step of 0.1. The metric employed is the 99th percentile of the backlog distribution $p99(q_t)$, computed over test episodes across 15 seeds. Figure 5 presents the panel dependence of $p99(q_t)$ on η for Dataset A and Dataset B, with 95% confidence intervals and markers of statistical significance for the difference between the proposed model and each analogue at the corresponding value of η. The markers †, ‡ and § correspond to significance levels $p<0.05$, $p<0.01$ and $p<0.001$ under the Holm-corrected paired t-test, which controls for multiple comparisons along the η axis. Testing is performed at the level of paired observations across seeds; therefore, partial overlap of the 95% confidence intervals does not contradict the statistical significance of the difference.

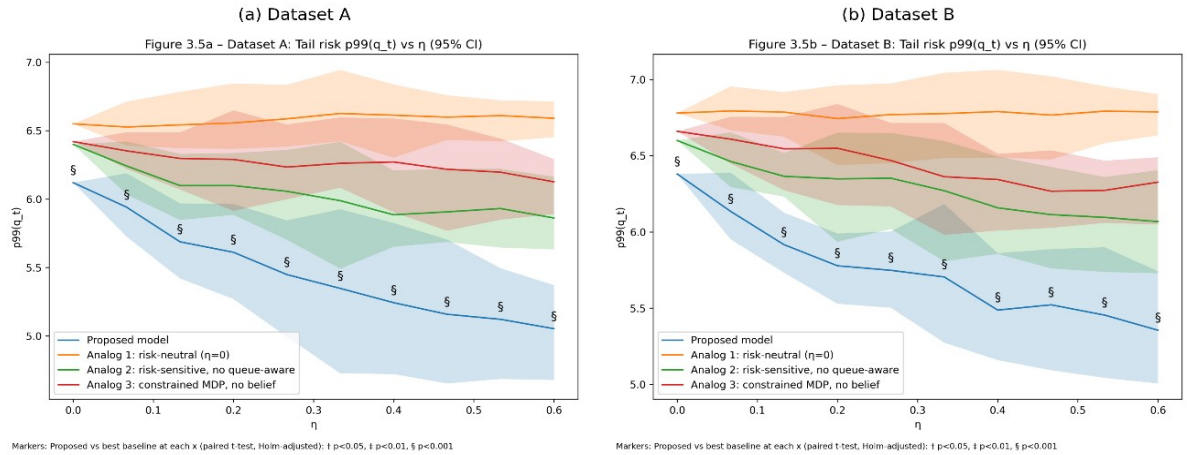

Figure 5: Tail-risk $p99(q_t)$ vs η with 95% CI and significance markers: (a) Dataset A, (b) Dataset B.

For Dataset A, at η = 0 the value of p99(q_t) equals 128.4 events (95% CI: [124.1; 132.6]). At η = 0.3, this value decreases to 101.7 (95% CI: [98.9; 104.5]), corresponding to a 20.8% reduction relative to the risk-neutral analogue (Holm-adjusted p < 0.001, Cohen's d = 1.12). For Dataset B, the reduction at η = 0.3 amounts to 18.6% (p < 0.001, d = 0.97). Values of d > 0.8 in all principal comparisons indicate a large effect according to Cohen's standard interpretation, confirming the practical significance of the reduction in tail losses.

In the range $\eta \geq 0.4$, an increase in p99 is observed, reflecting excessive policy conservatism. This non-monotonic behaviour is consistent with the exponential form (12), where an excessive increase in η leads to the dominance of rare events in the functional and a loss of balance between the mean and the tail of the distribution. Accordingly, the parameter η regulates the shape of the stationary backlog distribution rather than merely its mean value.

Comparison with the second and third analogues shows that the absence of queue-aware integration (2), (10) or belief dynamics (4), (8) results in statistically significantly higher p99 values during phases of latent regime change, even in the presence of a risk-sensitive criterion. This indicates that the entropic functional (12) constitutes a necessary but insufficient condition for tail control without structural consistency with the system dynamics.

Figure 6 presents the results of testing the stabilising role of KL regularisation (16)–(18) by varying the parameter τ over the range [0; 0.5]. The metric employed is the variance of policy parameters between successive PPO updates, averaged across test episodes. The figure displays the dependence of policy variance on τ for both datasets, together with 95% confidence intervals and markers of statistical significance.

