

When Doing Nothing Is the Optimal Cyber Defense: Quantum-Inspired Abstention as a First-Class Security Action

Michael Nguyen Phuc^{1*}

¹RMIT University, Melbourne, VIC, Australia

Abstract. Cyber security usually favors action: when an abnormality is detected, systems react visibly (block, alert, rotate) even when evidence is weak. However, work on abstention and selective prediction shows that deferring commitment can be rational under uncertainty, trading coverage for lower expected risk. At the same time, operational realities such as alert overload and adaptive adversaries imply that visible defensive reactions can backfire by increasing analyst burden and providing feedback that supports attacker probing and policy inference. This paper presents quantum-inspired abstention as a first-class security action. Using a quantum decision-theoretic lens, I model defensive commitment as a “measurement” that collapses an uncertain belief state into an externally observable response, and I define abstention as deliberate non-commitment that suppresses or delays measurement when uncertainty and leakage risk are high. I integrate these ideas into a conceptual framework and a minimal loss decomposition separating security loss, operational cost, and leakage-driven adversarial learning. I illustrate the approach through SOC triage and network intrusion detection scenarios, and I provide a lightweight simulation that instantiates the trade-offs among these losses—without requiring quantum hardware.

Keywords cyber defense, abstention, selective prediction, quantum-inspired, adversarial learning, information leakage.

1 Introduction

Modern cyber defense is frequently designed around an implicit assumption: when a threat signal appears, the defender should respond. In practice, this “action-first” posture manifests as visible interventions (blocking, quarantining, patching) or visible signaling (alerts, escalations, and active response playbooks). Yet, action is not free. In security operations centres (SOCs), persistent alarm volume and low signal-to-noise ratios can produce analyst overload and decision degradation, commonly discussed as alert fatigue [16]. At the network perimeter, intrusion detection and response mechanisms are increasingly confronted by adaptive adversaries whose probing strategies can exploit defensive feedback to infer thresholds, policies, and response logic [17]. In these settings, the defender’s observable reaction can become part of the attack surface.

A parallel line of research in statistical decision theory and machine learning provides a useful counterpoint: under uncertainty, “not deciding” can be rational. The reject option

*Corresponding author: michael.ng5724@gmail.com

and its descendants formalize abstention as a controllable decision outcome, trading coverage for reduced expected risk [1,2]. Subsequent developments extend rejection to modern learning settings, including margin-based formulations, selective prediction for deep networks, integrated reject architectures, and calibrated selective classification [3–7]. This literature establishes an important principle: deferring commitment can dominate forced decisions when uncertainty is high, costs are asymmetric, or errors are particularly harmful [8–11]. However, these results are typically framed for classification performance and reliability, rather than adversarial security environments where the defender’s actions are observed and exploited.

Cyber security research has long acknowledged that influencing an attacker’s ideas is a protective mechanism. Defensive deception clearly sees information management as a security measure and lists ways to influence, hide, or change how attackers think [12]. Game-theoretic methods also indicate that defense policy choices may be improved when the attacker sees signals and changes their conduct appropriately [13]. Proactive strategies like moving target defense (MTD) also try to level the playing field for attackers by constantly modifying the system’s attack surface, but they may be expensive to run and build [14,15]. A first-class treatment of abstention is not always clear in this literature. Abstention is not just “doing nothing,” but a purposeful strategy that keeps or delays externally visible defensive commitment when the information value to an adversary is high.

This study presents quantum-inspired abstention as a premier security measure. We use a quantum decision-theoretic framework whereby the commitment to a discernible defensive reaction is seen as a “measurement” that reduces an ambiguous belief state to an observable consequence for an opponent [18,19]. In this context, abstention means putting off or stopping measurement, which keeps confusion about defense policy and cuts down on information leakage that may help the enemy learn faster. Our contribution does not pertain to quantum hardware or physical quantum processes; instead, we employ established quantum-inspired decision formalisms as a succinct means to model commitment, observability, and information leakage in adversarial contexts, aligning with the broader evidence that quantum-inspired methodologies can yield practical modeling and algorithmic benefits even within classical systems [20].

