Computers & Security 121 (2022) 102844

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

TC 11 Briefing Papers

RADAMS: Resilient and adaptive alert and attention management strategy against Informational Denial-of-Service (IDoS) attacks☆,☆☆

Linan Huang*, Quanyan Zhu

Department of Electrical and Computer Engineering, New York University, Brooklyn, NY, 11201, USA

ARTICLE INFO

Article history:
Received 18 November 2021
Revised 27 June 2022
Accepted 17 July 2022
Available online 20 July 2022

Keywords:
Human attention vulnerability
Feint attacks
Reinforcement learning
Risk analysis
Cognitive load
Alert fatigue

ABSTRACT

Attacks exploiting human attentional vulnerability have posed severe threats to cybersecurity. In this work, we identify and formally define a new type of proactive attentional attacks called Informational Denial-of-Service (IDoS) attacks that generate a large volume of feint attacks to overload human operators and hide real attacks among feints. We incorporate human factors (e.g., levels of expertise, stress, and efficiency) and empirical psychological results (e.g., the Yerkes-Dodson law and the sunk cost fallacy) to model the operators' attention dynamics and their decision-making processes along with the real-time alert monitoring and inspection. To assist human operators in dismissing the feints and escalating the real attacks timely and accurately, we develop a Resilient and Adaptive Data-driven alert and Attention Management Strategy (RADAMS) that de-emphasizes alerts selectively based on the abstracted category labels of the alerts. RADAMS uses reinforcement learning to achieve a customized and transferable design for various human operators and evolving IDoS attacks. The integrated modeling and theoretical analysis lead to the Product Principle of Attention (PPoA), fundamental limits, and the tradeoff among crucial human and economic factors. Experimental results corroborate that the proposed strategy outperforms the default strategy and can reduce the IDoS risk by as much as 20%. Besides, the strategy is resilient to large variations of costs, attack frequencies, and human attention capacities. We have recognized interesting phenomena such as attentional risk equivalency, attacker's dilemma, and the half-truth optimal attack strategy.

1. Introduction

Human vulnerability and human-induced security threats have been a long-standing and fast-growing problem for the security of Industrial Control Systems (ICSs). According to Verizon (Bassett et al., 2021), 85% data breaches involve human errors. Attentional vulnerability is one of the representative human vulnerabilities. Adversaries have exploited human inattention to launch social engineering attacks and phishing attacks toward employees and users. According to the report (Tessian, 2020), 29% of employees fall for a phishing scam, and 36% send a misdirected email, owing to lack of attention. These attentional attacks are *reactive* as they exploit the existing human attention patterns. On the contrary, *proactive* attentional attacks can strategically change the attention pattern of a human operator or a network administrator. For example, an attacker can launch feint attacks to trigger a large volume of alerts and overload the human operators so that operators fail to inspect the alert associated with real attacks (Hitzel, 2019). We refer to this new type of attacks as the Informational Denial-of-Service (IDoS) attacks, which aim to deplete the limited attention resources of human operators to prevent them from accurate detection and timely defense.

IDoS attacks bring significant security challenges to ICSs for the following reasons. First, alert fatigue has already been a serious problem in the age of infobesity with terabytes of unprocessed data or manipulated information. According to the Ponemon Institute research report (LLC, 2015), organizations spend nearly 21,000 hours each year analyzing false alarms, which costs organizations an average of $1.27 million per year. IDoS attacks exacerbate the problem by generating feints to intentionally increase the percentage of false-positive alerts. Second, IDoS attacks directly target the human operators and security analysts in the Security Operations Center (SOC) that acts as the 'central immune system' in ICSs. Third, as ICSs become increasingly complicated and time-critical,

☆ This work was supported in part by the National Science Foundation (NSF) under Grants ECCS-1847056, CNS-2027884, and BCS-2122060; and in part by Army Research Office (ARO) under Grant W911NF-19-1-0041 and DOE-NE under Grant 20–19829.

☆☆ A preliminary version of this work Huang and Zhu (2021a) was presented at the 12-th Conference on Decision and Game Theory for Security

* Corresponding author.
E-mail addresses: lh2328@nyu.edu (L. Huang), qz494@nyu.edu (Q. Zhu).

https://doi.org/10.1016/j.cose.2022.102844
0167-4048/

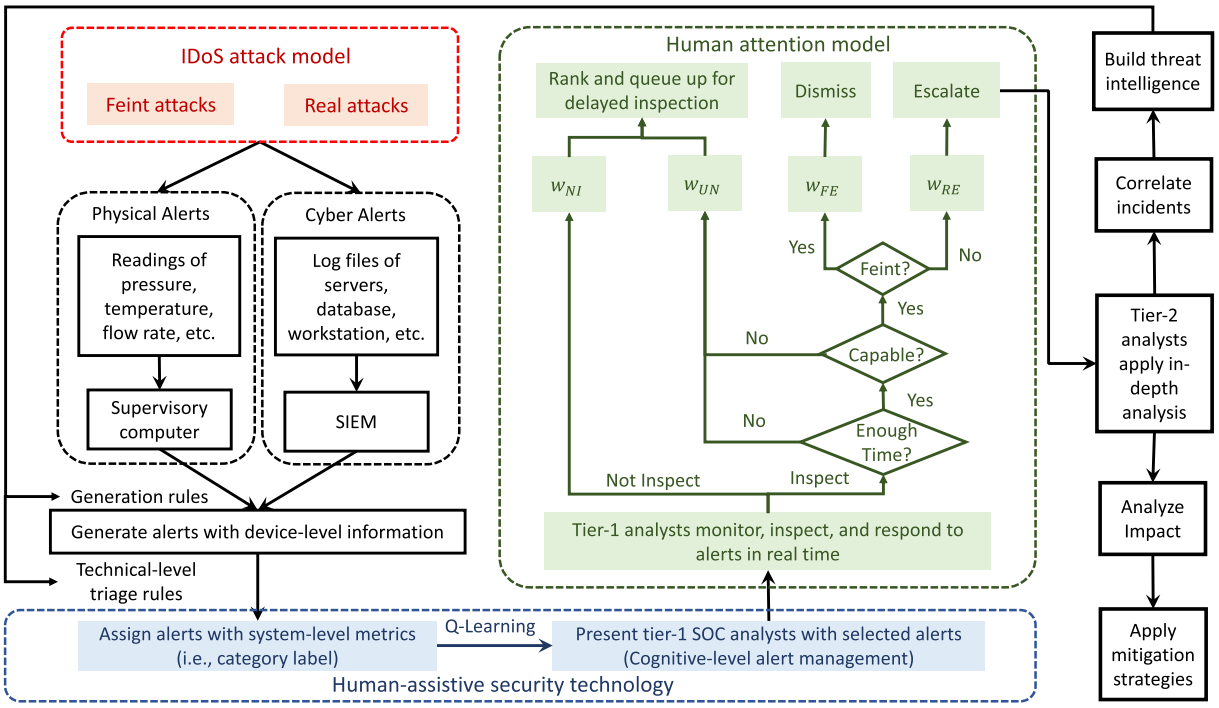


Fig. 1. The overview diagram of RADAMS against IDoS in ICS, which incorporates the IDoS attack model, the human attention model, and the human-assistive security technology in the red, green, and blue boxes, respectively. RADAMS consolidates the *technical-level* (i.e., generation rules and triage rules in black) and the *cognitive-level* (data-driven human-aware alert de-emphasis in blue) alert management before the manual inspection in green to reduce the operators' cognitive load. The modern SOC adopts a hierarchical alert analysis process. The tier-1 SOC analysts, also referred to as the operators, are in charge of real-time alert monitoring and inspections. The tier-2 SOC analysts are in charge of the in-depth analysis. All processes in black are not the focus of this work. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the human operators require higher expertise levels to understand the domain information and detect feints (Stouffer et al., 2011) in time to avoid life-threatening failures or huge economic losses. The SOCs in ICSs are usually understaffed, due to these high-standard requirements. Fourth, since human operators behave differently, and IDoS attacks are a broad class of adaptive attacks, it is challenging (yet highly desirable) to develop a customized and resilient defense. Due to the above factors, including the huge economic loss, there is an apparent need to understand this class of proactive attentional attacks, quantify its consequences and risks, and develop associated mitigation strategies.

To this end, we establish a holistic model of the IDoS attacks, the alert generations, and the human operators' alert responses. In the IDoS attack model, we adopt a Markov renewal process to characterize the sequential arrival of feints and real attacks that target different ICS assets. We define a *revelation probability* to abstract the alert generation and triage process of existing detection systems. The revelation probability maps the attacks' hidden types and targets stochastically to the associated alerts' observable category labels. To model the human operators' attention dynamics and alert responses under the IDoS attacks, we directly incorporate the operators' levels of expertise, stress, and efficiency into the security design based on the existing results from the literature in psychology, including the Yerkes-Dodson law (Yerkes et al., 1908) and the sunk cost fallacy (Arkes and Blumer, 1985). To assist human operators in alert inspection and response, compensate for their attentional vulnerabilities, and combat IDoS attacks, we develop human-centered technologies that selectively make some alerts less noticeable based on their category labels. Reinforcement learning is applied to make the human-assistive security technology *resilient, automatic*, and *adaptive* to various human models and attack scenarios.

Fig. 1 illustrates the overview diagram of Resilient and Adaptive Alert and Attention Management Strategy (RADAMS). We use the following control room scenario to elaborate on the entire process of RADAMS under IDoS attacks. Supervisory computers and Security Information and Event Management (SIEM) continuously monitor the physical readings and cyber log files, respectively, to generate alerts with *device-level* information. Since manual inspection and response of these alerts (illustrated in green) are indispensable for ICSs at the current stage, RADAMS adopts the following technical-level and cognitive-level automated alert selection schemes, illustrated in black and blue, respectively, to assist manual alert inspection. The technical-level alert selection scheme focuses on selecting and prioritizing alerts based on the *device-level* information and abstract *system-level* metrics. Although the above alert triage process significantly reduces the workload of the manual inspection, a sizeable number of alerts remain to be inspected, especially under a large volume of feints. To this end, RADAMS incorporates the cognitive-level alert selection to accommodate the operators' cognition limitation in the subsequent alert inspections. After the technical-level and cognitive-level alert management, RADAMS presents the selected alerts to the tier-1 SOC analysts in the control room for real-time monitoring and response. The alerts associated with the real attack will be identified and escalated to tier-2 analysts for in-depth analysis. The analysis outcomes of tier-2 analysts are used to mitigate the current threats and improve the generation rules and technical-level triage rules.

RADAMS enriches the existing alert selection frameworks with the IDoS attack model, the human attention model, and the human-assistive security technology highlighted in red, green, and blue, respectively. Through the integrated modeling and theoretical analysis, we obtain the *Product Principle of Attention* (PPoA), which states that the Attentional Deficiency Level (ADL), i.e., the proba-

Table 1
Summary of notations in Section 3.

Variable	Meaning
$t^k \in [0, \infty)$	Arrival time of the k-th attack.
$\tau^k = t^{k+1} - t^k \in [0, \infty)$	Inter-arrival time at attack stage $k \in \mathbb{Z}^{0+}$.
$\kappa_{AT} \in \mathcal{K}_{AT}$	Transition kernel of attacks.
$z \in \mathcal{Z}$	Probability Density Function (PDF) of the inter-arrival time.
$\theta^k \in \Theta := \{\theta_{FE}, \theta_{RE}\}$	Attack's type at attack stage $k \in \mathbb{Z}^{0+}$.
$\phi^k \in \Phi$	Attack's target at attack stage $k \in \mathbb{Z}^{0+}$.
$s^k \in \mathcal{S}$	Alert's category label at attack stage k.
$o(s^k \mid \theta^k, \phi^k)$	Revelation kernel of category labels.
$b(\theta^k, \phi^k)$	Steady-state distribution.
$\kappa_{CL} \in \mathcal{K}_{CL}$	Transition kernel of category labels.

bility of incomplete alert responses, and the risk of IDoS attacks depend on the product of the supply and the demand of human attention resources. The closed-form expressions under mild assumptions lead to several fundamental limits, including the minimum ADL and the maximum length of de-emphasized alerts to reduce IDoS risk. We explicitly characterize the tradeoff among crucial factors such as the ADL, the reward of alert attention, and the impact of alert inattention.

Finally, we propose an algorithm to learn the adaptive Attention Management (AM) strategy based on the operator's alert inspection outcomes. We present several case studies based on the simulation of different IDoS attacks and alert inspecting processes. The numerical results show that the proposed optimal AM strategy outperforms the default strategy and can effectively reduce the IDoS risk by as much as 20%. The strategy is also resilient to a large range of cost variations, attack frequencies, and human attention capacities. We have observed the phenomenon of *attentional risk equivalency*, which states that the deviation from the optimal to sub-optimal strategies for some category labels can reduce the risk under the default strategy to approximately the same level. The results also corroborate that RADAMS can adapt to different category labels to strike a balance of quantity (i.e., inspect more alerts) and quality (i.e., complete alert responses to dismiss feints and escalate real attacks). We identify the *attacker's dilemma* where destructive IDoS attacks induce unbearable costs to the attacker. We also identify the *half-truth attack strategy* as the optimal IDoS attack strategy when feints are generated at a high cost.

1.1. Contribution, notations, and organization of the paper

Our main contributions are fourfold. First, we have formally defined a new type of attentional attacks called IDoS attacks. Second, we propose a consolidated alert and attention management strategy that is explicitly aware of human cognition limitations to defend against IDoS attacks. Third, we provide theoretical underpinnings of RADAMS under IDoS attacks and propose a learning algorithm to implement RADAMS in real time. Fourth, we present comprehensive case studies to demonstrate the effectiveness, adaptiveness, robustness, and resilience of the proposed assistive strategies.

The rest of the paper is organized as follows. The related works are presented in Section 2. Sections 3, 4, and 5 introduce the IDoS attack model, the human operator model, and the human-assistive security technology, respectively. We summarize main notations for these three sections in Table 1, 2, and 3, respectively. We analyze the attentional deficiency level and the risk of IDoS attacks in closed form for the class of ambitious operators in Section 6, where the main notations are summarized in Table 4. Section 7 presents a case study of alert inspection under IDoS attacks and the adaptive AM strategies. Section 8 concludes the paper.

2. Related work

2.1. Alert management

Previous works have applied various alert management methods during the alert generation, detection, and response processes to mitigate alert fatigue and enhance cybersecurity, as shown in the following three subsections.