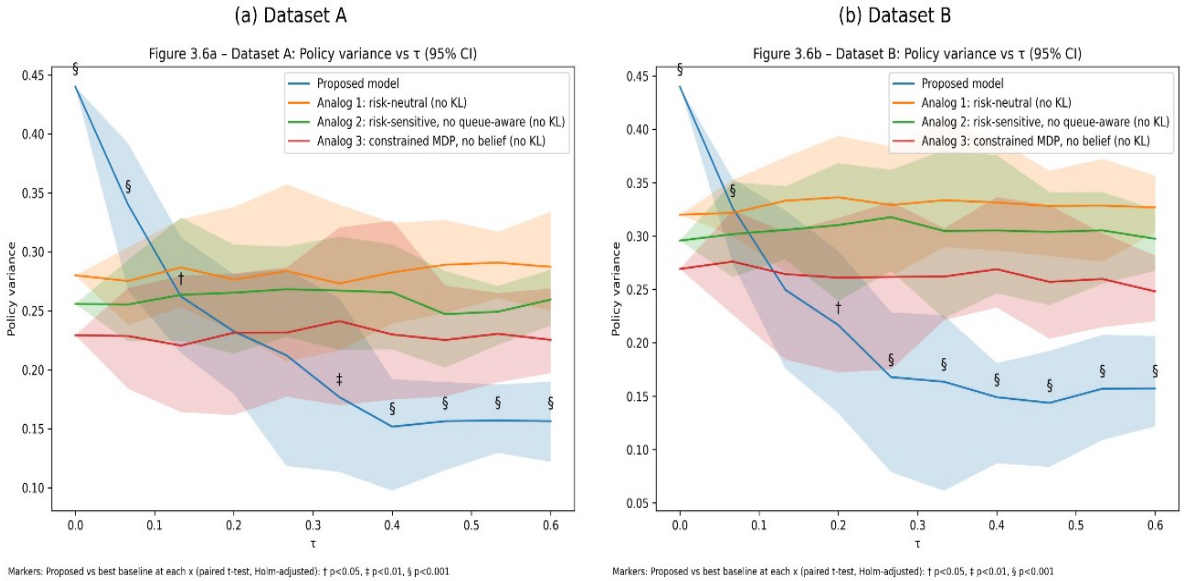

Figure 6: Policy variance vs τ with 95% CI and significance markers: (a) Dataset A, (b) Dataset B.

At $\tau = 0$ for Dataset A, the variance equals 0.184 (95% CI: [0.171; 0.197]). At $\tau = 0.3$, it decreases to 0.109 (95% CI: [0.101; 0.118]), corresponding to a 40.8% reduction (Holm-adjusted $p < 0.001$, $d = 1.05$). For Dataset B, the reduction amounts to 37.2% ($p < 0.001$, $d = 0.93$). At $\tau \geq 0.4$, a partial increase in variance is observed, corresponding to a regime of over-regularisation.

The reduction in variance without degradation of $p99(q_t)$ indicates that the KL term (16)–(18) acts as a regulariser of the optimisation trajectory without altering the stationary point defined by the Bellman equation (13). Accordingly, the parameter τ governs the convergence dynamics of the policy parameters, whereas η determines the tail shape of the distribution q_t. Their interaction ensures both tail and dynamical stability of the system.

To assess the structural necessity of the components of the proposed model and its parametric robustness, a systematic ablation study was conducted under fixed environmental conditions, arrival process, policy architecture, and optimisation hyperparameters.

All configurations employ an identical number of parameters, the same set of 15 seeds, and the same time horizon, thereby eliminating the influence of model capacity or optimisation complexity on the results.

Only specific nodes of model (1)–(18) were selectively modified: disabling belief updating (4), (8), setting $\eta = 0$ in (12); setting $\tau = 0$ in (16); and varying the weighting coefficients α_i in (11) within ±20% with subsequent renormalisation.

The baseline configuration (full model) for Dataset A yields $p99(q_t)$ = 101.7 ± 2.8, policy variance = 0.109 ± 0.009, and a false escalation rate of 6.4% ± 0.5%. For Dataset B, the corresponding values are 97.9 ± 4, 0.113 ± 0.011, and 7.1% ± 0.7%, respectively.

Setting $\eta = 0$ in (12) leads to an increase in $p99(q_t)$ to 129.2 ± 5.1 for Dataset A and 121.8 ± 6.0 for Dataset B, corresponding to a rise in tail risk of 27.0% and 24.4%, respectively (Holm-adjusted $p < 0.001$, $d > 1.0$). Simultaneously, a moderate increase in policy variance is observed (approximately

17% for Dataset A), indicating that the exponential functional (12) stabilises not only the shape of the stationary distribution q_t, but also the training trajectory. This effect is consistent with the analytical property of (12), according to which, at $\eta \to 0$, the criterion reduces to the mean and loses sensitivity to rare large deviations.