I make four contributions. Initially, we conceptualize abstention as a primary action within cyber defense policy, separate from mere inaction: abstention represents a purposeful decision that influences observability and operational workload, rather than an inability or neglect to respond. Second, we present a streamlined conceptual framework that synthesizes (i) uncertainty-aware rejection strategies derived from selective prediction [1–11], (ii) adversarial inference considerations rooted in defensive deception and game-theoretic security [12,13], and (iii) a quantum-inspired decision structure that differentiates between commitment (measurement) and non-commitment (abstention) [18,19]. Third, we demonstrate the application of this model to practical systems—namely, SOC triage and network intrusion detection—where abstention can mitigate self-inflicted damage and impede adversary inference without replacing proactive countermeasures such as MTD [14–17]. Fourth, we augment this conceptual model with a lightweight simulation study that assesses the trade-offs among security loss, operational cost, and leakage-driven adversarial learning under various defender policies and instantiates the proposed loss decomposition. Additionally, we offer practitioner-oriented implementation guidelines that map abstention to SOC/NIDS decision protocols (e.g., SOAR-style playbooks), including triggers, safeguards, and escalation criteria. Together, these contributions reconceptualize cyber defense as not merely a matter of determining what actions to undertake, but also of discerning when to refrain from acting.

The remaining portion of the paper is structured as follows. Selective prediction, defensive deception, and related security decision frameworks are examined in Section 2. The loss decomposition, which encompasses security loss, operational cost, and leakage risk, is

introduced in Section 3, which formalizes abstention as an explicit defensive action. The corresponding policy structure and the quantum-inspired commitment/abstention abstraction are presented in Section 4. The scenarios of SOC triage and NIDS in Section 5 demonstrate how abstention mitigates self-inflicted injury and adversary inference. A lightweight simulation-based evaluation is reported in Section 6, which quantifies important trade-offs and motivates operational design choices. Limitations and practical considerations are addressed in Section 7, while Section 8 concludes.

2 Background & Related Work

This section positions quantum-inspired abstinence at the convergence of three bodies of literature. Initially, research on reject-option and selective prediction demonstrates that refraining from making a decision may be a logical, cost-sensitive strategy in the face of uncertainty. Secondly, cybersecurity efforts concerning deception, moving target defense, and SOC operations demonstrate that observable defensive measures may create operational difficulties and potentially disclose information to adaptive adversaries. Third, quantum decision-making and quantum-inspired computing provide a concise framework for formalizing "commitment" vs "non-commitment" choices without the need of quantum hardware.

2.1 Reject option / abstention / selective classification

The idea that “refusing to decide” can be optimal is a classical result in statistical decision theory. Early work formalised the error–reject trade-off and showed that introducing an explicit reject option can reduce expected risk when the cost of a wrong decision is high relative to the cost of rejection [1]. Later formulations generalised this perspective into classification frameworks where a model may either predict a label or reject, clarifying the conditions under which rejection improves overall decision quality compared to forced prediction [2].

As machine learning techniques advanced, reject-option concepts were included into optimization-based models. For instance, variations of support vector machines with a reject option explicitly integrate abstinence into the learning aim, seeing rejection as a structured choice rather than a mere afterthought [3]. Related research on learning with rejection further organizes the framework by concurrently learning a prediction rule and a rejection rule, often examined via risk–coverage trade-offs, where the objective is to sustain low risk on the fraction of occurrences it elects to address [4].

Recently, selective prediction has expanded these concepts to deep neural networks. Research on selective classification for deep models examined methods to regulate coverage while maintaining limited error on accepted predictions [5]. SelectiveNet advanced the integration of selection into model training, enabling the system to learn calibrated acceptance zones instead of depending only on post-hoc confidence levels [6]. Further research underscores the need for abstention to be reliability-aware: calibrated selective classification and one-sided prediction frameworks enhance the criteria for triggering abstention, ensuring that risk guarantees on the non-abstained group remain significant [7,8]. Optimal methods for reject-option classifiers provide theoretical insights for formulating rejection policies that transcend ordinary heuristics [9], whereas surveys integrate these perspectives into a unified understanding of abstention as a trust mechanism for contemporary predictors [10].

A fundamental recurrent assumption in these studies is that choices to abstain should be influenced by uncertainty. Practical methods for estimating uncertainty, such as dropout-based Bayesian approximations, are often used to determine when a model should wait or refrain from making a possibly erroneous choice [11]. Nevertheless, the majority of this work focuses

on predicted risk in non-adversarial contexts. In cybersecurity, the environment is hostile, and the defender's actions are observable; hence, abstention may provide supplementary benefits as a strategy to mitigate information leakage and impede attacker adaptation, in addition to standard error-risk assessments.