2.1.1. Source management

On the one hand, proactive defense (Huang and Zhu, 2020a) and deception techniques, including honeypots (Huang and Zhu, 2019; 2020b) and moving target defense (Jajodia et al., 2011), have managed to reduce alerts at the outset by deterring, delaying, and preventing attacks. On the other hand, previous works have designed incentive mechanisms (e.g., Casey et al. (2016); Liu et al. (2009)) and information mechanisms (e.g., Huang and Zhu (2021b, 2022)) to enhance insiders' compliance, reduce users' misbehavior, and consequently reduce false positives.

2.1.2. Detection management

A rich literature has attempted to develop detection systems capable of reducing false positives while maintaining the ability to detect malicious behaviors. Methods include statistical analysis (Spathoulas and Katsikas, 2010), fuzzy inference (Elshoush and Osman, 2010), kernel density estimation (Su et al., 2019), and machine learning approaches (Bouzar-Benlabiod et al., 2020; Goeschel, 2016; Ohta et al., 2008; Pietraszek and Tanner, 2005). Alert aggregation and correlation methods (Salah et al., 2013) have also been applied to dismiss repeated and innocuous alerts and generate alerts of system-level threat information. Recently, the authors in Bryant and Saiedian (2020) have implemented a hybrid kill-chain based classification model to boost detection rates, improve alert description, and lower the number of false-positive alerts. There is a rich literature on alert filtering and selection, and we refer the readers to Cotroneo et al. (2017) for the empirical analysis and validation of these state-of-the-art filtering techniques.

2.1.3. Response management

Despite the significant advances in alert reduction methods introduced in Section 2.1.1 and 2.1.2, the demand for alert inspection still exceeds the operators' capacity. To this end, researchers have developed various alert triage and prioritization approaches that can be classified into the following three categories.

The first category ranks alerts based on rules. These rules can be generated through fuzzy logic (Alsubhi et al., 2012; Newcomb et al., 2016) and attack graphs (Noel and Jajodia, 2008). Many works have attempted to learn from security experts and automate the process of mining triage rules out of cybersecurity analysts' operation traces (Zhong et al., 2016; 2018b). The second category assigns scores to alerts and quantitatively optimizes the

Table 2
Summary of notations in Section 4.

Variable	Meaning
$w_{FE}, w_{RE}, w_{UN}, w_{NI}$	Alert dismissal, alert escalation, uninspected alerts, and inadequate alert response.
$w^k \in \mathcal{W}$	Operator's alert response at attack stage k.
$\kappa_{SW}^{\Delta k}(s^{k+\Delta k}\|s^k)$	Operator's default switching probability.
$D_{\max}(s^k) \in \mathbb{R}^+$	Maximum Allowable Delay (MAD) for responding to alerts of category label $s^k \in \mathcal{S}$.
$t_{AoI}^k = t - t^k$	k-th alert's Age of Information (AoI).
$y_{EL} \in \mathcal{Y}_{EL}$	Operator's expertise level.
$\bar{d}(y_{EL}, s^k, \theta^k, \phi^k) \in \mathbb{R}^+$	Average inspection time to reach a complete alert response w_{FE} or w_{RE}.
$d(y_{EL}, s^k, \theta^k, \phi^k)$	Actual Inspection Time Needed (AITN).
$n^t \in \mathbb{Z}^{0+}$	Number of alerts that arrive during the current inspection up to time $t \in [0, \infty)$.
$y_{SL}^t = f_{SL}(n^t) \in \mathbb{R}^+$	Operator's stress level at time t.
$\omega^t = f_{LOE}(y_{SL}^t) \in [0, 1]$	Operator's Level of Operational Efficiency (LOE) at time t.
$\tilde{n}(y_{EL}, s^k) \in \mathbb{R}^{0+}$	Attention threshold.
$\tilde{\omega}^{t_1,t_2} := \int_{t_1}^{t_2} \omega^t dt$	Effective Inspection Time (EIT) during inspection time $[t_1, t_2]$.
$p_{SP}(y_{EL}, s^k, \theta^k, \phi^k)$	Probability of a complete response.

Table 3
Summary of notations in Section 5.

Variable	Meaning
$I_h \in \mathbb{Z}^{0+}$, $t^{I_h} \in [0, \infty)$	Index and time of the alert under the h-th inspection (i.e., inspection stage $h \in \mathbb{Z}^{0+}$).
$a_m \in \mathcal{A}$	Attention management (AM) strategy of period $m \in \mathbb{Z}^+$.
$a^h \in \mathcal{A}$	AM action at inspection stage $h \in \mathbb{Z}^{0+}$.
$\tilde{\kappa}_{SW}^{I_{h+1}-I_h, a^h}(s^{I_{h+1}}\|s^{I_h})$	Operator's switching probability under a^h.
$\tilde{c}(w^k, s^k) \in \mathbb{R}$	Stage cost.
$c(s^{I_h}, a^h) \in \mathbb{R}$	Expected Consolidated Cost (ECoC).
$\tilde{c}(s^{I_h}, a^h) \in \mathbb{R}$	Consolidated Cost (CoC).
$\sigma^0, \sigma^* \in \Sigma$	Default and optimal AM strategy.

Table 4
Summary of Notations in Section 6.

Variable	Meaning
$p_{UN}(s^{I_h}, a^h)$	Attentional Deficiency Level (ADL).
$\beta > 0$	Poisson arrival rate.
$\tilde{z}$	PDF of Erlang distribution with shape $m+1$ and rate β.
$p_{SD}^h(w^{I_h}\|s^{I_h}, a^h; \theta^{I_h}, \phi^{I_h})$	Probability that the operator makes alert response w^{I_h} at inspection stage h.
$\lambda(s^{I_h}, m, \phi^{I_h})$	Expected reward of a complete alert response.

alert triage process by minimizing the cyber risk. The score can be computed through a causal dependency graph of an alert event (Hassan et al., 2019), game-theoretic approaches (Laszka et al., 2017), and the Quantitative Value Function (QVF) hierarchy process (Shah et al., 2019a). The authors in Ganesan et al. (2016); Shah et al. (2019a) further incorporate organization-specific factors and constraints into the design of the optimal alert selection. The third category relies on data and learning methods. Supervised learning (Bierma et al., 2016; Renners et al., 2017), deep learning (Aminanto et al., 2020; McElwee et al., 2017), and adversarial reinforcement learning Tong et al. (2020) are used to prioritize alerts. The authors in Zhong et al. (2018a) have developed a triage operation retrieval system to provide novice analysts with on-the-job suggestions using relevant data triage operations conducted by senior analysts.

The above three categories of *rule-based, risk-aware*, and *data-driven* alert triage methods rank alerts based on their contextual information and organizational factors. Our *human-centered* approach generalizes these classical alert triage approaches by explicitly modeling the attentional behaviors of human operators and selecting alerts based on human cognitive capacity.

2.2. Feint attacks and human attentional models

Feints have been widely studied in sports, military, and biology (Project, 2017). They are recently used to attack detection systems Corona et al. (2013). In particular, the authors in Mutz et al. (2003); Patton et al. (2001) have developed tools that can generate false positives by matching detection signatures. The tools are tested on SNORT (Roesch et al., 1999), and the empirical results verify the feasibility of feint attacks on detection systems. Compared to these empirical practices of feint attacks that exploit the vulnerability of detection systems, we focus on the attentional vulnerabilities and the impact of feints on human operators. Moreover, we abstract models to formally characterize cyber feint attacks, quantify the risk, and develop human-assistive security technologies.

We can classify human vulnerabilities into *acquired* vulnerabilities (e.g., lack of security awareness and noncompliance) and *innate* ones (e.g., bounded attention and rationality) based on whether they can be mitigated through short-term training and security rules. Many works (e.g., Casey et al. (2016); Huang and Zhu (2021b); Wang et al. (2021)) have emphasized the urgency and necessity to reduce acquired human vulnerability and proposed human-assistive strategies. However, few works have focused on mitigation strategies for innate vulnerabilities. Visual support systems have been used for rapid cyber event triage (Miserendino et al., 2017) and alert investigations (Franklin et al., 2017), and eye-tracking data have been incorporated to enhance attention for phishing identification (Huang et al., 2022). The authors in Sundaramurthy et al. (2015) perform an anthropological study in a corporate SOC to model and mitigate security analyst burnout. These works lay the foundations of empirical solutions to mitigate human attentional vulnerabilities. Our work combines real-time human behavioral and decision data with the well-identified human factors to enable quantitative characterizations of the empirical relationship such as the Yerkes-Dodson law (Yerkes et al.,

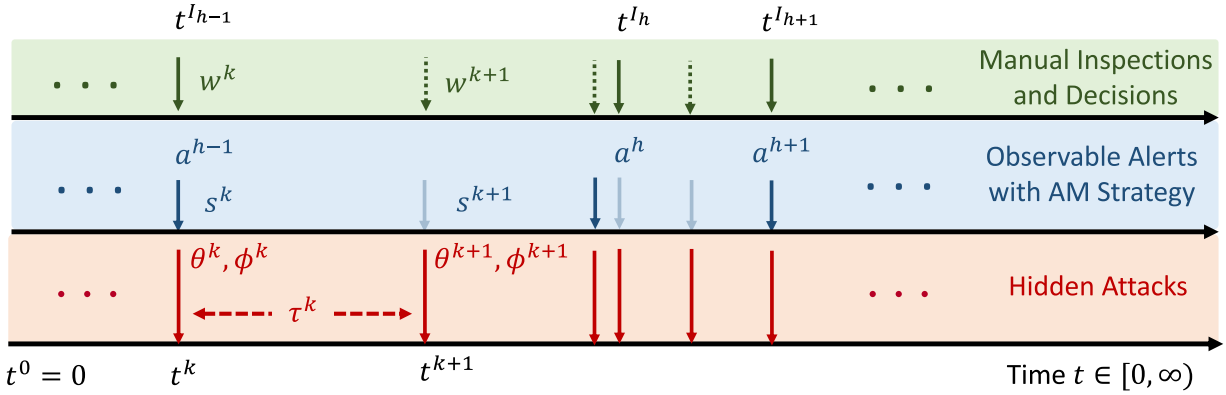


Fig. 2. The timelines of an IDoS attack, alerts under AM strategies, and manual inspections are depicted in red, blue, and green, respectively. The inspection stage $h \in \mathbb{Z}^{0+}$ is equivalent to the attack stage $I_h \in \mathbb{Z}^{0+}$. The red arrows represent the sequential arrivals of feints and real attacks. The semi-transparent blue and the dashed green arrows represent the de-emphasized alerts and the alerts without inspections, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

1908). The learning-based method for attention management also makes our human-assistive technology adaptive and transferable to various human-technical systems.

3. IDoS attacks and sequential alert arrivals

As illustrated in the first column of Fig. 1, after the IDoS attacker has generated feint and real attacks, the detection system monitors the readings from physical layers and log files from cyber layers and generates alerts according to the *generation rules*. Then, the alerts are sent to the SOC and a triage system automatically generates their category labels (e.g., the alerts' criticality) based on the *technical-level triage rules*. The rules for alert generation and triage are pre-defined and their designs are not the focus of this work.

3.1. Feint and real attacks of heterogeneous targets

After the essential preparation stages (e.g., initial intrusion, privilege escalation, and lateral movement), IDoS attacks identify the vulnerable assets as the attack targets and gain control of the ICS to launch feint and real attacks sequentially, as illustrated by the solid red arrows in Fig. 2. With a deliberate goal of triggering alerts, feint attacks require fewer resources to craft. Although feints have limited impacts on the target system, they aggravate the alert fatigue by depleting human attention resources and preventing human operators from a timely response to real attacks. For example, the attacker can attempt to access a database with wrong credentials intentionally, and in the meantime, gradually changes the temperature of the reactor of a nuclear power plant. The repeated log-in attempts trigger an excessive number of alerts so that the overloaded human operators fail to pay sustained attention and respond timely to the sensor alerts of the temperature deviation.

We denote feint and real attacks as θ_{FE} and θ_{RE}, respectively, where $\Theta := \{\theta_{FE}, \theta_{RE}\}$ is the set of attacks' types. Each feint or real attack can target cyber assets (e.g., servers, databases, and workstations) or physical assets (e.g., sensors of pressure, temperature, and flow rate) in the ICS. We define Φ as the set of the potential attack targets. The stochastic arrival of these attacks is modeled as a Markov renewal process where $t^k, k \in \mathbb{Z}^{0+}$, is the time of the k-th arrival. We refer to the k-th attack equivalently as the attack at *attack stage* $k \in \mathbb{Z}^{0+}$ and let $\theta^k \in \Theta$ and $\phi^k \in \Phi$ be the attack's type and target at attack stage $k \in \mathbb{Z}^{0+}$, respectively. Define $\kappa_{AT} \in \mathcal{K}_{AT} : \Theta \times \Phi \times \Theta \times \Phi \mapsto [0, 1]$ as the transition kernel, where $\kappa_{AT}(\theta^{k+1}, \phi^{k+1} | \theta^k, \phi^k)$ denotes the probability that the $(k+1)$-th attack has type $\theta^{k+1} \in \Theta$ and target $\phi^{k+1} \in \Phi$ when the k-th attack has type $\theta^k \in \Theta$ and target $\phi^k \in \Phi$. The inter-arrival time $\tau^k := t^{k+1} - t^k$ is a continuous random variable with support $[0, \infty)$ and Probability Density Function (PDF) $z \in \mathcal{Z} : \Theta \times \Phi \times \Theta \times \Phi \mapsto \mathbb{R}^{0+}$, where $z(t | \theta^{k+1}, \phi^{k+1}, \theta^k, \phi^k)$ is the probability that the inter-arrival time is t when the attacks' types and targets at attack stage k and $k+1$ are θ^k, ϕ^k and θ^{k+1}, ϕ^{k+1}, respectively. The values of $\kappa_{AT} \in \mathcal{K}_{AT}$ and $z \in \mathcal{Z}$ are unknown to human operators and the designer of RADAMS. Attackers can adapt κ_{AT} and z to different ICSs and alert inspection schemes to achieve the attack goals. We formally define IDoS attacks in Definition 1.

Definition 1 (IDoS Attacks)**.** An IDoS attack is a sequence of feint and real attacks of heterogeneous targets, which can be characterized by the 4-tuple $(\Theta, \Phi, \mathcal{K}_{AT}, \mathcal{Z})$.