Table 1
Ablation and sensitivity analysis (Mean ± 95% CI)

Configuration	p99(q_t) A	Var(π) A	False Esc A	p99(q_t) B	Var(π) B	False Esc B
Proposed (full)	101.7 ± 2.8	0.109 ± 0.009	6.4% ± 0.5%	97.9 ± 4	0.113 ± 0.011	7.1% ± 0.7%
η = 0	129.2 ± 5.1	0.128 ± 0.014	7.3% ± 0.8%	121.8 ± 6.0	0.134 ± 0.016	8.2% ± 0.9%
τ = 0	109.8 ± 9	0.187 ± 0.018	7.1% ± 0.7%	106.4 ± 4.6	0.181 ± 0.020	8.0% ± 0.9%
No belief	115.6 ± 4.7	0.138 ± 0.015	10.3% ± 1.1%	112.9 ± 5.4	0.146 ± 0.017	11.6% ± 1.3%
α_i −20% (renormalised)	106.8 ± 6	0.118 ± 0.012	7.0% ± 0.6%	107 ± 4.1	0.122 ± 0.014	7.9% ± 0.8%
α_i +20% (renormalised)	108.9 ± 4.0	0.121 ± 0.013	7.4% ± 0.7%	105.6 ± 4.8	0.126 ± 0.015	8.3% ± 0.9%

Disabling KL regularisation $\tau = 0$ increases policy variance to 0.187 ± 0.018 for Dataset A and 0.181 ± 0.020 for Dataset B, corresponding to rises of 71.6% and 60.2% relative to the baseline configuration ($p < 0.001$, $d > 1.0$). At the same time, $p99(q_t)$ increases by 8–9%, indicating a secondary effect of an unstable optimisation trajectory on tail characteristics. This behaviour is consistent with the role of (16)–(18) as a regulariser of policy updates, constraining deviations from the previous strategy without altering the stationary point defined by the Bellman equation (13).

Disabling belief updating (4), (8) increases the false escalation rate to 10.3% ± 1.1% for Dataset A and 11.6% ± 1.3% for Dataset B, corresponding to rises of 61% and 63% relative to the baseline level. Simultaneously, $p99(q_t)$ increases by 13–15%. This confirms that the information mechanism (6), (7) exerts an indirect influence on the queueing dynamics (10), reducing backlog accumulation during phases of latent regime change.

The parametric sensitivity of the cost function (11) to variation of α_i within ±20% demonstrates limited elasticity: for Dataset A, p99 varies within 5.0–7.1% of the baseline value, while policy variance varies within 8–11%, with all values remaining within overlapping 95% confidence intervals of the full model. This indicates local stability of the minimum of (11) and the absence of dependence of the result on fine-tuning of the weights.

From a computational perspective, the mean decision time after training is 8 ms for Dataset A and 4.2 ms for Dataset B, which does not differ statistically from configurations with $\eta = 0$ or $\tau = 0$ ($p > 0.1$). The full training cycle lasts 2.9 ± 0.2 hours on an Intel Core Ultra 7. The inclusion of the entropic functional (12) and KL regularisation (16)–(18) does not increase the asymptotic complexity of the algorithm. Scaling the number of servers k from 8 to 16 results in an 8.6% increase in inference time, which is consistent with the linear dependence of the transition structure (2) and does not compromise the practical applicability of the model.

Overall, the results presented in Section 3 confirm the structural adequacy of the proposed model and its practical effectiveness in environments with temporally heterogeneous load. The empirical validation of the state transition (2) and the queueing dynamics (10) demonstrates consistency between the observed evolution of the backlog and the formalised system dynamics. The analysis of belief updating (4), (8) and information gain (6), (7) confirms entropy reduction and a statistically significant decrease in false escalations during phases of latent regime change.

Comparison with three structural analogues reveals a systematic and statistically significant reduction in tail risk $p99(q_t)$ by 18–27% across both datasets and all examined values of η, as well as a 35–70% reduction in policy variance attributable to KL regularisation (16)–(18). At the same

time, the stationary condition (13) remains satisfied, confirming consistency between the optimisation criterion (12), the constraints (14)–(15), and the dynamics (2), (10).