2.2 Cybersecurity: deception, MTD, and operational cost

Cybersecurity defenses are not just about accuracy; they are also about controlling what the enemy learns and limiting the operational cost of responding. Defensive deception research clearly defines security as an information issue, presenting a taxonomy of deception strategies and demonstrating how defenders may modify attacker assumptions, raise attacker uncertainty, and limit attacker efficacy [12]. Game-theoretic models substantiate this perspective by examining optimum defender tactics while adversaries adjust according to observable signals and results, indicating that a defense's actions may significantly affect attacker behavior [13].

Moving target defense (MTD) is a significant area of research designed to hinder attacker surveillance and exploitation by perpetually altering system configurations or surfaces [14,15]. Although MTD may enhance attacker unpredictability, it simultaneously raises technical complexity and operational burdens. This underscores a wider practical limitation: defenders often need to reconcile proactive change, reactive action, and organizational capability.

The operational capacity is particularly evident in SOC situations. Research on alert fatigue underscores that excessive warning volumes, false positives, and continual escalation may overwhelm analysts and diminish response quality, indicating that an increase in responses may paradoxically impair defensive efficacy [16]. Research at the detection layer reveals that realistic adversarial assaults on network intrusion detection systems enable adversaries to design probes and adaptive techniques that exploit the behavior of detectors and responders, transforming defender feedback into a learning mechanism for the attacker [17]. Together, these literatures motivate treating visibility and operational burden as first-class factors in cyber defense policy design—exactly where abstention becomes meaningful: not as neglect, but as a deliberate choice to delay or suppress observable commitment in regimes where the cost of action (including leakage and human overload) exceeds the benefit.

2.3 Quantum decision and quantum-inspired computing

Quantum decision and cognition models provide a compact formal language for representing decision-making under uncertainty. In these frameworks, a decision-maker can be modeled as having a superposed belief state that is not fully resolved until a “measurement” (i.e., a commitment to an outcome) occurs [18,19]. Importantly, these models are not used here to claim that defenders operate as physical quantum systems; rather, they offer a structured abstraction for reasoning about commitment, uncertainty, and the consequences of collapsing uncertainty into a visible choice.

This metaphor is inherently compatible with cybersecurity contexts, when undertaking a conspicuous defensive measure may disclose sensitive information. If defensive commitment is seen as a metric, then abstention equates to postponing measurement, maintaining uncertainty about the defender's policy and diminishing the adversary's capacity to deduce thresholds or reaction rationale. To enhance the validity of using this perspective, research on quantum-inspired computing has shown that quantum-inspired methodologies may provide practical modeling and algorithmic advantages even on conventional hardware and in non-quantum environments [20]. This substantiates the conceptualization of “quantum-inspired abstention” as a methodological instrument: a means to formalize and convey a choice frame-

work of commitment vs non-commitment, while being anchored in existing security and machine learning evidence.

3 Problem Statement

consider a defender operating a cyber defense system that continuously monitors an environment and must decide how to respond to uncertain threat evidence. At each decision point, the system observes a signal x , where x summarises available telemetry such as network features, host logs, alerts, anomaly scores, or probe patterns. Based on x , the defender selects an action a from an action set A . The objective is not only to reduce the likelihood and impact of successful attacks, but also to manage operational burden and limit information leakage that can be exploited by adaptive adversaries.

A key modelling choice in this work is to treat abstention as a first-class security action rather than as an absence of control. Accordingly, we define the defender's action set as:

$$A = \{\text{intervene, signal, **abstain**\}. \quad (1)$$

Here, intervene includes direct mitigations that change system state, such as blocking traffic, quarantining endpoints, resetting credentials, or applying patches. Signal includes actions primarily intended to record, escalate, or influence investigation and adversary perception, such as generating alerts, increasing logging, deploying decoys, or triggering deception mechanisms. Abstain denotes a deliberate choice to delay, suppress, or rate-limit externally observable responses while continuing to collect evidence (e.g., silent logging, delayed escalation, or non-committal handling), with the explicit purpose of avoiding premature commitment when uncertainty and leakage risk are high. In this formulation, abstention is not “doing nothing” in an operational sense; it is a controlled policy action that trades immediate responsiveness for reduced harm from false positives and reduced feedback to the adversary.