3.2. Technical-level alert triage and system-level metrics

The alerts triggered by IDoS attacks contain *device-level* contextual information, including the software version, hardware parameters, existing vulnerabilities, and security patches. The alert triage process consists of rules that map the device-level information to *system-level* metrics, which helps human operators make timely responses. Some essential metrics are listed as follows.

- **Source** $s_{SO} \in \mathcal{S}_{SO}$: The ICS sensors or the cyber assets that the alerts are associated with.
- **Time Sensitivity** $s_{TS} \in \mathcal{S}_{TS}$: The length of time that the potential attack needs to achieve its attack goals.
- **Complexity** $s_{CO} \in \mathcal{S}_{CO}$: The degree of effort that a human operator takes to inspect the alert.
- **Susceptibility** $s_{SU} \in \mathcal{S}_{SU}$: The likelihood that the attack succeeds and inflicts damage on the protected system.
- **Criticality** $s_{CR} \in \mathcal{S}_{CR}$: The consequence or the impact of the attack's damage.

These alert metrics are observable to the human operators and the RADAMS designer and form the *category label* of an alert. We define the category label associated with the k-th alert as $s^k := (s^k_{SO}, s^k_{TS}, s^k_{CO}, s^k_{SU}, s^k_{CR}) \in \mathcal{S}$, where $\mathcal{S} := \mathcal{S}_{SO} \times \mathcal{S}_{TS} \times \mathcal{S}_{CO} \times \mathcal{S}_{SU} \times \mathcal{S}_{CR}$. The joint set $\mathcal{S}$ can be adapted to suit the organization's needs in the security practice. For example, we have $\mathcal{S}_{TS} = \emptyset$ if time sensitivity is unavailable or unimportant.

The technical-level alert triage process establishes a stochastic connection between the hidden types and targets of the IDoS attacks and the observable category labels of the associated alerts. Let $o(s^k | \theta^k, \phi^k)$ be the probability of obtaining category label $s^k \in \mathcal{S}$, when the associated attack has type $\theta^k \in \Theta$

and target $\phi^k \in \Phi$. The revelation kernel o reflects the quality of the alert triage. For example, feints with lightweight resource consumption usually have a limited impact. Thus, a high-quality triage process should classify the associated alert as low criticality with a high probability. Letting $b(\theta^k, \phi^k)$ denote the probability that the k-th attack has type θ^k and target ϕ^k at the steady-state, we can compute the steady-state distribution b in closed form based on κ_{AT}. Then, the transition of category labels at different attack stages is also Markov and is represented by $\kappa_{CL} \in \mathcal{K}_{CL} : \mathcal{S} \times \mathcal{S} \mapsto [0, 1]$. We can compute $\kappa_{CL} = \frac{\Pr(s^{k+1}, s^k)}{\sum_{s^{k+1} \in \mathcal{S}} \Pr(s^{k+1}, s^k)}$ based on κ_{AT}, o, b, where $\Pr(s^{k+1}, s^k) = \sum_{\theta^k, \theta^{k+1} \in \Theta} \sum_{\phi^k, \phi^{k+1} \in \Phi} \kappa_{AT}(\theta^{k+1}, \phi^{k+1} | \theta^k, \phi^k) o(s^k | \theta^k, \phi^k) o(s^{k+1} | \theta^{k+1}, \phi^{k+1}) b(\theta^k, \phi^k)$. In this work, we focus on the case where the detection system introduces the same delay between attacks and their triggered alerts. Since the sequences of attacks and alerts have a one-to-one mapping, we can consider zero delay time without loss of generality. Hence, the sequence of alerts associated with an IDoS attack $(\Theta, \Phi, \mathcal{K}_{AT}, \mathcal{Z})$ is also a Markov renewal process characterized by the 3-tuple $(\mathcal{S}, \mathcal{K}_{CL}, \mathcal{Z})$.

4. Human attention model under IDoS attacks

An SOC typically adopts a hierarchical alert analysis (Zimmerman, 2014). The attention model in this section applies to the tier-1 SOC analysts, or the operators, who are in charge of monitoring, inspecting, and responding to alerts in real time. As illustrated by the green box in Fig. 1, the operators choose to inspect certain alerts, dismiss the feints, and escalate the real attacks to tier-2 SOC analysts for in-depth analysis. The in-depth analysis can last hours to months, during which the tier-2 analysts correlate incidents from different assets in the ICS over long periods to build threat intelligence and analyze the impact. The threat intelligence is then incorporated to form and update the generation rules of the detection system and triage rules of the triage process.

4.1. Alert responses

Due to the high volume of alerts and the potential short-term surge arrivals, human operators cannot inspect all alerts in real time. The uninspected alerts receive an alert response w_{NI}. Whether the operator chooses to inspect an alert depends on the switching probability in Section 4.2.

When the operator inspects an alert, he can be distracted by the arrival of new alerts and switch to newly-arrived alerts without completing the current inspection. We elaborate on the attention dynamics in Section 4.3. The alert with incomplete inspection is labeled by w_{UN}. Besides the insufficient inspection time, the operator's cognitive capacity constraint can also prevent him from determining whether the alert is triggered by a feint or a real attack. In this work, we consider prudent operators. When they cannot determine the attack's type after a full inspection, the associated alert is labeled as w_{UN}, as shown in the green flowchart of Fig. 1. We elaborate on how the insufficient inspection time and the operator's cognitive capacity constraint lead to w_{UN}, i.e., referred to as the *inadequate alert response*, in Section 4.4. The alerts labeled as w_{NI} and w_{UN} are ranked and queued up for delayed inspections at later stages.

When the operator successfully completes the alert inspection with a deterministic decision, he either dismisses the alert (denoted by w_{FE}) or escalates the alert to tier-2 SOC analysts for in-depth analysis (denoted by w_{RE}), as shown in Fig. 1. We use $w^k \in \mathcal{W} := \{w_{FE}, w_{RE}, w_{UN}, w_{NI}\}$ to denote the operator's response to the alert at attack stage $k \in \mathbb{Z}^{0+}$. We can extend the set $\mathcal{W}$ to suit the organization's security practice. For example, some organizations let the operators report their estimations and confidence levels concerning incomplete alert inspection, i.e., divide the label w_{UN} into finer subcategories. Then at later stages, the delayed inspection can prioritize the alerts based on the estimations and confidence levels.

4.2. Probabilistic switches within allowable delay

Alerts are monitored in real time when they arrive. When the category label of the new alert indicates higher time sensitivity, susceptibility, or criticality, the operator can delay the current inspection (i.e., label the alert under inspection as w_{UN}) and switch to inspect the new alert. We denote $\kappa_{SW}^{\Delta k}(s^{k+\Delta k} | s^k)$ as the operator's default switching probability when the previous alert at attack stage k and the new alert at stage $k + \Delta k$, $\Delta k \in \mathbb{Z}^+$, have category label $s^k \in \mathcal{S}$ and $s^{k+\Delta k} \in \mathcal{S}$, respectively. As a probability measure,

$$\sum_{\Delta k=1}^{\infty} \sum_{s^{k+\Delta k} \in \mathcal{S}} \kappa_{SW}^{\Delta k}(s^{k+\Delta k} | s^k) \equiv 1, \forall k \in \mathbb{Z}^{0+}, \forall s^k \in \mathcal{S}. \tag{1}$$

Since the operator cannot observe the attack's hidden type and hidden target, the switching probability $\kappa_{SW}^{\Delta k}$ is independent of θ^k, ϕ^k and θ^{k+1}, ϕ^{k+1}. The switching probability depends on the time that the operator has already spent on the current inspection. For example, an operator becomes less likely to switch after spending a long time inspecting an alert of low criticality or beyond his capacity, which can lead to the Sunk Cost Fallacy (SCF).

We denote $D_{\max}(s^k) \in \mathbb{R}^+$ as the Maximum Allowable Delay (MAD) for alerts of category label $s^k \in \mathcal{S}$. At time $t \geq t^k$, the k-th alert's Age of Information (AoI) (Yates et al., 2021) is defined as $t_{AoI}^k := t - t^k$. This work focuses on time-critical ICSs where a defensive response for the k-th alert of category label $s^k \in \mathcal{S}$ is only effective if the alert's AoI is within the MAD, i.e., $t_{AoI}^k \leq D_{\max}(s^k)$. Therefore, the operator will be reminded when an alert's AoI exceeds the MAD so that he can switch to monitor and inspect new alerts. The MAD and the reminder scheme help mitigate the SCF when the operators are occupied with old alerts and miss the chance to monitor and inspect new alerts in real time.

4.3. Attentional factors

We identify the following human and environmental factors affecting operators' alert inspection and response processes.

- The operator's expertise level denoted by $y_{EL} \in \mathcal{Y}_{EL}$.
- The k-th alert's category label $s^k \in \mathcal{S}$.
- The k-th attack's type θ^k and target ϕ^k.
- The operator's stress level $y_{SL}^t \in \mathbb{R}^+$, which changes with time t as new alerts arrive.

The first three factors are the static attributes of the analyst, the alert, and the IDoS attack, respectively. They determine the average inspection time, denoted by $\bar{d}(y_{EL}, s^k, \theta^k, \phi^k) \in \mathbb{R}^+$, to reach a *complete response* w_{FE} or w_{RE}. For example, if the inspected alert is of low complexity, the operator can reach a complete response in a shorter time. Also, it takes a senior operator less time on average to reach a complete alert response than a junior one does. We use $d(y_{EL}, s^k, \theta^k, \phi^k)$ to represent the Actual Inspection Time Needed (AITN) when the operator is of expertise level y_{EL}, the alert is of category label s^k, and the attack has type θ^k and target ϕ^k. AITN $d(y_{EL}, s^k, \theta^k, \phi^k)$ is a random variable with mean $\bar{d}(y_{EL}, s^k, \theta^k, \phi^k)$.

The fourth factor reflects the temporal aspect of human attention during the inspection process. Evidence has shown that the continuous arrival of the alerts can increase the stress level of human operators (Ancker et al., 2017), and 52% of employees attribute

their mistakes to stress (Tessian, 2020). We denote $n^t \in \mathbb{Z}^{0+}$ as the number of alerts that arrives during the current inspection up to time $t \in [0, \infty)$ and model the operator's stress level y_{SL}^t as an increasing function f_{SL} of n^t, i.e., $y_{SL}^t = f_{SL}(n^t)$. At time $t \in [0, \infty)$, the human operator's Level of Operational Efficiency (LOE), denoted by $\omega^t \in [0, 1]$, is a function f_{LOE} of the stress level y_{SL}^t, i.e.,

$$\omega^t = f_{LOE}(y_{SL}^t) = (f_{LOE} \circ f_{SL})(n^t), \forall t \in [0, \infty). \quad (2)$$

Based on the Yerkes-Dodson law, the function f_{LOE} follows an inverse U-shape that contains the following two regions. In region one, a small number of alerts result in a moderate stress level and allow human operators to inspect the alert efficiently. In region two, the LOE starts to decrease when the number of alerts to inspect is beyond some threshold $\bar{n}(y_{EL}, s^k) \in \mathbb{R}^{0+}$, and the human operator is overloaded. The value of the *attention threshold* $\bar{n}(y_{EL}, s^k)$ depends on the operator's expertise level $y_{EL} \in \mathcal{Y}_{EL}$ and the alert's category label $s^k \in \mathcal{S}$. For example, it requires more (resp. fewer) alerts (i.e., higher (resp. lower) attention threshold) to overload a senior (resp. an inexperienced) operator. We can also adapt the value of $\bar{n}(y_{EL}, s^k)$ to different scenarios. In the extreme case where all alerts are of high complexity and create a heavy cognitive load, we let $\bar{n}(y_{EL}, s^k) = 0, \forall y_{EL} \in \mathcal{Y}_{EL}, s^k \in \mathcal{S}$, and the LOE decreases monotonously with the number of alert arrivals during an inspection.

4.4. Alert responses under time and capacity limitations

After we identify attentioinal factors in Section 4.3, we illustrate their impacts on the operators' alert responses as follows. We define the Effective Inspection Time (EIT) during inspection time $[t_1, t_2]$ as the integration $\tilde{\omega}^{t_1,t_2} := \int_{t_1}^{t_2} \omega^t dt$. When the operator is overloaded and has a low LOE during $[t_1, t_2]$, the EIT $\tilde{\omega}^{t_1,t_2}$ is much shorter than the actual inspection time $t_2 - t_1$.

Suppose that the operator of expertise level y_{EL} inspects the k-th alert for a duration of $[t_1, t_2]$. If the EIT has exceed the AITN $d(y_{EL}, s^k, \theta^k, \phi^k)$, then the operator can reach a complete response w_{FE} or w_{RE} with a high success probability denoted by $p_{SP}(y_{EL}, s^k, \theta^k, \phi^k) \in [0, 1]$. However, when $\tilde{\omega}^{t_1,t_2} < d(y_{EL}, s^k, \theta^k, \phi^k)$, it indicates that the operator has not completed the inspection, and the alert response concerning the k-th alert is $w^k = w_{UN}$. The success probability p_{SP} depends on the operator's capacity to identify attacks' types, which leads to the definition of the capacity gap below.

Definition 2 (Capacity Gap)**.** For an operator of expertise level $y_{EL} \in \mathcal{Y}_{EL}$, we define $p_{CG}(y_{EL}, s^k, \theta^k, \phi^k) := 1 - p_{SP}(y_{EL}, s^k, \theta^k, \phi^k)$ as his capacity gap to inspect an alert with category label $s^k \in \mathcal{S}$, type $\theta^k \in \Theta$, and target $\phi^k \in \Phi$ defined in Section 3.

5. Human-assistive security technology for cognitive-level alert management

As illustrated in Section 4, the frequent arrival of alerts triggered by IDoS attacks can overload the human operator and reduce the LOE and the EIT. To compensate for the human's attentional limitation, we can intentionally make some alerts less noticeable, e.g., without sounds or in a light color, based on their category labels. As illustrated by the blue box in Fig. 1, based on the category labels from the technical-level triage process, RADAMS automatically emphasizes and de-emphasizes alerts, referred to as the cognitive-level alert management, and then presents them to the tier 1 SOC analysts.