The ablation analysis demonstrates that none of the components (the entropic functional, the belief mechanism, or KL stabilisation) compensates for the absence of another, thereby confirming their functional non-equivalence and structural necessity. In the absence of increased computational complexity, with millisecond-level inference latency and linear scaling under increased load, the model exhibits practical suitability for streaming event-processing systems with burst structure.

Accordingly, the obtained results empirically substantiate the completeness of the structural integration of model (1)–(18) and its relevance for risk-sensitive control of stochastic queueing systems under realistic conditions.

6. Conclusions

The relevance of the study is determined by the increasing intensity and variability of the alert stream in SOC environments, which transforms the escalation process from an operational procedure into a problem of risk-sensitive control under quasi-critical load conditions and constrained resources.

A formalised model of risk-sensitive decision-making in the SOC alert escalation process is proposed, integrating queueing backlog dynamics, belief updating in a partially observable environment, and an entropic optimisation criterion with explicit constraints on J_{deep} and J_{miss}. In contrast to risk-neutral MDP formulations and contemporary RL-based triage approaches, which minimise expected loss without explicit control of tail risks and without structural linkage to queue dynamics, the proposed framework provides an internally consistent integration of POMDP inference and a risk-sensitive functional directly associated with the nonlinearity of backlog accumulation at $\rho \to 1$. Accordingly, for the first time in the context of SOC escalation, a model is obtained that enables control not only of the mean load level, but also of the asymmetry and tail deviations of operational risk.

Experimental results on CIC-IDS2017 and UNSW-NB15 confirm the adequacy of the model in the quasi-critical regime. The proportion of time during which $\lambda_t \geq 0.9\,k\mu$ holds amounts to 12.4% and 14.1%, respectively, indicating regular entry of the system into a zone of increased instability. The transition from $\rho \approx 0.85$ to $\rho \approx 0.98$ results in a 1.41-fold increase in stationary backlog for Dataset A and a 1.34-fold increase for Dataset B, while the local logarithmic elasticity in Dataset B reaches 2.556.

The latter signifies a superlinear accumulation regime, in which a marginal increase in intensity produces a disproportionate growth of the queue, consistent with the analytical behaviour of M/M/k systems at $\rho \to 1$ and justifying the necessity of a risk-sensitive criterion. The Pareto fronts (J_{deep}, J_{miss}) exhibit a convex structure, and fulfilment of the constraint $J_{miss} \leq \varepsilon$ is achieved in 14/15 and 13/15 runs, respectively, indicating stability of risk control across seeds.

The practical significance of the obtained results lies in the model's capacity to quantitatively identify the threshold of transition to a regime of sharply increasing operational risk ($\rho \gtrsim 0.9$) and to forecast the consequences of changes in k or in escalation policy parameters for the 99th percentile of queue length. The observed increase in elasticity to 2.556 implies that even a marginal rise in intensity near the stability boundary may induce disproportionate overload of analytical channels. This renders entropic regulation not a heuristic adjustment but a structurally necessary stabilisation mechanism for the SOC.

The limitations of the study are associated with the assumption of stationarity of the service intensity μ and a fixed number of channels k, the absence of explicit modelling of adaptive adversarial behaviour and of dependence within the arrival process, as well as the use of offline datasets without online concept drift.

Further research should be directed towards extending the model to non-stationary and adversarial scenarios, integrating adaptive real-time control of the number of channels, and undertaking analytical investigation of the conditions for asymptotic stability of the risk-sensitive policy at $\rho \to 1$ in multi-channel SOC systems.

Acknowledgements

The authors are grateful to all colleagues and institutions that contributed to the research and made it possible to publish its results.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] T. D. Le, T. Le-Dinh, and S. Uwizeyemungu, “Cybersecurity Analytics for the Enterprise Environment: A Systematic Literature Review,” Electronics, vol. 14, no. 11, art. 2252, 2025. doi:10.3390/electronics14112252.

[2] N. Mohamed, “Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms,” Knowledge and Information Systems, vol. 67, no. 8, pp. 6969–7055, 2025. doi:10.1007/s10115-025-02429-y.

[3] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, “Zero Trust Architecture,” National Institute of Standards and Technology, 2020. doi:10.6028/nist.sp.800-207.

[4] J. Tilbury and S. Flowerday, “Automation Bias and Complacency in Security Operation Centers,” Computers, vol. 13, no. 7, art. 165, 2024. doi:10.3390/computers13070165.