I assume an adaptive threat model where the attacker may probe the defender's system and observe aspects of the defender's responses. Specifically, the attacker can generate queries or probes (e.g., scanning patterns, crafted traffic, or low-cost test intrusions) and then use the presence, absence, or timing of defender responses as feedback. This aligns with security perspectives that treat defender behavior as part of a strategic interaction in which adversaries update beliefs and strategies based on observed signals and outcomes [12,13]. Under this model, repeated interactions can form a learning loop: probes \rightarrow observed responses \rightarrow attacker belief update \rightarrow refined probing or exploitation strategy. Work on realistic adversarial attacks against intrusion detection systems further supports the assumption that attackers can adapt their behaviour to the characteristics of detection and response mechanisms, including by exploiting feedback channels [17].

The central problem addressed by this paper is therefore: **how should a defender choose among intervene, signal, and abstain when observations are uncertain and when defensive commitment can create both operational cost and exploitable information leakage?** By explicitly including abstention in the action set and by modelling attacker adaptation through observable responses, we set up the foundation for a decision model in which “not committing yet” can be optimal in specific regimes—even when traditional action-biased defenses would force an immediate, visible response.

4 Quantum-Inspired Abstention Model

A central claim of this paper is that abstention should be modelled as a deliberate security action, not as an omission. To make this precise in adversarial settings, we introduce a quantum-inspired decision abstraction that distinguishes *commitment* from *non-commitment*. The purpose of this abstraction is not to invoke quantum hardware or physical quantum effects, but to provide a clean formal language for representing uncertainty, commitment, and information leakage in defender decisions.

4.1 Mapping: “measurement” as commitment to an observable defense

I start with a conceptual framework inspired by quantum decision-making and cognitive models [18,19]. The defender sustains a belief state amidst ambiguity, grounded on accessible data such as telemetry, alarms, and detected probes. In several operational systems, once the defender decides on a response—such as blocking, escalating, or implementing a visible countermeasure—this decision generates an externally detectable signal. In our context, this is essential: visible reactions may serve as input that an adaptive adversary use to deduce defense policies and thresholds.

We therefore define measuring as the act of committing to a discernible defensive choice. In this framework, measurement “collapses” a partly ambiguous belief state into a definitive action that can be witnessed (either directly or indirectly) by the attacker. Conversely, non-measurement pertains to abstention: the defense intentionally postpones or conceals outwardly evident commitment while persistently gathering evidence. Abstention so maintains ambiguity from the attacker’s viewpoint, possibly delaying inference and reducing the likelihood of self-inflicted damage.

4.2 Minimal formalism: state, measurement, and loss decomposition

I adopt a minimal superposition-style representation to express “act” versus “abstain” as competing decision modes. Let the defender’s internal decision state be:

$$|\psi\rangle; =; \alpha, |act\rangle; +; \beta, |abstain\rangle, \quad (2)$$

where α and β are coefficients reflecting the defender’s current propensity to commit or to abstain given available evidence (e.g., uncertainty level, perceived threat severity, operational constraints). In this representation, commitment is realised through a measurement operator M that maps the internal decision state to an externally realised outcome, including any observable side-effects:

$$M : |\psi\rangle; \mapsto; (a, y), \quad (3)$$

where $a \in A$ is the selected action and y captures what becomes observable externally (e.g., timing, intensity, or presence/absence of a response). In adversarial environments, y is the channel through which attacker inference occurs.

To evaluate decisions, we decompose the defender’s total loss into three components:

$$L; =; L_S; +; \lambda, L_O; +; \mu, L_\ell. \quad (4)$$

Here, L_S is security loss, capturing miss/attack impact (e.g., successful intrusion, lateral movement, or data loss). L_O is operational loss, capturing costs such as false-positive

disruption, response overhead, and analyst burden—motivated by SOC realities including alert fatigue [16]. Finally, L_ℓ is leakage (observability) loss, capturing the extent to which the defender’s observable behaviour helps an attacker infer policy or optimise probing and exploitation. This aligns with defensive deception and strategic-adversary perspectives where observable defender actions shape attacker learning [12,13], and with evidence that adversaries can adapt to detection and response mechanisms in realistic NIDS settings [17].

The defender’s objective is to minimise expected total loss:

$$\min_{\pi} \mathbb{E}[L_s; +; \lambda, L_o; +; \mu, L_\ell], \quad (5)$$

where π is the defender policy that selects when to commit (measure) and which action to take, versus when to abstain (delay or suppress observable commitment). This formulation makes abstention relevant even when it does not immediately reduce L_s : abstention can reduce L_o and L_ℓ , which can dominate in regimes where operational burden or attacker inference is the primary threat multiplier.