5.1. Adaptive attention management strategy

In this work, we focus on the class of AM strategies, denoted by $\mathcal{A} := \{a_m\}_{m \in \{0,1,\cdots,M\}}$, that de-emphasize consecutive alerts. As explained in Section 4.1, the operator can only inspect some alerts in real time. Thus, we use $I_h \in \mathbb{Z}^{0+}$ and $t^{I_h} \in [0, \infty)$ to denote the index and the time of the alert under the h-th inspection; i.e., the inspection stage $h \in \mathbb{Z}^{0+}$ is equivalent to the attack stage $I_h \in \mathbb{Z}^{0+}$. Whenever the operator starts a new inspection at inspection stage $h \in \mathbb{Z}^{0+}$, RADAMS determines the AM action $a^h \in \mathcal{A}$ for the h-th inspection based on the stationary strategy $\sigma \in \Sigma : \mathcal{S} \mapsto \mathcal{A}$ that is adaptive to the category label of the h-th alert. We illustrate the timeline of the manual inspections and the AM strategies in green and blue, respectively, in Fig. 2. The solid and dashed green arrows indicate the inspected and uninspected alerts, respectively. The non-transparent and semi-transparent blue arrows indicate the emphasized and de-emphasized alerts, respectively. At inspection stage h, if $a^h = a_m$, RADAMS will make the next m alerts less noticeable; i.e., the alerts at attack stages $I_h + 1, \cdots, I_h + m$ are de-emphasized. Denote $\bar{\kappa}_{SW}^{I_{h+1}-I_h,a^h}(s^{I_{h+1}}|s^{I_h})$ as the operator's switching probability to these de-emphasized alerts under the AM action $a^h \in \mathcal{A}$. Analogously to (1), the following holds for all $h \in \mathbb{Z}^{0+}$ and $a^h \in \mathcal{A}$, i.e.,

$$\sum_{I_{h+1}=I_h+1}^{\infty} \sum_{s^{I_{h+1}} \in \mathcal{S}} \bar{\kappa}_{SW}^{I_{h+1}-I_h,a^h}(s^{I_{h+1}}|s^{I_h}) \equiv 1, \forall s^{I_h} \in \mathcal{S}. \quad (3)$$

The deliberate de-emphasis on selective alerts brings the following tradeoff. On the one hand, these alerts do not increase the operator's stress level, and the operator can pay sustained attention to the alert under inspection with high LOE and EIT. On the other hand, these alerts do not draw the operator's attention, and the operator is less likely to switch to them during the real-time monitoring and inspections.

Since the operator may switch to inspect a de-emphasized alert with switching probability $\bar{\kappa}_{SW}^{I_{h+1}-I_h,a^h}$ (e.g., the h-inspection in Fig. 2), RADAMS recomputes the AM strategy and implements the new strategy whenever the operator has started to inspect a new alert. Although the operator can switch unpredictably, Proposition 1 shows that the transition of the inspected alerts' category labels is Markov.

Proposition 1. *For a stationary AM strategy $\sigma \in \Sigma$, the set of random variables $(\mathbf{S}^{\mathbf{I}_h}, \mathbf{T}^{\mathbf{I}_h})_{h \in \mathbb{Z}^{0+}}$ is a Markov renewal process.*

Proof. The sketch of the proof includes two steps. First, we prove that the state transition from s^{I_h} to $s^{I_{h+1}}$ is Markov for all $h \in \mathbb{Z}^{0+}$. Due to the uncertainty of switching in inspection, the transition stage $\mathbf{I}_{h+1}$ is also a random variable for all $h \in \mathbb{Z}^{0+}$, and we can represent the transition probability as

$$\Pr(\mathbf{S}^{\mathbf{I}_{h+1}} = s^{I_{h+1}}|s^{I_h}) = \sum_{l=1}^{\infty} \Pr(\mathbf{I}_{h+1} = I_h + l) \cdot \Pr(\mathbf{S}^{\mathbf{I}_{h+1}} = s^{I_{h+1}}|s^{I_h}),$$

where $\Pr(\mathbf{I}_{h+1} = I_h + l)$ is the probability that the $(h+1)$-th inspection happens at attack stage $I_h + l$. The term $\Pr(\mathbf{S}^{\mathbf{I}_{h+1}} = s^{I_{h+1}}|s^{I_h})$ is Markov and can be computed based on κ_{CL}. The term $\Pr(\mathbf{I}_{h+1} = I_h + l)$ depends on $d(y_{EL}, s^{I_h+l'}, \theta^{I_h+l'}, \phi^{I_h+l'})$, $\kappa_{SW}^{l'}$, $\bar{\kappa}_{SW}^{l'}$, $\tau^{l'}$, for all $l' \in \{1, \cdots, l\}$. Since $s^{I_h+l'}, \theta^{I_h+l'}, \phi^{I_h+l'}, l' \in \{1, \cdots, l\}$, are all stochastically related to s^{I_h} and $s^{I_{h+1}}$ based on o, κ_{AT} and κ_{CL}, the term $\Pr(\mathbf{I}_{h+1} = I_h + l)$ depends on s^{I_h} and $s^{I_{h+1}}$ for all $l \in \mathbb{Z}^+$.

Then, we show that the distribution of the inter-arrival time $\tau_{IN}^{\mathbf{I}_h,m} := \mathbf{T}^{\mathbf{I}_{h+1}} - \mathbf{T}^{\mathbf{I}_h}$ only depends on s^{I_h} and $s^{I_{h+1}}$. Analogously, the cumulative distribution function of $\tau_{IN}^{\mathbf{I}_h,m}$ is

$$\Pr(\tau_{IN}^{\mathbf{I}_h,m} \le t) = \textstyle\sum_{l=1}^{\infty} \Pr(\mathbf{I}_{h+1} = I_h + l) \cdot \Pr(\tau_{IN}^{I_h,m} \le t),$$

and hence we arrive at the Markov property. □

5.2. Stage cost and expected cumulative cost

For each alert at attack stage $k \in \mathbb{Z}^{0+}$, RADAMS assigns a stage cost $\bar{c}(w^k, s^k) \in \mathbb{R}$ to evaluate the outcomes of alert response $w^k \in$

$\mathcal{W}$ under the category label $s^k \in \mathcal{S}$. The value of the cost varies under different scenarios. In this work, we can estimate it using the salary of SOC analysts and the estimated loss of the associated attack. For example, $\bar{c}(w_{UN}, s^{I_h})$ and $\bar{c}(w_{NI}, s^{I_h})$ are positive costs as those alerts without a complete response incur additional workloads. The delayed inspections also expose the organization to the threats of time-sensitive attacks. On the other hand, $\bar{c}(w_{FE}, s^{I_h})$ and $\bar{c}(w_{RE}, s^{I_h})$ are negative costs because the alerts with complete alert response w_{FE} and w_{RE} reduce the workload of tier 2 SOC analysts and enable them to obtain threat intelligence.

When the operator starts a new inspection at inspection stage $h+1$, RADAMS will evaluate the effectiveness of the AM strategy for the h-th inspection. The performance evaluation is reflected by the Expected Consolidated Cost (ECoC) $c : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ at each inspection stage $h \in \mathbb{Z}^{0+}$. We denote the realization of $c(s^{I_h}, a^h)$ as the Consolidated Cost (CoC) $\tilde{c}^{I_h}(s^{I_h}, a^h)$. Since the AM strategy σ at each inspection stage can affect the future human inspection process and the alert responses, we define the Expected Cumulative Cost (ECuC) $u(s^{I_h}, \sigma) := \sum_{h=0}^{\infty} \gamma^h c(s^{I_h}, \sigma(s^{I_h}))$ under adaptive strategy $\sigma \in \Sigma$ as the long-term performance measure. The goal of the assistive technology is to design the optimal adaptive strategy $\sigma^* \in \Sigma$ that minimizes the ECuC u under the presented IDoS attack based on the category label $s^{I_h} \in \mathcal{S}$ at each inspection stage h. We define $v^*(s^{I_h}) := \min_{\sigma \in \Sigma} u(s^{I_h}, \sigma)$ as the optimal ECuC when the category label is $s^{I_h} \in \mathcal{S}$. We refer to the *default AM strategy* $\sigma^0 \in \Sigma$ as the one when no AM action is applied under all category labels, i.e., $\sigma^0(s^{I_h}) = a_0, \forall s^{I_h} \in \mathcal{S}$.

5.3. Reinforcement learning

Due to the absence of the following exact model parameters, RADAMS has to learn the optimal AM strategy $\sigma^* \in \Sigma$ based on the operator's alert responses in real time.

- Parameters of the IDoS attack model (e.g., κ_{AT} and z) and the alert generation model (e.g., o) in Section 3.
- Parameters of the human attention model (e.g., f_{LOE} and f_{SI}), inspection model (e.g., $\kappa_{SW}^{\Delta k}$, $\bar{\kappa}_{SW}^{I_{h+1}-I_h, a^h}$, and d), and alert response model (e.g., y_{EL} and p_{SP}) in Section 4.

Define $Q^h(s^{I_h}, a^h)$ as the estimated ECuC during the h-th inspection when the category label is $s^{I_h} \in \mathcal{S}$ and the AM action is a^h. Based on Proposition 1, the state transition is Markov, which enables Q-learning as follows.

$$Q^{h+1}(s^{I_h}, a^h) := (1-\alpha^h(s^{I_h}, a^h))Q^h(s^{I_h}, a^h) + \alpha^h(s^{I_h}, a^h)[\tilde{c}^{I_h}(s^{I_h}, a^h) + \gamma \min_{a' \in \mathcal{A}} Q^h(s^{I_{h+1}}, a')], \quad (4)$$

where s^{I_h} and $s^{I_{h+1}}$ are the observed category labels of the alerts at the attack stage I_h and I_{h+1}, respectively. When the learning rate $\alpha^h(s^{I_h}, a^h) \in (0, 1)$ satisfies $\sum_{h=0}^{\infty} \alpha^h(s^{I_h}, a^h) = \infty, \sum_{h=0}^{\infty} (\alpha^h(s^{I_h}, a^h))^2 < \infty, \forall s^{I_h} \in \mathcal{S}, \forall a^h \in \mathcal{A}$, and all state-action pairs are explored infinitely, $\min_{a' \in \mathcal{A}} Q^h(s^{I_h}, a')$ converges to the optimal ECuC $v^*(s^{I_h})$ with probability 1 as $h \to \infty$. At each inspection stage $h \in \mathbb{Z}^{0+}$, RADAMS selects AM strategy $a^h \in \mathcal{A}$ based on the ϵ-greedy policy; i.e., RADAMS chooses a random action with a small probability $\epsilon \in [0, 1]$, and the optimal action $arg\min_{a' \in \mathcal{A}} Q^h(s^{I_h}, a')$ with probability $1 - \epsilon$.

We present the algorithm to learn the adaptive AM strategy based on the operator's real-time alert monitoring and inspection process in Algorithm 1.

Each simulation run corresponds to the operator's work shift of 24 hours at the SOC. Since the SOC can receive over 10 thousand of alerts in each work shift, we can use infinite horizon to approximate the total number of attack stages $K > 10,000$. Whenever the operator starts to inspect a new alert at inspection stage

Algorithm 1: Algorithm to Learn the Adaptive AM strategy based on the Operator's Real-Time Alert Inspection.

1 **Input** K: The total number of attack stages;
2 **Initialize**The operator starts the h-th inspection under AM action $a^h \in \mathcal{A}$; $I_h = k_0$; $\tilde{c}^{I_h}(s^{I_h}, a^h) = 0$;
3 **for** $k \leftarrow k_0 + 1$ **to** K **do**
4 **if** *The operator has finished the I_h-th alert (i.e., EIT > AITN),* **then**
5 **if** *Capable (i.e., rand $\le p_{SP}(y_{EL}, s^k, \theta^k, \phi^k)$)* **then**
6 Dismiss (i.e., $w^{I_h} = w_{FE}$) or escalate (i.e., $w^{I_h} = w_{RE}$) the I_h-th alert;
7 **else**
8 Queue up the I_h-th alert, i.e., $w^{I_h} = w_{UN}$;
9 **end**
10 $\tilde{c}^{I_h}(s^{I_h}, a^h) = \tilde{c}^{I_h}(s^{I_h}, a^h) + \bar{c}(w^{I_h}, s^{I_h})$;
11 $I_{h+1} \leftarrow k$; The operator starts to inspect the k-th alert with category label $s^{I_{h+1}}$;
12 Update $Q^{h+1}(s^{I_h}, a^h)$ via (**??**) and obtain the AM action a^{h+1} by ϵ-greedy policy;
13 $\tilde{c}^{h+1}(s^{I_{h+1}}, a^{h+1}) = 0$; $h \leftarrow h + 1$;
14 **else**
15 **if** *The operator chooses to switch* ***or*** *The MAD is reached, i.e., $t^k - t^{I_h} \ge D_{\max}(s^{I_h})$* **then**
16 Queue up the I_h-th alert (i.e., $w^{I_h} = w_{UN}$);
17 $\tilde{c}^{I_h}(s^{I_h}, a^h) = \tilde{c}^{I_h}(s^{I_h}, a^h) + \bar{c}(w_{UN}, s^{I_h})$;
18 $I_{h+1} \leftarrow k$; The operator starts to inspect the k-th alert with category label $s^{I_{h+1}}$;
19 Update $Q^{h+1}(s^{I_h}, a^h)$ via (**??**) and obtain the AM action a^{h+1} by ϵ-greedy policy;
20 $\tilde{c}^{h+1}(s^{I_{h+1}}, a^{h+1}) = 0$; $h \leftarrow h + 1$;
21 **else**
22 The operator continues the inspection of the I_h-th alert with decreased LOE;
23 The k-th alert is queued up for delayed inspection (i.e., $w^k = w_{NI}$);
24 $\tilde{c}^{I_h}(s^{I_h}, a^h) = \tilde{c}^{I_h}(s^{I_h}, a^h) + \bar{c}(w_{NI}, s^k)$;
25 **end**
26 **end**
27 **end**
28 **Return** $Q^h(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A}$;

I_{h+1}, RADAMS applies Q-learning in (4) based on the category label $s^{I_{h+1}}$ of the newly arrived alert and determines the AM action a^{h+1} for the $h+1$ inspection based on the ϵ-greedy policy as shown in lines 12 and 19 of Algorithm 1. The CoC $\tilde{c}^{I_h}(s^{I_h}, a^h)$ of the h-th inspection under the AM action $a^h \in \mathcal{A}$ and the category label s^{I_h} of the inspected alert can be computed iteratively based on the stage cost $\bar{c}(w^k, s^k)$ of the alerts during the attack stage $k \in \{I_h, \cdots, I_{h+1} - 1\}$, as shown in lines 13, 20, and 24 of Algorithm 1.