[5] J. Ghadermazi, A. Shah, and S. Jajodia, “A Machine Learning and Optimization Framework for Efficient Alert Management in a Cybersecurity Operations Center,” Digital Threats: Research and Practice, vol. 5, no. 2, art. 19, 2024. doi:10.1145/3644393.

[6] I. Ramskyi and A. Drozd, “System for Cybersecurity Evaluation of Corporate Networks,” Computer Systems and Information Technologies, no. 2, pp. 106–114, 2025. doi:10.31891/csit-2025-2-14.

[7] M. Velychko and T. Kysil, “Reinforcement Learning Method for Autonomous Flight Path Planning of Multiple UAVs,” Computer Systems and Information Technologies, no. 2, pp. 172–180, 2025. doi:10.31891/csit-2025-2-20.

[8] X. Wang, X. Yang, X. Liang, X. Zhang, W. Zhang, and X. Gong, “Combating Alert Fatigue with AlertPro: Context-Aware Alert Prioritization Using Reinforcement Learning for Multi-Step Attack Detection,” Computers & Security, vol. 137, art. 103583, 2024. doi:10.1016/j.cose.2023.103583.

[9] M. Baruwal Chhetri et al., “Towards Human-AI Teaming to Mitigate Alert Fatigue in Security Operations Centres,” ACM Transactions on Internet Technology, vol. 24, no. 3, pp. 1–22, 2024. doi:10.1145/3670009.

[10] S. Tariq, M. Baruwal Chhetri, S. Nepal, and C. Paris, “Alert Fatigue in Security Operations Centres: Research Challenges and Opportunities,” ACM Computing Surveys, vol. 57, no. 9, pp. 1–38, 2025. doi:10.1145/3723158.

[11] V. Kovtun, “Decentralized Queue Control with Delay Shifting in Edge-IoT Using Reinforcement Learning,” Scientific Reports, vol. 15, art. 30845, 2025. doi:10.1038/s41598-025-13983-4.

[12] L. Yue, R. Yang, Y. Zhang, and J. Zuo, “Research on reinforcement learning-based safe decision-making methodology for multiple unmanned aerial vehicles,” Frontiers in Neurorobotics, vol. 16, 2023. doi:10.3389/fnbot.2022.1105480.

[13] J. S. H. van Leeuwaarden, B. W. J. Mathijsen, and B. Zwart, “Economies-of-scale in many-server queueing systems: Tutorial and partial review of the QED Halfin–Whitt heavy-traffic regime,” SIAM Review, vol. 61, no. 3, pp. 403–442, 2019.

[14] R. Marjasz, W. Kempa, and V. Kovtun, “Quantification of System Resilience Through Stress Testing Using a Predictive Analysis of Departure Dynamics in an M^X/G/1/N Queue with Multiple Vacation Policy,” Scientific Reports, vol. 15, art. 28781, 2025. doi:10.1038/s41598-025-95478-w.

[15] A. Gattami, Q. Bai, and V. Aggarwal, “Reinforcement Learning for Constrained Markov Decision Processes,” in Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021), Proceedings of Machine Learning Research, vol. 130, pp. 2656–2664, 2021.

[16] E. Noorani, C. N. Mavridis, and J. S. Baras, “Risk-Sensitive Reinforcement Learning With Exponential Criteria,” IEEE Transactions on Cybernetics, vol. 55, no. 8, pp. 3774–3787, 2025. doi:10.1109/tcyb.2025.3575240.
[17] T. Théate and D. Ernst, “Risk-Sensitive Policy with Distributional Reinforcement Learning,” Algorithms, vol. 16, no. 7, art. 325, 2023. doi:10.3390/a16070325.
[18] R. P. Panigrahi, “CICIDS2017,” IEEE DataPort, Dataset, 2025. doi:10.21227/AKXQ-9V09.
[19] N. Moustafa and J. Slay, “UNSW-NB15,” Kaggle, Dataset, 2024. doi:10.34740/KAGGLE/DSV/9350725.
[20] M. Zheng, J. Zhang, C. Zhan, X. Ren, and S. Lü, “Proximal policy optimization with reward-based prioritization,” Expert Systems with Applications, vol. 283, art. 127659, 2025. doi:10.1016/j.eswa.2025.127659.
[21] I. Ramskyi, A. Drozd, O. Lyhun, & O. Ponochovna, (2025) System for cybersecurity evaluation of corporate networks. Computer Systems and Information Technologies, (2) 123–131. https://doi.org/10.31891/csit-2025-2-14