4.3 Abstention policy: conceptual decision rules

Building on selective prediction and uncertainty-aware decision-making, we define abstention as a policy choice that becomes preferable in three broad regimes.

High uncertainty regime. When the defender’s uncertainty is high (e.g., weak evidence, ambiguous anomaly signals, low-confidence classification), abstaining can prevent forced low-quality responses. This is consistent with selective classification and calibrated abstention principles, where deferring decisions helps control risk on accepted decisions [5–7,10], and with uncertainty estimation approaches that support abstention when confidence is unreliable [11].

High leakage regime. When committing to an observable response would provide high-value feedback to the attacker—e.g., revealing thresholds, response logic, or enabling rapid policy inference—abstaining can be preferred to reduce the information channel available to adaptive adversaries. This aligns with defensive deception and game-theoretic security perspectives that treat information shaping as a core defensive mechanism [12,13].

High operational cost regime. When the operational cost of action is high—such as during alert storms, analyst overload, or when false positives are expensive—abstaining (or delaying commitment) can reduce self-inflicted harm. This is directly motivated by SOC research on alert fatigue and the operational consequences of excessive signalling and response [16].

In summary, the quantum-inspired abstraction provides a compact way to represent a defender’s choice between committing (measurement) and intentionally not committing (abstention), while explicitly incorporating operational cost and adversarial leakage alongside traditional security impact.

5 Illustrative Cyber Defense Scenarios

This section offers qualitative validation via exemplary cases. The objective is not to assert actual optimality, but to illustrate how abstention arises as a logical primary action when we explicitly consider uncertainty, operational costs, and knowledge leaking to adaptive adversaries. For each case, we delineate the observation background, the defender’s choice, the insights an attacker may get from the defender’s reaction, and the advantages of abstention.

5.1 Scenario A: SOC triage under alert overload

Observation. A SOC receives a surge of alerts produced by multiple detection sources (signature rules, anomaly detectors, EDR signals). Many alerts are weakly supported and partially redundant, and the SOC has limited analyst capacity. In this setting, the raw volume of alerts increases the probability of slow triage, missed escalation, and inconsistent response.

Decision. An action-biased policy tends to “signal” by escalating and notifying widely, or “intervene” by applying broad blocks and containment steps whenever an alert crosses a simple threshold. In contrast, an abstention-aware policy deliberately delays or suppresses outward-facing actions for low-confidence alerts while continuing silent evidence collection (e.g., enhanced logging, correlation, and time-window aggregation). Operationally, this can be implemented as rate-limited escalation, deferred ticket creation, or quiet enrichment without immediate analyst interruption.

What the attacker learns. Visible, immediate actions (rapid blocks, obvious containment, or predictable escalation timing) can reveal which signals trigger response and how quickly the defender reacts. Even if the attacker cannot see internal SOC workflows, externally observable changes (connection resets, blocked IPs, altered service behaviour) can serve as feedback. Abstention reduces the amount of immediate feedback available to an attacker and reduces the probability that a low-confidence alert triggers a noisy, externally detectable response pattern.

Why abstention helps. SOC alert fatigue is a documented operational risk: excessive signalling creates cognitive overload and can degrade detection and response quality [16]. By abstaining on low-confidence cases i.e., withholding premature visible commitment the defender reduces operational loss (fewer unnecessary escalations and disruptions) and avoids producing clear reaction patterns that could be exploited. In this regime, abstention functions as a capacity-preserving action that maintains response quality for the subset of events that truly warrant commitment.

5.2 Scenario B: NIDS probing and adversarial threshold inference

Observation. A network intrusion detection system (NIDS) identifies anomalous traffic patterns. An adaptive attacker, suspecting that the defense employs threshold-based anomaly scoring or rule triggers, initiates the transmission of meticulously designed probes to ascertain the conditions that activate blocks, throttling, or alarms. These probes are economical and intended to delineate decision limits (e.g., packet sizes, timing, protocol fields, request rates).

Decision. An action-biased defense commits early: it blocks or rate-limits when the score crosses a threshold, or it generates visible warnings that cause measurable traffic effects. An abstention-aware defense avoids committing to externally observable actions when the confidence is uncertain and when it appears that the traffic may be part of boundary testing. Instead, the defender chooses abstention actions such as silent logging, delayed responses, randomized response timing, or accumulating additional evidence before committing.