6. Theoretical analysis

In Section 6, we focus on the class of ambitious operators who attempt to inspect all alerts, i.e., $\kappa_{SW}(s^{k+\Delta k}|s^k) = \mathbf{1}_{\{\Delta k = 1\}}, \forall s^k, s^{k+\Delta k} \in \mathcal{S}, \forall \Delta k \in \mathbb{Z}^+$. To assist this class of operators, the implemented AM action $a_m, m \in \{0, 1, \cdots, M\}$, chooses to make the selected alerts fully unnoticeable. Then, under $a_m \in \mathcal{A}$, the operator at inspection stage h can pay sustained attention to inspect the alert of category label $s^{I_h} \in \mathcal{S}$ for $m+1$ attack stages. Moreover, the operator switches to the new alert at attack stage I_{h+1}, i.e.,

$\sum_{s^{I_h+m+1}\in\mathcal{S}} \bar{\kappa}_{SW}^{I_{h+1}-I_h,a_m}(s^{I_h+m+1}|s^{I_h}) = \mathbf{1}_{\{I_{h+1}-I_h=m+1\}}$. Throughout the section, we omit the variable of the expertise level y_{EL} in functions $d, \bar{d}, p_{SP}$, and p_{CG} because y_{EL} is a constant for all attack stages.

6.1. Security metrics

We propose two security metrics in Definition 3 to evaluate the performance of ambitious operators under IDoS attacks and different AM strategies. The first metric, denoted as $p_{UN}(s^{I_h}, a^h)$, is the probability that the operator chooses w_{UN} during the h-th inspection under the category label $s^{I_h} \in \mathcal{S}$ and AM action $a^h \in \mathcal{A}$. This metric reflects the Attentional Deficiency Level (ADL) of the IDoS attack. For example, as the attackers generate more feints at a higher frequency, the operator is persistently distracted by the new alerts, and it becomes unlikely for him to fully respond to an alert. The ADL $p_{UN}(s^{I_h}, a^h)$ is high in this scenario. We use the ECuC $u(s^{I_h}, \sigma)$ as the second metric that evaluates the *IDoS risk* under the category label $s^{I_h} \in \mathcal{S}$ and the AM strategy $\sigma \in \Sigma$. For both metrics, smaller values are preferred.

Definition 3 (Attentional Deficiency Level and Risk)**.** Under category label $s^{I_h} \in \mathcal{S}$ and the stationary AM strategy $\sigma \in \Sigma$, we define $p_{UN}(s^{I_h}, \sigma(s^{I_h}))$ and $u(s^{I_h}, \sigma)$ as the Attentional Deficiency Level (ADL) and the risk of the IDoS attacks defined in Section 3, respectively.

6.2. Closed-form computations

The Markov renewal process that characterizes the IDoS attack or the associated alert sequence follows a Poisson process when Condition 1 holds.

Condition 1 (Poisson Arrival)**.** The inter-arrival times $\tau^k, \forall k \in \mathbb{Z}^{0+}$, are independent and exponentially distributed random variables with the same arrival rate denoted by $\beta > 0$, i.e., $z(\tau|\theta^{k+1}, \phi^{k+1}, \theta^k, \phi^k) = \beta e^{-\beta\tau}, \tau \in [0, \infty)$ for all $\theta^{k+1}, \theta^k \in \Theta$ and $\phi^{k+1}, \phi^k \in \Phi$.

Recall that random variable $\mathbf{T}_{IN}^{I_h,m}$ represents the inspection time of the I_h-th alert under the AM action $a^h = a_m \in \mathcal{A}$. For the ambitious operators under AM action $a_m \in \mathcal{A}$ at inspection stage h, the next inspection happens at attack stage $I_{h+1} = I_h + m + 1$. Thus, I_{h+1} is no longer a random variable. As a summation of $m+1$ i.i.d. exponential distributed random variables of rate β, $\mathbf{T}_{IN}^{I_h,m}$ follows an *Erlang distribution* denoted by PDF function $\bar{z}$ with shape $m+1$ and and rate $\beta > 0$ when condition 1 holds, i.e., $\bar{z}(\tau) = \frac{\beta^{m+1}\tau^m e^{-\beta\tau}}{m!}, \tau \in [0, \infty)$.

Denote $p_{SD}^h(w^{I_h}|s^{I_h}, a^h; \theta^{I_h}, \phi^{I_h})$ as the probability that the operator makes alert response w^{I_h} at inspection stage h. To obtain a theoretical underpinning, we consider the case where the AITN equals the average inspection time, i.e., $d(s^k, \theta^k, \phi^k) = \bar{d}(s^k, \theta^k, \phi^k)$. Then, the operator under AM action a_m makes a complete alert response (i.e., $w^{I_h} \in \{w_{FE}, w_{RE}\}$) at inspection stage h for category label s^{I_h} if the inspection time $\tau_{IN}^{I_h,m}$ is greater than the AITN. The probability of the above event can be represented as $\int_{d(s^{I_h},\theta^{I_h},\phi^{I_h})}^{\infty} p_{SP}(s^{I_h}, \theta^{I_h}, \phi^{I_h})\bar{z}(\tau)d\tau = p_{SP}(s^{I_h}, \theta^{I_h}, \phi^{I_h}) \cdot \sum_{n=0}^{m} \frac{1}{n!} e^{-\beta d(s^{I_h},\theta^{I_h},\phi^{I_h})}(\beta d(s^{I_h}, \theta^{I_h}, \phi^{I_h}))^n$, which leads to

$$p_{SD}^h(w_{UN}|s^{I_h}, a_m; \theta^{I_h}, \phi^{I_h}) = 1 - p_{SP}(s^{I_h}, \theta^{I_h}, \phi^{I_h}) \cdot \sum_{n=0}^{m} \frac{1}{n!} e^{-\beta d(s^{I_h},\theta^{I_h},\phi^{I_h})}(\beta d(s^{I_h}, \theta^{I_h}, \phi^{I_h}))^n. \tag{5}$$

Then, the ADL $p_{UN}(s^{I_h}, a^h)$ can be computed as

$$\sum_{\theta^{I_h}\in\Theta,\phi^{I_h}\in\Phi} \Pr(\theta^{I_h}, \phi^{I_h}|s^{I_h}) \cdot p_{SD}^h(w_{UN}|s^{I_h}, a^h; \theta^{I_h}, \phi^{I_h}), \tag{6}$$

where the conditional probability $\Pr(\theta^{I_h}, \phi^{I_h}|s^{I_h})$ can be computed via the Bayesian rule, i.e., $\Pr(\theta^{I_h}, \phi^{I_h}|s^{I_h}) = \frac{o(s^{I_h}|\theta^{I_h},\phi^{I_h})b(\theta^{I_h},\phi^{I_h})}{\sum_{\theta^{I_h}\in\Theta,\phi^{I_h}\in\Phi} o(s^{I_h}|\theta^{I_h},\phi^{I_h})b(\theta^{I_h},\phi^{I_h})}$.

We can compute the ECoC $c(s^{I_h}, a_m)$ explicitly as

$$c(s^{I_h}, a_m) = m\bar{c}(w_{NI}, s^{I_h}) + \sum_{\theta^{I_h}\in\Theta,\phi^{I_h}\in\Phi} \Pr(\theta^{I_h}, \phi^{I_h}|s^{I_h}) \cdot \sum_{w^{I_h}\in\mathcal{W}} p_{SD}^h(w^{I_h}|s^{I_h}, a_m; \theta^{I_h}, \phi^{I_h})\bar{c}(w^{I_h}, s^{I_h}). \tag{7}$$

For prudent operators in Section 4.1, we have

$$p_{SD}^h(w_i|s^{I_h}, a^h; \theta_i, \phi^{I_h}) = 1 - p_{SD}^h(w_{UN}|s^{I_h}, a^h; \theta_i, \phi^{I_h}), \tag{8}$$

for all $i \in \{FE, RE\}, s^{I_h} \in \mathcal{S}, a^h \in \mathcal{A}, \phi^{I_h} \in \Phi, h \in \mathbb{Z}^{0+}$. Plugging (8) into (7), we can simplify the ECoC $c(s^{I_h}, a_m)$ as

$$c(s^{I_h}, a_m) = \sum_{\phi^{I_h}\in\Phi} \sum_{i\in\{FE,RE\}} \Pr(\theta_i, \phi^{I_h}|s^{I_h}) \cdot p_{SD}^h(w_i|s^{I_h}, a_m; \theta_i, \phi^{I_h}) \cdot [\bar{c}(w_i, s^{I_h}) - \bar{c}(w_{UN}, s^{I_h})] + m\bar{c}(w_{NI}, s^{I_h}) + \bar{c}(w_{UN}, s^{I_h}). \tag{9}$$

As shown in Proposition 2, the ADL and the risk are monotone function of $\beta d(s^{I_h}, \theta^{I_h}, \phi^{I_h})$ for each AM strategy.

Proposition 2. *If condition 1 holds, then the ADL $p_{UN}(s^{I_h}, \sigma(s^{I_h}))$ and the risk $u(s^{I_h}, \sigma)$ of an IDoS attack under category label $s^{I_h} \in \mathcal{S}$ and AM strategy $\sigma \in \Sigma$ increase in the value of the product $\beta d(s^{I_h}, \theta^{I_h}, \phi^{I_h})$.*

Proof. First, since $p_{SD}^h(w_{UN})$ in (5) increases monotonously with respect to the product $\beta d(s^{I_h}, \theta^{I_h}, \phi^{I_h})$, the values of $p_{SD}^h(w_{FE})$ and $p_{SD}^h(w_{RE})$ in (8) decrease monotonously with respect to the product. Plugging (5) into (6), we obtain that $p_{UN}(s^{I_h}, a_m)$ in (10) under any $a_m \in \mathcal{A}$ and $s^{I_h} \in \mathcal{S}$ is a summation of functions increasing in $\beta d(s^{I_h}, \theta^{I_h}, \phi^{I_h})$.

$$p_{UN}(s^{I_h}, a_m) = \sum_{\phi^{I_h}\in\Phi} \sum_{i\in\{FE,RE\}} \Pr(\theta_i, \phi^{I_h}|s^{I_h})[1 - p_{SP}(s^{I_h}, \theta_i, \phi^{I_h}) \cdot \sum_{n=0}^{m} \frac{1}{n!} e^{-\beta d(s^{I_h},\theta_i,\phi^{I_h})}(\beta d(s^{I_h}, \theta_i, \phi^{I_h}))^n]. \tag{10}$$

Second, since $\bar{c}(w_{FE}, s^{I_h})$ and $\bar{c}(w_{RE}, s^{I_h})$ are negative, and $\bar{c}(w_{UN}, s^{I_h})$ is positive, the ECoC in (9) decreases with $\beta d(s^{I_h}, \theta^{I_h}, \phi^{I_h})$ under any $a_m \in \mathcal{A}$ and $s^{I_h} \in \mathcal{S}$. Then, the risk also decreases with the product, due to the monotonicity of the Bellman operator Bertsekas and Tsitsiklis (1996). □

Remark 1 (Product Principle of Attention (PPoA))**.** On the one hand, as β increases, the feint and real attacks arrive at a higher frequency on average, resulting in a higher demand of attention resources from the human operator. On the other hand, as $d(s^{I_h}, \theta^{I_h}, \phi^{I_h})$ increases, the human operator requires a longer inspection time to determine the attack's type, leading to a lower supply of attention resources. Proposition 2 characterizes the PPoA; i.e., for any stationary AM strategy $\sigma \in \Sigma$, the ADL and the risk of IDoS attacks depend on the product of the supply and demand of attention resources.

6.3. Fundamental limits under AM strategies

Section 6.3 aims to show the fundamental limits of the IDoS attack's ADL, the ECoC, and the risk under different AM strategies. Define the shorthand notation: $\underline{p}(s^{I_h}) := \sum_{\phi^{I_h}\in\Phi} \sum_{i\in\{FE,RE\}} \Pr(\theta_i, \phi^{I_h}|s^{I_h}) p_{CG}(s^{I_h}, \theta_i, \phi^{I_h})$.

Lemma 1. *If Condition 1 holds and $M \to \infty$, then for each $s^{I_h} \in \mathcal{S}$, the ADL $p_{UN}(s^{I_h}, a_m)$ decreases strictly to $\underline{p}(s^{I_h})$ as m increases.*

Proof. Since $\frac{1}{n!} e^{-\beta d(s^{I_h},\theta^{I_h},\phi^{I_h})}(\beta d(s^{I_h}, \theta^{I_h}, \phi^{I_h}))^n > 0$ for all $m \in \{0, \cdots, M\}$, the value of $p_{UN}(s^{I_h}, a_m)$ in

(10) strictly decreases as m increases. Moreover, since $\lim_{m\to\infty}\sum_{n=0}^{m}\frac{1}{n!}e^{-\beta d(s^{I_h},\theta^{I_h},\phi^{I_h})})(\beta d(s^{I_h},\theta^{I_h},\phi^{I_h}))^n = 1$, we have $\min_{m\in\{0,\cdots,M\}} p_{UN}(s^{I_h},a_m) = \underline{p}(s^{I_h})$ for all $s^{I_h}\in\mathcal{S}$. □

Remark 2 (Fundamental Limit of ADL)**.** Lemma 1 characterizes that the minimum ADL under all AM strategies $a_m\in\mathcal{A}$ is $\underline{p}(s^{I_h})$. The value of $\underline{p}(s^{I_h})$ depends on the operator's capacity gap $p_{CG}(s^{I_h},\theta_{FE},\phi^{I_h})$ and the frequency of feint and real attacks with different targets, i.e., $\Pr(\theta^{I_h},\phi^{I_h}|s^{I_h}),\forall\theta^{I_h}\in\Theta,\phi^{I_h}\in\Phi$.

Denote the expected reward of making a complete alert response (i.e., the rewards to dismiss feints and escalate real attacks) as

$$\lambda(s^{I_h},m,\phi^{I_h}) := \sum_{i\in\{FE,RE\}}\bar{c}(w_i,s^{I_h})\cdot\Pr(\theta_i,\phi^{I_h}|s^{I_h}) \cdot p^h_{SP}(s^{I_h},\theta_i,\phi^{I_h})\cdot\left[\sum_{n=0}^{m}\frac{1}{n!}e^{-\beta d(s^{I_h},\theta_i,\phi^{I_h})}(\beta d(s^{I_h},\theta_i,\phi^{I_h}))^n\right].$$

Combining (9) and (10), we can rewrite ECoC as a combination of the following three terms in (11).

$$c(s^{I_h},a_m) = p_{UN}(s^{I_h},a_m)\bar{c}(w_{UN},s^{I_h}) + m\bar{c}(w_{NI},s^{I_h}) + \sum_{\phi^{I_h}\in\Phi}\lambda(s^{I_h},m,\phi^{I_h}). \tag{11}$$

Based on Lemma 1, the first term $p_{UN}(s^{I_h},a_m)\bar{c}(w_{UN},s^{I_h})$ and the third term $\sum_{\phi^{I_h}\in\Phi}\lambda(s^{I_h},m,\phi^{I_h})$ decrease in m, while the second term $m\bar{c}(w_{NI},s^{I_h})$ in (11) increases in m linearly at the rate of $\bar{c}(w_{NI},s^{I_h})$. The tradeoff among the three terms is summarized below.

Remark 3 (Tradeoff among ADL, Reward of Alert Attention, and Impact for Alert Inattention)**.** Based on Lemma 1 and (11), increasing m reduces the ADL and achieves a higher reward of completing the alert response. However, the increase of m also linearly increases the impact for alert inattention represented by $m\bar{c}(w_{NI},s^{I_h})$, the cost of uninspected alerts. Thus, we need to strike a balance among these terms to reduce the IDoS risk.

Define $\lambda_{\min}(s^{I_h},\phi^{I_h}) := \sum_{i\in\{FE,RE\}}\bar{c}(w_i,s^{I_h})\Pr(\theta_i,\phi^{I_h}|s^{I_h})p^h_{SP}(s^{I_h},\theta_i,\phi^{I_h})$, $\lambda^{\epsilon_0}_{\max}(s^{I_h},\phi^{I_h}) := (1-\epsilon_0)\lambda_{\min}(s^{I_h},\phi^{I_h})$, $c_{\min}(s^{I_h}) := \sum_{\phi^{I_h}\in\Phi}\lambda_{\min}(s^{I_h},\phi^{I_h}) + \underline{p}(s^{I_h})\bar{c}(w_{UN},s^{I_h}) + m\bar{c}(w_{NI},s^{I_h})$, and $c^{\epsilon_0}_{\max}(s^{I_h}) := \sum_{\phi^{I_h}\in\Phi}\lambda^{\epsilon_0}_{\max}(s^{I_h},\phi^{I_h}) + [\underline{p}(s^{I_h}) + \epsilon_0(1-\underline{p}(s^{I_h}))]\bar{c}(w_{UN},s^{I_h}) + m\bar{c}(w_{NI},s^{I_h})$.

Proposition 3. *Consider the scenario where Condition 1 holds and $M > \underline{m}(s^{I_h})$. For any $\epsilon_0\in(0,1]$ and $s^{I_h}\in\mathcal{S}$, there exists $\underline{m}(s^{I_h})\in\mathbb{Z}^+$ such that $c(s^{I_h},a_m)\in[c_{\min}(s^{I_h}),c^{\epsilon_0}_{\max}(s^{I_h})],\forall a_m\in\mathcal{A}$, when $m\geq\underline{m}(s^{I_h})$. Moreover, the lower bound $c_{\min}(s^{I_h})$ and the upper bound $c^{\epsilon_0}_{\max}(s^{I_h})$ increase in m linearly at the same rate $\bar{c}(w_{NI},s^{I_h})$.*

Proof. For any $\epsilon_0\in(0,1]$, there exists $\underline{m}(s^{I_h})\in\mathbb{Z}^+$ such that $\sum_{n=0}^{m}\frac{1}{n!}e^{-\beta d(s^{I_h},\theta^{I_h},\phi^{I_h})}(\beta d(s^{I_h},\theta^{I_h},\phi^{I_h}))^n\in[1-\epsilon_0,1]$ when $m\geq\underline{m}(s^{I_h})$. Based on Lemma 1, if $m>\underline{m}(s^{I_h})$, then $p_{UN}(s^{I_h},a_m)\in[\underline{p}(s^{I_h}),\underline{p}(s^{I_h})+\epsilon_0(1-\underline{p}(s^{I_h}))]$. Plugging it into (11), we obtain the results. □

Let $\sigma^{\underline{m}}\in\Sigma$ denote the AM strategy that chooses to deemphasize the next $m\geq\underline{m}(s^{I_h})$ alerts for all category label $s^{I_h}\in\mathcal{S}$. The monotonicity of the Bellman operator Bertsekas and Tsitsiklis (1996) leads to the following corollary.

Corollary 1. *Consider the scenario where Condition 1 holds and $M > \underline{m}(s^{I_h})$. For any $\epsilon_0\in(0,1]$ and $s^{I_h}\in\mathcal{S}$, the upper and lower bounds of the risk $u(s^{I_h},\sigma^{\underline{m}})$ increase in m linearly at the same rate of $\bar{c}(w_{NI},s^{I_h})$.*

Table 5
Benchmark values of the average inter-arrival time $\mu(\theta^k,\theta^{k+1}) = 1/\beta(\theta^k,\theta^{k+1}),\forall\theta^k,\theta^{k+1}\in\Theta$.

Average inter-arrival time from feints to real attacks	6s
Average inter-arrival time from real attacks to feints	10s
Average inter-arrival time between feints	15s
Average inter-arrival time between real attacks	8s

Table 6
Benchmark values of the average inspection time $\bar{d}(s^k_{CR},\theta^k),\forall\theta^k\in\Theta,s^k_{CR}\in\mathcal{S}_{CR}$.

Average time to inspect feints of low criticality	6s
Average time to inspect feints of high criticality	8s
Average time to inspect real attacks of low criticality	15s
Average time to inspect real attacks of high criticality	20s

Remark 4 (Fundamental Limit of ECoC and Risk)**.** Proposition 3 and Corollary 1 show that the maximum length of the deemphasized alerts for any $s^{I_h}\in\mathcal{S}$ should not exceed $\underline{m}(s^{hm})$ to reduce the ECoC and the risk of IDoS attacks.

7. Case study

The following section presents case studies to demonstrate the impact of IDoS attacks on human operators' alert inspections and alert responses, and further illustrate the effectiveness of RADAMS. Throughout the section, we adopt the attention model in Section 4.

7.1. Experiment setup

We consider an IDoS attack targeting either the Programmable Logic Controllers (PLCs) in the physical layer or the data centers in the cyber layer of an ICS. We denote these two targets as ϕ_P and ϕ_C, respectively. They constitute the binary set of attack targets $\Phi=\{\phi_P,\phi_C\}$ defined in Section 3.1. The SOC of the ICS is in charge of monitoring, inspecting, and responding to both the cyber and the physical alerts. We consider two system-level metrics defined in Section 3.2, the source $\mathcal{S}_{SO}=\{s_{SO,P},s_{SO,C}\}$ and the criticality $\mathcal{S}_{CR}=\{s_{CR,L},s_{CR,H}\}$, i.e., $\mathcal{S}=\mathcal{S}_{SO}\times\mathcal{S}_{CR}$. Let $s_{SO,P}$ and $s_{SO,C}$ represent the source of physical and cyber layers, respectively. We assume that the alert triage process can accurately identify the source of attacks, i.e., $\Pr(s_{SO,i}|\phi_j)=\mathbf{1}_{\{i=j\}},\forall i,j\in\{P,C\}$. Let $s_{CR,L}$ and $s_{CR,H}$ represent low and high criticality, respectively. We assume that the triage process cannot accurately identify feints as low criticality and real attacks as high criticality. The revelation kernel is separable and takes the form of $o(s_{SO},s_{CR}|\theta_i,\phi_j)=\Pr(s_{SO}|\phi_j)\cdot\Pr(s_{CR}|\theta_i),s_{SO}\in\mathcal{S}_{SO},s_{CR}\in\mathcal{S}_{CR},i\in\{FE,RE\},j\in\{P,C\}$. We choose the values of o so that the attack is more likely to be feint (resp. real) when the criticality level is low (resp. high).

The inter-arrival time at attack stage $k\in\mathbb{Z}^{0+}$ follows an exponential distribution with rate $\beta(\theta^k,\theta^{k+1})$ parameterized by the attack's type θ^k,θ^{k+1}. Thus, the average inter-arrival time $\mu(\theta^k,\theta^{k+1}) := 1/\beta(\theta^k,\theta^{k+1})$ also depends on the attack's type at the current and the next attack stages as shown in Table 5. We choose the benchmark values based on the literature (e.g., Shah et al. (2019a,b) and the references within) and attacks can change these values in different IDoS attacks.

The average inspection time $\bar{d}$ in Section 4.3 depends on the criticality s^k_{CR} and attack's type θ^k at attack stage $k\in\mathbb{Z}^{0+}$, as shown in Table 6. We choose the benchmark values of $\bar{d}(s^k_{CR},\theta^k)$ based on Shah et al. (2019a), and these values can change for different human operators and IDoS attacks. We add a random noise uniformly distributed in $[-5,5]$ to the average inspection time to simulate the AITN.

The stage cost $\bar{c}(w^k,s^k_{SO})$ at attack stage $k\in\mathbb{Z}^{0+}$ in Section 5.2 depends on the alert response $w^k\in\mathcal{W}$ and the

Table 7
The benchmark values of the stage cost $\bar{c}(w^k, s^k_{SO}), \forall w^k \in \mathcal{W}, s^k_{SO} \in \mathcal{S}_{SO}$.

Reward of dismissing feints w_{FE}	\$80
Reward of identifying real attacks w_{RE} in physical layer	\$500
Reward of identifying real attacks w_{RE} in cyber layer	\$100
Cost of incomplete alert response w_{UN} or w_{NI}	\$300

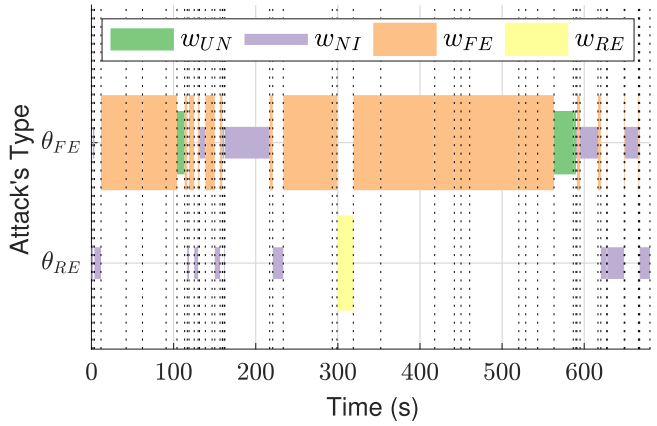


Fig. 3. Alert response $w^k \in \mathcal{W}$ for the k-th attack whose type is shown in the y-axis. The k-th vertical dash line represents the k-th alert's arrival time t^k.

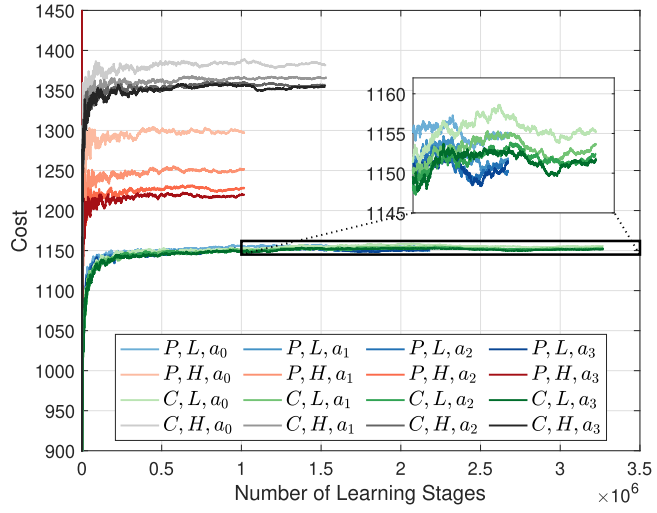


Fig. 4. The convergence of the estimated ECuC $Q^h(s^{I_h}, a^h)$ vs. the number of inspection stages.

source $s^k_{SO} \in \mathcal{S}_{SO}$. We determine the benchmark values of $\bar{c}(w^k, s^k_{SO})$ per alert in Table 7 based on the salary of the SOC analysts and the estimated loss of the associated attacks.

7.2. Analysis of numerical results

We plot the dynamics of the operator's alert responses in Fig. 3 under the benchmark experiment setup in Section 7.1. We use green, purple, orange, and yellow to represent w_{UN}, w_{NI}, w_{FE}, and w_{RE}, respectively. The heights of squares are also used to distinguish the four categories.

7.2.1. Adaptive learning during the real-Time monitoring and inspection

Based on Algorithm 1, we illustrate the learning process of the estimated ECuC $Q^h(s^{I_h}, a^h)$ for all $s^{I_h} \in \mathcal{S}$ and $a^h \in \mathcal{A}$ at each inspection stage $h \in \mathbb{Z}^{0+}$ in Fig. 4. We choose $\alpha^h(s^{I_h}, a^h) = \frac{k_c}{k_{TI}(s^{I_h})-1+k_c}$ as the learning rate, where $k_c \in (0, \infty)$ is a constant parameter and $k_{TI}(s^{I_h}) \in \mathbb{Z}^{0+}$ is the number of visits to $s^{I_h} \in \mathcal{S}$ up to stage $h \in \mathbb{Z}^{0+}$. Here, the AM action a^h is implemented randomly at each inspection stage h, i.e., $\epsilon = 1$. Thus, all four AM actions ($M = 3$) are explored equally on average for each $s^{I_h} \in \mathcal{S}$ as shown in Fig. 4. Since the number of visits to different category labels depends on the transition probability κ_{AT}, the learning stages for four category labels are of different lengths.