What the attacker learns. If the defender reacts consistently, the attacker can use observed outcomes (blocked vs not blocked, latency spikes, connection resets) as labels to infer the defender’s decision rule. Realistic adversarial settings for NIDS explicitly consider that attackers can adapt to and exploit detector behaviour, including by crafting traffic that evades detection while learning what triggers responses [17]. From a strategic perspective, this is exactly the kind of feedback channel that defensive deception and game-theoretic security models treat as pivotal: the defender’s observable actions influence adversary beliefs and future strategy [12,13].

Why abstention helps. Abstention reduces the clarity and immediacy of the feedback channel. By delaying or suppressing visible commitment during suspected probing, the defender forces the attacker to operate with higher uncertainty about which behaviours are safe, increasing attacker cost and slowing adversarial learning. Even when abstention does not immediately reduce the probability of attack, it can reduce leakage loss by withholding the information that would otherwise allow the attacker to rapidly approximate the defender's policy.

5.3 Scenario C: Moving Target Defense versus abstention under high operational cost

Observation. A defender considers deploying moving target defense (MTD) techniques such as rotating IP addresses, randomising service configurations, or frequently changing attack surfaces. The environment, however, has constraints: frequent changes can break dependencies, incur downtime risk, or require significant operational effort. The defender faces a trade-off between proactive change and stable operations.

Decision. MTD is designed to increase attacker uncertainty by continually shifting the target, but it can be costly to operate at high frequency or across complex systems [14,15]. In high-cost regimes, an abstention-aware strategy may choose to “do less” visibly on low-confidence threat signals rather than triggering expensive rotations or disruptive interventions. For example, instead of immediate rotation in response to ambiguous scanning, the defender may abstain by quietly collecting evidence, limiting outward signals, and only committing to MTD actions once evidence crosses a stronger threshold.

What the attacker learns. Frequent, deterministic rotations in response to ambiguous signals can inadvertently teach the attacker which probes trigger movement, enabling them to adapt probes that avoid or exploit rotation patterns. Conversely, if the defender abstains from visible commitment under uncertainty, the attacker receives less reliable information about whether they have triggered a defensive state transition.

Why abstention helps. In environments where the operational cost of MTD is high, abstention can be the dominant choice for low-confidence or ambiguous events because it avoids expensive “move/rotate” actions that may not be justified by evidence [14,15]. Abstention does not replace MTD; rather, it acts as a gatekeeping action that preserves scarce operational capacity and reduces unnecessary defensive churn. This highlights a practical role for abstention: it can complement proactive defenses by preventing overreaction when uncertainty is high and the operational cost of action is substantial.

6 Quantitative Evaluation (Simulation Study)

To address the primarily conceptual nature of our contribution and to provide measurable outcomes, we instantiate the loss decomposition in Eq. (4) in a lightweight simulation. The goal is not to claim empirical optimality, but to demonstrate how abstention can shift the trade-off among security loss, operational cost, and leakage-driven adversarial learning.

6.1 Setup

I consider a streaming detection setting with a noisy detector score s_t generated from benign or adversarial events. At each step, the defender chooses an observable action $a_t \in \{\textit{intervene}, \textit{none}, \textit{abstain}\}$. Interventions represent overt responses (e.g., block, quarantine, escalation), whereas abstention denotes deliberate non-commitment (no externally

Table 1. Lightweight simulation instantiation of Eq. (4) (seed 42; $T = 20,000$; $p_{\text{probe}} = 0.02$; $p_{\text{attack}} = 0.06$; $\tau = 0.7$; $b = 0.08$; $c_{\text{FN}} = 1.0$; $c_{\text{int}} = 0.2$; $\lambda = 1.0$; $\mu = 0.5$). L_s and L_o are per-step averages; $L_\ell \in [0, 1]$ is the fraction of probes that yield an informative observation (abstention yields none).

Policy	L_s	L_o	L_ℓ	L
π_{thresh}	0.0151	0.0120	1.0000	0.5271
π_{rand}	0.0168	0.0131	1.0000	0.5299
π_{abst}	0.0295	0.0059	0.0050	0.0379
π_{always}	0.0000	0.2000	1.0000	0.7000

informative response) under uncertainty. In addition to natural events, the adversary occasionally issues probing events designed to learn the defender's decision boundary from observable outcomes.

I simulate $T = 20,000$ steps with probe rate $p_{\text{probe}} = 0.02$ and base attack rate $p_{\text{attack}} = 0.06$. Detector scores follow clipped Gaussian mixtures: $(\mu_{\text{benign}}, \sigma_{\text{benign}}) = (0.35, 0.15)$ for benign events and $(\mu_{\text{attack}}, \sigma_{\text{attack}}) = (0.78, 0.12)$ for attacks. Probing scores are sampled near the boundary $\tau = 0.7$ from $\mathcal{N}(\tau, 0.03^2)$. We report results for seed 42 and use abstention-band width $b = 0.08$.

6.2 Defender policies

I compare four simple policies: (i) π_{thresh} : intervene if $s_t > \tau$, else none; (ii) π_{rand} : probabilistic intervention with probability increasing in s_t around τ ; (iii) π_{abst} : abstain within a band $|s_t - \tau| < b$, intervene if $s_t > \tau + b$, else none; and (iv) π_{always} : always intervene. These baselines are intentionally minimal to isolate the effect of abstention on leakage and operational load.

6.3 Metrics mapped to Eq. (4)

I operationalize Eq. (4) using simple proxies. Security loss L_s penalizes missed adversarial events (false negatives) with cost $c_{\text{FN}} = 1.0$. Operational cost L_o counts overt interventions with cost $c_{\text{int}} = 0.2$ (representing analyst time, disruption, or automation cost). Leakage loss L_ℓ measures adversarial learning from probes as the fraction of probing steps that produce an informative observation (*intervene/none*); abstention yields no informative observation. Total loss uses $L = L_s + \lambda L_o + \mu L_\ell$ with $\lambda = 1.0$ and $\mu = 0.5$.

6.4 Results

Table 1 reports one representative run under the above parameterization. Compared to π_{thresh} , the abstention-band policy π_{abst} sharply suppresses leakage from probing while also reducing operational interventions, at the cost of increased missed attacks when abstention is applied too aggressively. This supports the paper's central claim: abstention is a meaningful defensive action that can be tuned to trade coverage for reduced leakage and reduced self-inflicted operational harm, rather than being treated as mere inaction.

7 Discussion

This section examines the circumstances in which abstention is suitable, when it is inappropriate, and the practical and ethical limitations that must follow any defense strategy that is cognizant of abstention. The primary assertion is that abstinence is a proactive measure in the face of uncertainty, rather than a mere directive to refrain from action.

7.1 When abstention should not be used

When there is strong evidence that a deal is happening or will happen soon and there are bad effects for waiting, abstention is not a good idea. In situations with a high chance of being true, the small benefit of keeping the attacker guessing is usually outweighed by the security loss that comes with delaying control. As an example, abstention should not be used when (i) signs come from different sources that agree with each other (for example, multiple telemetry streams converge on the same event), (ii) the event fits known high-severity patterns (for example, confirmed credential theft, active lateral movement), or (iii) the system is in a safety-critical state where delay could cause too much harm. In these situations, the guard should commit to either intervening or at least sending strong signals, even if others can see what they are doing.

Abstention is also unsuitable when the defender's environment lacks adequate sensing to "abstain safely." If abstention simply means suppressing response without improving evidence collection, it can become indistinguishable from neglect. Therefore, abstention should be coupled with continued monitoring, enrichment, correlation, and readiness to escalate once uncertainty reduces.

7.2 Abuse risk and ethics: abstention is not an excuse

If you treat silence as an official activity, it could be abused. To cut down on work, operational teams may be drawn to name tough or resource-intensive events as "abstain," even when they need to take action. This is a problem with ethics and governance because it puts the risk on users, customers, or partners further down the line. This type of failure can be avoided by limiting abstention with clear rules and ways to be held accountable. Some examples are (i) time limits on abstention that must be reviewed, (ii) writing down the reasons (like uncertainty and expected costs), and (iii) being able to check the decisions and results of abstention.

A related concern is fairness and harm distribution: systematic abstention in certain contexts (e.g., noisy segments, low-visibility assets, or specific user populations) could create uneven protection. Even though this work is conceptual, any real deployment should treat abstention not only as an optimization decision but also as a governance-controlled action that is monitored for unintended consequences.

7.3 Operational implementation guidelines (SOAR-ready)

Abstention should be implemented as an explicit playbook state rather than a passive absence of response. Practically, "abstain" means withholding externally informative actions while continuing internal sensing, evidence collection, and readiness to commit once uncertainty reduces.