We denote category labels $(s_{SO,P}, s_{CR,L})$, $(s_{SO,P}, s_{CR,H})$, $(s_{SO,C}, s_{CR,L})$, and $(s_{SO,C}, s_{CR,H})$ in blue, red, green, and black, respectively. To distinguish four AM actions, a deeper color represents a larger $m \in \{0, 1, 2, 3\}$ for each category label $s_{SO,i}, s_{CR,j}, i \in \{P, C\}, j \in \{H, L\}$. The inset black box magnifies the selected area. The optimal strategy $\sigma^* \in \Sigma$ is to take a_3 for all category labels. The risk $v^*(s^{I_h}) = u(s^{I_h}, \sigma^*)$ under the optimal strategy has the approximated values of \$1153, \$1221, \$1154, and \$1358 for the above category labels in blue, red, green, and black, respectively. Based on Algorithm 1, we also simulate the operator's real-time monitoring and inspection under IDoS attacks when AM strategy is not applied. The risks $v^0(s^{I_h}) := u(s^{I_h}, \sigma^0)$ under the default AM strategy $\sigma^0 \in \Sigma$ have the approximated values of \$1377, \$1527, \$1378, and \$1620 for the category label $(s_{SO,P}, s_{CR,L})$, $(s_{SO,P}, s_{CR,H})$, $(s_{SO,C}, s_{CR,L})$, and $(s_{SO,C}, s_{CR,H})$, respectively. These results illustrate that the optimal AM strategy $\sigma^* \in \Sigma$ can significantly reduce the risk under IDoS attacks for all category labels and the reduction percentage can be as high as 20%.

We further investigate the IDoS risk under the optimal AM strategy σ^* as follows. As illustrated in Fig. 4, when the criticality level is high (i.e., the attack is more likely to be real), the attacks targeting cyber layers (denoted in black) result in a higher risk than the one targeting physical layers (denoted in red). This asymmetry results from the different rewards of identifying real attacks in physical or cyber layers denoted in Table 7. Since dismissing feints brings the same reward in physical and cyber layers, the attacks targeting physical or cyber layers result in similar IDoS risks when the criticality level is low. Within physical or cyber layers, high-criticality alerts result in a higher risk than low-criticality alerts do.

The value of $Q^h(s^{I_h}, a_m), m \in \{0, 1, 2\}$, represents the risk when RADAMS deviates to sub-optimal AM action a_m for a single category label $s^{I_h} \in \mathcal{S}$. As illustrated by the red and black lines in Fig. 4, this single deviation can increase the risk under alerts of high criticality. However, it hardly increases the risk under alerts of low criticality as illustrated by the green and blue lines in the inset black box of Fig. 4. These results illustrate that we can deviate from the optimal AM strategy to sub-optimal ones for some category labels with approximately equivalent risk, which we refer to as the *attentional risk equivalency* in Remark 5.

Remark 5 (Attentional Risk Equivalency)**.** The above results illustrate that we can contain the IDoS risk by selecting proper sub-optimal strategies. If applying the optimal AM strategy σ^* is costly, then RADAMS can choose not to apply AM strategy for $(s_{SO,C}, s_{CR,L})$ or $(s_{SO,P}, s_{CR,L})$ without significantly increasing the IDoS risks.

7.2.2. Optimal AM strategy and resilience margin under different stage costs

We define *resilience margin* as the difference of the risks under the optimal and the default AM strategies. We investigate how the cost of incomplete alert response in Table 7 affects the optimal AM strategy and the resilience margin in Fig. 5.

As shown in the upper figure, the optimal strategy remains to choose AM action a_3 when the alert is of high criticality. When the alert is of low criticality, then as the cost increases, the optimal AM strategy changes sequentially from a_3, a_2, and a_1 to a_0; i.e., RADAMS gradually decreases $m \in \{0, 1, 2, 3\}$, the number of de-emphasized alerts. As shown in the lower figure, the resilience margin increases monotonously with the cost. The optimal strat-

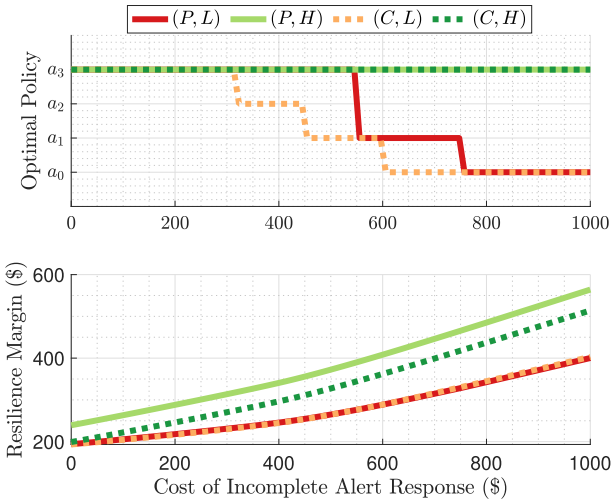


Fig. 5. The optimal AM strategy and the risk vs. the cost of an incomplete alert response under category label $(s_{SO,P}, s_{CR,L})$, $(s_{SO,P}, s_{CR,H})$, $(s_{SO,C}, s_{CR,L})$, and $(s_{SO,C}, s_{CR,H})$ in solid red, solid green, dashed yellow, and dashed green, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

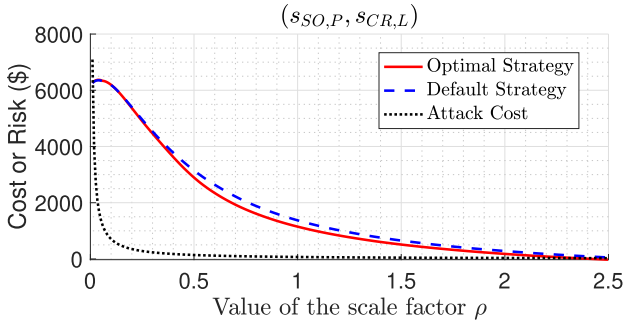


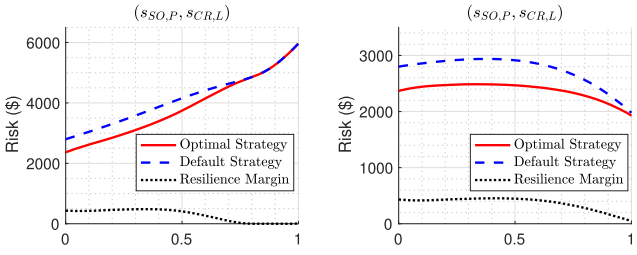
Fig. 6. IDoS risk vs. ρ under the optimal and the default AM strategies in solid red and dashed blue, respectively. The black line represents the attack cost per work shift of 24 hours. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

egy for alerts of high criticality yields a larger resilience margin than the one for low criticality.

Remark 6 (Tradeoff of Monitoring and Inspection)**.** The results show that the optimal strategy strikes a balance between real-time monitoring a large number of alerts and inspecting selected alerts with high quality. Moreover, the optimal strategy is resilient for a large range of cost values ([$0,$1000]). If the cost is high, and the alert is of low (resp. high) criticality, then the optimal strategy encourages monitoring (resp. inspecting) by choosing a small (resp. large) m. However, when the cost of an incomplete alert response is relatively low, the optimal strategy is a_4 for all alerts because the high-quality inspection outweighs the high-quantity monitoring.

7.2.3. Arrival frequency of IDos attacks

As stated in Section 3.1, feint attacks with the goal of triggering alerts require fewer resources to craft. Thus, we let $\hat{c}_{RE} = \$0.04$ and $\hat{c}_{FE} \in (0, \hat{c}_{RE})$ denote the cost to generate a real attack and a feint, respectively. With $\hat{c}_{RE}$ and $\hat{c}_{FE}$, we can compute the attack cost of feint and real attacks per work shift of 24 hours. Let ρ be the scaling factor for the arrival frequency, and in Section 7.2.3, the average inter-arrival time is $\hat{\mu}(\theta^k, \theta^{k+1}) = \rho\mu(\theta^k, \theta^{k+1}), \forall \theta^k, \theta^{k+1} \in \Theta$. We investigate how the scale factor $\rho \in (0, 2.5]$ affects the IDoS risk and the attack cost in Fig. 6. As ρ decreases, the attacker generates feint and real attacks at a higher frequency. Then, the risks under both the optimal and the default strategies increase. However, the optimal AM strategy can reduce the increase rate for a large range of $\rho \in [0.5, 2]$.



(a) Low-cost feints $\hat{c}_{FE} = 1$. **(b)** High-cost feints $\hat{c}_{FE} = 5$.

Fig. 7. IDoS risk vs. $\eta_{FE} \in [0, 1]$ under the optimal and the default AM strategies in red and blue, respectively. The black line represents the resilience margin. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Remark 7 (Attacker's Dilemma)**.** From the attacker's perspective, although increasing the attack frequency can induce a high risk to the organization, and the attacker can gain from it, the frequency increase also increases the attack cost exponentially, as shown by the dotted black line in Fig. 6. Thus, the attacker has to strike a balance between the attack cost and the attack gain (represented by the IDoS risk). Moreover, attackers with a limited budget are not capable to choose small values of ρ (i.e., high attack frequencies).

7.2.4. Percentage of feint and real attacks

Consider the case where κ_{AT} independently generates feints and real attacks with probability η_{FE} and $\eta_{RE} = 1 - \eta_{FE}$, respectively. We consider the case where the attacker has a limited budge $\hat{c}_{\max} = \$270$ per work shift (i.e., 86400s) and generates feint and real attacks at the same rate $\hat{\beta}$, i.e., $\beta(\theta^k, \theta^{k+1}) = \hat{\beta}, \forall \theta^k, \theta^{k+1} \in \Theta$. Consider the attack cost in Section 7.2.3, the attacker has the following budget constraint, i.e.,

$$86,400 \cdot \hat{\beta} \cdot (\eta_{FE}\hat{c}_{FE} + \eta_{RE}\hat{c}_{RE}) \leq \hat{c}_{max}. \tag{12}$$

The budget constraint results in the following tradeoff. If the attacker chooses to increase the probability of real attack η_{RE}, then he has to reduce the arrival frequency $\hat{\beta}$ of feint and real attacks. We investigate how the probability of feints affects the IDoS risk in Fig. 7 under the optimal and the default AM strategies in red and blue, respectively. The feints are of low and high costs in Fig. 7a and 7 b, respectively.

As shown in Fig. 7a, when the feints are of low cost, i.e., $\hat{c}_{FE} = \hat{c}_{RE}/10$, generating feints with a higher probability monotonously increases the IDoS risks for both AM strategies. When the probability of feints is higher than 80%, the resilience margin is zero; i.e., the optimal and the default AM strategies both induce high risks. However, as the probability of feint decreases, the resilience margin increases to around $500; i.e., the default strategy can moderately reduce the risk, but the optimal strategy can excessively reduce the risk.

Remark 8 (Half-Truth Attack for High-Cost Feints)**.** As shown in Fig. 7b, when the feints are of high cost, i.e., $\hat{c}_{FE} = \hat{c}_{RE}/2$, then the optimal attack strategy is to deceive with *half-truth*, i.e., generating feint and real attacks with approximately equal probability to induce the maximum IDoS risk. As the probability of feints decreases from $\eta_{FE} = 1$, the risk increases significantly under the default AM strategy but moderately under the optimal one.

The figures in Fig. 7 show that the optimal attack strategy under the budget constraint (12) needs to adapt to the cost of feint generation. Regardless of the attack strategy, the optimal AM strategy can reduce the risk and achieve a positive resilient margin for all category labels $(s_{SO,i}, s_{CR,j}), i \in \{P, C\}, j \in \{L, H\}$. More-

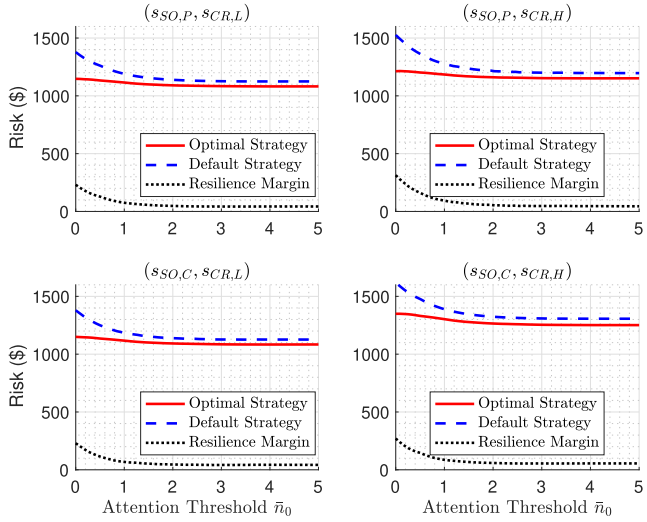


Fig. 8. Risk vs. attention threshold under the optimal and the default AM strategies in red and blue, respectively. The black dotted line represents the resilience margin. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

over, higher feint generation cost reduces the arrival frequency of IDoS attacks due to (12). Thus, comparing to Fig. 7a, the risk in Fig. 7b is lower for the same η_{FE} under the optimal or the default AM strategies, especially when η_{FE} is close to 1.

7.2.5. The Operator's attention capacity

We consider the following attention function $f_{LOE} \circ f_{SL}$ with a constant attention threshold, i.e., $\bar{n}(y_{EL}, s^k) = \bar{n}_0, \forall y_{EL}, s^k \in \mathcal{S}$. Consider the following trapezoid attention function. If $n^t \leq \bar{n}_0$, the LOE $\omega^t = 1$; i.e., the operator can retain the high LOE when the number of distractions is less than the attention threshold $\bar{n}_0$. If $n^t > \bar{n}_0$, the LOE ω^t gradually decreases as n^t increases. Then, a larger value of $\bar{n}_0$ indicates a high attention capacity. We investigate how the value of $\bar{n}_0$ affects the risk in Fig. 8.

As the operator's attention capacity increases, the risks under the optimal and the default AM strategies decrease for all category labels. The resilience margin decreases from around \$200 to \$50 as $\bar{n}_0$ increases from 0 to 2 and then maintains the value of around \$50. Thus, the optimal strategy suits operators with a large range of attention capacity, especially for the ones with limited attention capacity.

8. Conclusion

Attentional human vulnerabilities exploited by attackers lead to a new class of proactive attacks called the Informational Denial-of-Service (IDoS) attacks. IDoS attacks generate a large number of feint attacks on purpose to deplete the limited human attention resources and exacerbate the alert fatigue problem. In this work, we have formally defined IDoS attacks as a sequence of feint and real attacks of heterogeneous targets, which can be characterized by the Markov renewal process. We have abstracted the alert generation and technical-level triage processes as a revelation probability to establish a stochastic relationship between the IDoS attack's hidden types and targets and the associated alert's observable category labels. We have explicitly incorporated human factors (e.g., levels of expertise, stress, and efficiency) and empirical results (e.g., the Yerkes-Dodson law and the sunk cost fallacy) to model the operators' attention dynamics and the processes of alert monitoring, inspection, and response in real time. Based on the system-scientific human attention and alert response model, we have developed a Resilient and Adaptive Data-driven alert and Attention Management Strategy (RADAMS) to assist human operators in combating IDoS attacks. We have proposed a Reinforcement Learning (RL)-based algorithm to obtain the optimal assistive strategy according to the costs of the operator's alert responses in real time.

Through theoretical analysis, we have observed the *Product Principle of Attention* (PPoA), the fundamental limits of Attentional Deficiency Level (ADL) and risk, and tradeoff among the ADL, the reward of alert attention, and the impact of alert inattention. Through the experimental results, we have corroborated the *effectiveness, adaptiveness, robustness*, and *resilience* of the proposed assistive strategies as follows. First, the optimal AM strategy outperforms the default strategy and can effectively reduce the IDoS risk by as much as 20%. Second, the strategy adapts to different category labels to strike a balance of monitoring and inspections. Third, the optimal AM strategy is robust to deviations. We can apply suboptimal strategies at some category labels without significantly increasing the IDoS risk. Finally, the optimal AM strategy is resilient to a large variations of costs, attack frequencies, and human attention capacities.

The current work uses Industrial Control Systems (ICS) as a quintessential example to illustrate the IDoS attacks and the associated human-aware alert and attention management strategies. RADAMS can also be applied to broad types of scenarios (e.g., healthcare, public transport control, and weather warning) that require human operators of limited attention resources to monitor and manage massive alerts in real time with a high level of situational awareness. RADAMS adopts the "less is more" principle by restricting the amount of information processed by the human operators to be within their attention capacities. Such principle is transferable to other assailable cognitive resources of human operators, including memory, reasoning, and learning capacity. The future work would incorporate more generalized models (e.g., the spatio-temporal self-excited process) to capture the history-dependent temporal arrival of IDoS attacks, the spatial location of the alerts, their impacts on human attention, and the associated human-assistive security technologies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Linan Huang: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft. **Quanyan Zhu:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

References

Alsubhi, K., Aib, I., Boutaba, R., 2012. Fuzmet: a fuzzy-logic based alert prioritization engine for intrusion detection systems. Int. J. Network Manage. 22 (4), 263–284.
Aminanto, M.E., Ban, T., Isawa, R., Takahashi, T., Inoue, D., 2020. Threat alert prioritization using isolation forest and stacked auto encoder with day-forward-chaining analysis. IEEE Access 8, 217977–217986.
Ancker, J.S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., Kaushal, R., 2017. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 17 (1), 1–9.
Arkes, H.R., Blumer, C., 1985. The psychology of sunk cost. Organ Behav Hum Decis Process 35 (1), 124–140.
Bassett, G., Hylender, C.D., Langlois, P., Pinto, A., Widup, S., 2021. Data Breach Investigations Report. Technical Report. Verizon DBIR Team.
Bertsekas, D.P., Tsitsiklis, J.N., 1996. Neuro-dynamic Programming. Athena Scientific.
Bierma, M., Doak, J.J.E., Hudson, C., 2016. Learning to rank for alert triage. In: 2016 IEEE Symposium on Technologies for Homeland Security (HST). IEEE, pp. 1–5.
Bouzar-Benlabiod, L., Rubin, S.H., Belaidi, K., Haddar, N.E., 2020. RNN-VED for reducing false positive alerts in host-based anomaly detection systems. In: 2020 IEEE

21st International Conference on Information Reuse and Integration for Data Science (IRI). IEEE, pp. 17–24.
Bryant, B.D., Saiedian, H., 2020. Improving SIEM alert metadata aggregation with a novel kill-chain based classification model. Computers & Security 94, 101817.
Casey, W., Morales, J.A., Wright, E., Zhu, Q., Mishra, B., 2016. Compliance signaling games: toward modeling the deterrence of insider threats. Comput Math Organ Theory 22 (3), 318–349.
Corona, I., Giacinto, G., Roli, F., 2013. Adversarial attacks against intrusion detection systems: taxonomy, solutions and open issues. Inf Sci (Ny) 239, 201–225. doi:10.1016/j.ins.2013.03.022.
Cotroneo, D., Paudice, A., Pecchia, A., 2017. Empirical analysis and validation of security alerts filtering techniques. IEEE Trans Dependable Secure Comput 16 (5), 856–870.
Elshoush, H., Osman, I., 2010. Reducing false positives through fuzzy alert correlation in collaborative intelligent intrusion detection systems-a review. In: IEEE Int. Conf. Fuzzy Syst.. IEEE, pp. 1–8.
Franklin, L., Pirrung, M., Blaha, L., Dowling, M., Feng, M., 2017. Toward a visualization-supported workflow for cyber alert management using threat models and human-centered design. In: 2017 IEEE Symposium on Visualization for Cyber Security (VizSec). IEEE, pp. 1–8.
Ganesan, R., Jajodia, S., Shah, A., Cam, H., 2016. Dynamic scheduling of cybersecurity analysts for minimizing risk using reinforcement learning. ACM Transactions on Intelligent Systems and Technology (TIST) 8 (1), 1–21.
Goeschel, K., 2016. Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive bayes for off-line analysis. In: SoutheastCon 2016. IEEE, pp. 1–6.
Hassan, W.U., Guo, S., Li, D., Chen, Z., Jee, K., Li, Z., Bates, A., 2019. NODOZE: combatting threat alert fatigue with automated provenance triage. Network and Distributed Systems Security Symposium.
Hitzel, B., 2019. The art of cyber war and cyber battle: Deception operations. https://www.networkdefenseblog.com/post/art-of-cyber-war-deception.
Huang, L., Jia, S., Balcetis, E., Zhu, Q., 2022. Advert: an adaptive and data-driven attention enhancement mechanism for phishing prevention. IEEE Transactions on Information Forensics and Security 17, 2585–2597. doi:10.1109/TIFS.2022.3189530.
Huang, L., Zhu, Q., 2019. Adaptive honeypot engagement through reinforcement learning of semi-Markov decision processes. In: International Conference on Decision and Game Theory for Security. Springer, pp. 196–216.
Huang, L., Zhu, Q., 2020. A dynamic games approach to proactive defense strategies against advanced persistent threats in cyber-physical systems. Computers & Security 89, 101660. doi:10.1016/j.cose.2019.101660.
Huang, L., Zhu, Q., 2020. Farsighted risk mitigation of lateral movement using dynamic cognitive honeypots. In: International Conference on Decision and Game Theory for Security. Springer, pp. 125–146.
Huang, L., Zhu, Q., 2021. Combating informational denial-of-service (IDoS) attacks: modeling and mitigation of attentional human vulnerability. In: International Conference on Decision and Game Theory for Security. Springer.
Huang, L., Zhu, Q., 2021. Duplicity games for deception design with an application to insider threat mitigation. IEEE Trans. Inf. Forensics Secur. 16, 4843–4856. doi:10.1109/TIFS.2021.3118886.
Huang, L., Zhu, Q., 2022. Zetar: modeling and computational design of strategic and adaptive compliance policies. arXiv preprint arXiv:2204.02294.
Jajodia, S., Ghosh, A.K., Swarup, V., Wang, C., Wang, X.S., 2011. Moving Target Defense: Creating Asymmetric uncertainty for cyber threats, Vol. 54. Springer Science & Business Media.
Laszka, A., Vorobeychik, Y., Fabbri, D., Yan, C., Malin, B., 2017. A game-theoretic approach for alert prioritization. AAAI Workshops.
Liu, D., Wang, X., Camp, L.J., 2009. Mitigating inadvertent insider threats with incentives. In: International Conference on Financial Cryptography and Data Security. Springer, pp. 1–16.
LLC, P.I., 2015. The Cost of Malware Containment. Technical Report.
McElwee, S., Heaton, J., Fraley, J., Cannady, J., 2017. Deep learning for prioritizing and responding to intrusion detection alerts. In: IEEE Military Communications Conference. IEEE, pp. 1–5.
Miserendino, S., Maynard, C., Davis, J., 2017. Threatvectors: Contextual workflows and visualizations for rapid cyber event triage. In: 2017 International Conference On Cyber Incident Response, Coordination, Containment & Control (Cyber Incident). IEEE, pp. 1–8.
Mutz, D., Vigna, G., Kemmerer, R., 2003. An experience developing an ids stimulator for the black-box testing of network intrusion detection systems. In: 19th Annual Computer Security Applications Conference, 2003. Proceedings.. IEEE, pp. 374–383.
Newcomb, E.A., Hammell, R.J., Hutchinson, S., 2016. Effective prioritization of network intrusion alerts to enhance situational awareness. In: 2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, pp. 73–78.
Noel, S., Jajodia, S., 2008. Optimal IDs sensor placement and alert prioritization using attack graphs. Journal of Network and Systems Management 16 (3), 259–275.
Ohta, S., Kurebayashi, R., Kobayashi, K., 2008. Minimizing false positives of a decision tree classifier for intrusion detection on the internet. Journal of network and systems management 16 (4), 399–419.
Patton, S., Yurcik, W., Doss, D., 2001. An achilles' heel in signature-based ids: Squealing false positives in snort. In: Proceedings of RAID, Vol. 2001. Citeseer.
Pietraszek, T., Tanner, A., 2005. Data mining and machine learning-towards reducing false positives in intrusion detection. Information security technical report 10 (3), 169–183.
Project, V. M., 2017. How does hepatitis b combat the immune system? https://vimeo.com/248010182.
Renners, L., Heine, F., Rodosek, G.D., 2017. Modeling and learning incident prioritization. In: 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Vol. 1. IEEE, pp. 398–403.
Roesch, M., et al., 1999. Snort: Lightweight intrusion detection for network. In: Lisa, Vol. 99, pp. 229–238.
Salah, S., Maciá-Fernández, G., Díaz-Verdejo, J.E., 2013. A model-based survey of alert correlation techniques. Comput. Networks 57 (5), 1289–1317.
Shah, A., Ganesan, R., Jajodia, S., Cam, H., 2019. A two-step approach to optimal selection of alerts for investigation in a csoc. IEEE Trans. Inf. Forensics Secur. 14 (7), 1857–1870. doi:10.1109/TIFS.2018.2886465.
Shah, A., Ganesan, R., Jajodia, S., Cam, H., 2019. Understanding tradeoffs between throughput, quality, and cost of alert analysis in a CSOC. IEEE Trans. Inf. Forensics Secur. 14 (5), 1155–1170. doi:10.1109/TIFS.2018.2871744.
Spathoulas, G., Katsikas, S., 2010. Reducing false positives in intrusion detection systems. Comput & Secur 29 (1), 35–44.
Stouffer, K., Falco, J., Scarfone, K., et al., 2011. Guide to industrial control systems (ics) security. NIST special publication 800 (82). 16–16
Su, Y.-H., Cho, M.C.Y., Huang, H.-C., 2019. False alert buster: an adaptive approach for nids false alert filtering. In: Proceedings of the 2nd International Conference on Computing and Big Data, pp. 58–62.
Sundaramurthy, S.C., Bardas, A.G., Case, J., Ou, X., Wesch, M., McHugh, J., Rajagopalan, S.R., 2015. A human capital model for mitigating security analyst burnout. In: Eleventh Symposium On Usable Privacy and Security (SOUPS 2015), pp. 347–359.
Tessian, 2020. The Psychology of Human Error. Technical Report.
Tong, L., Laszka, A., Yan, C., Zhang, N., Vorobeychik, Y., 2020. Finding needles in a moving haystack: Prioritizing alerts with adversarial reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, pp. 946–953.
Wang, Z., Zhu, H., Sun, L., 2021. Social engineering in cybersecurity: effect mechanisms, human vulnerabilities and attack methods. IEEE Access 9, 11895–11910.
Yates, R.D., Sun, Y., Brown, D.R., Kaul, S.K., Modiano, E.H., Ulukus, S., 2021. Age of information: an introduction and survey. IEEE J. Sel. Areas Commun. 39, 1183–1210.
Yerkes, R.M., Dodson, J.D., et al., 1908. The relation of strength of stimulus to rapidity of habit-formation. Punishment: Issues and experiments 27–41.
Zhong, C., Lin, T., Liu, P., Yen, J., Chen, K., 2018. A cyber security data triage operation retrieval system. Computers & Security 76, 12–31.
Zhong, C., Yen, J., Liu, P., Erbacher, R.F., 2016. Automate cybersecurity data triage by leveraging human analysts' cognitive process. In: 2016 IEEE 2nd International Conference on big data security on cloud (BigDataSecurity), IEEE International Conference on high performance and smart computing (HPSC), and IEEE International Conference on intelligent data and security (IDS). IEEE, pp. 357–363.
Zhong, C., Yen, J., Liu, P., Erbacher, R.F., 2018. Learning from experts' experience: toward automated cyber security data triage. IEEE Syst. J. 13 (1), 603–614.
Zimmerman, C., 2014. Ten strategies of a world-class cybersecurity operations center. The MITRE Corporation.

Linan Huang received the B.Eng. degree (Hons.) in Electrical Engineering from Beijing Institute of Technology, China, in 2016 and the Ph.D. degree in electrical engineering from New York University (NYU), Brooklyn, NY, USA, in 2022. His research interests include dynamic decision-making of the multi-agent system, mechanism design, artificial intelligence, security, and resilience for cyberphysical systems.

Quanyan Zhu (S'04-M'12) received B. Eng. in Honors Electrical Engineering with distinction from McGill University in 2006, M.A.Sc. from University of Toronto in 2008, and Ph.D. from the University of Illinois at Urbana-Champaign (UIUC) in 2013. After stints at Princeton University, he is currently an assistant professor at the Department of Electrical and Computer Engineering, New York University. He is a recipient of many awards including NSF CAREER Award, NYU Goddard Junior Faculty Fellowship, NSERC Postdoctoral Fellowship (PDF), NSERC Canada Graduate Scholarship (CGS), and Mavis Future Faculty Fellowships. He spearheaded and chaired INFOCOM Workshop on Communications and Control on Smart Energy Systems (CCSES), and Midwest Workshop on Control and Game Theory (WCGT). His current research interests include resilient and secure interdependent critical infrastructures, Internet of Things, cyber-physical systems, game theory, machine learning, network optimization and control. He has served as the general chair of the 7th Conference on Decision and Game Theory for Security (GameSec) in 2016, the 9th International Conference on NETwork Games, COntrol and OPtimisation (NETGCOOP) in 2018, and the 5th International Conference on Artificial Intelligence and Security (ICAIS 2019) in 2019.