Trigger conditions (when to abstain). Abstain when (i) evidence is ambiguous or conflicting; (ii) there are indicators of adversarial probing intended to infer thresholds; or (iii) operational load is saturated (e.g., alert flood), such that overt responses would amplify self-inflicted cost.

Safe-abstain safeguards. During abstention:

- increase internal telemetry and correlation;
- preserve high-fidelity logs for later attribution;
- apply rate-limited, minimally informative controls (e.g., soft throttling) that do not reveal decision boundaries; and
- enforce a maximum abstention window to avoid silent failure.

Escalation and exit criteria. Exit abstention and commit to an overt action when (i) confidence crosses a predefined threshold; (ii) multiple independent signals converge; (iii) safety-critical assets are implicated; or (iv) the abstention window expires. The playbook should support human override and post-incident review of abstention decisions. These operational choices align with the simulation trade-offs in Section 6: increasing the leakage penalty favors abstention, whereas increasing the miss penalty favors earlier commitment.

7.4 Implementation notes: mapping abstention to SOAR playbooks

One useful thing about this approach is that abstention fits easily into current models for security management and automation. In SOAR-style processes, refusal can be modeled as clear plan steps that control commitment and visibility, such as

- Delay: put off actions that affect other people (like blocking or notifying) for a limited time while you gather more proof (for example, by adding to it or finding a link between different sources).

- Increase data capture and investigative context without making changes that attackers can easily see (for example, better packet capture and more terminal logs).

- Rate-limited response: lower the number of alerts, escalations, or control steps to stop alert storms and avoid operating overload, but keep the option to raise the level of alerting when trust grows.

Practitioners already know about these processes; this paper's addition is to turn them from ad hoc tactical tactics into a category of actions that are based on principles and are clearly modeled. It's easier to think about trade-offs like security effect vs. practical cost vs. information leaks and to make rules that decide not only what to do but also when not to commit yet.

8 Conclusion

This paper argues that effective cyber defense is not only a question of *what* action to take, but also *when not to act*. By treating abstention as a first-class security action, we reframe “doing nothing” from an operational failure into a deliberate policy choice that can be optimal under uncertainty—particularly when visible defensive commitment imposes high operational cost or creates exploitable information leakage for adaptive adversaries. Using a quantum-inspired decision abstraction, we interpret commitment as a measurement-like collapse that produces observable signals, and we show conceptually how abstention can preserve uncertainty, reduce self-inflicted harm, and slow attacker inference without requiring quantum hardware.

In the future, these ideas should be tested in real-life situations. Some interesting directions to look into are (i) controlled experiments using SOC-style alert streams to figure out how much operational benefits abstention policies have, (ii) red-team simulations where attackers learn from defenders' responses to find out how much leakage is reduced, and (iii) formal extensions using explicit game-theoretic models to describe how equilibrium behaves when abstention is an action. These suggestions can help make the case for abstention-aware security

systems stronger and give useful advice on how to use uncertainty-preserving defense methods on a large scale.

References

- [1] C. K. Chow, On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory* **16**, 41–46 (1970). doi: 10.1109/TIT.1970.1054406
- [2] R. Herbei, M. H. Wegkamp, Classification with reject option. *Can. J. Stat.* **34**(4), 709–721 (2006). doi: 10.1002/cjs.5550340410
- [3] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, S. Canu, Support vector machines with a reject option. In *Advances in Neural Information Processing Systems (NeurIPS)* (2008). https://publications.idiap.ch/downloads/papers/2009/Grandvalet_NIPS_2008.pdf
- [4] C. Cortes, G. DeSalvo, M. Mohri, Learning with rejection (2016). <https://research.google/pubs/learning-with-rejection/>
- [5] Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)* (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf
- [6] Y. Geifman, R. El-Yaniv, SelectiveNet: A deep neural network with an integrated reject option. *arXiv arXiv:1901.09192* (2019). <https://arxiv.org/abs/1901.09192>
- [7] A. Fisch, T. Jaakkola, R. Barzilay, Calibrated selective classification. *arXiv arXiv:2208.12084* (2022). <https://arxiv.org/abs/2208.12084>
- [8] A. Gangrade, A. Kag, V. Saligrama, Selective classification via one-sided prediction. *arXiv arXiv:2010.07853* (2020). <https://arxiv.org/abs/2010.07853>
- [9] V. Franc, D. Průša, V. Voráček, Optimal strategies for reject option classifiers. *arXiv arXiv:2101.12523* (2021). <https://arxiv.org/abs/2101.12523>
- [10] M. Hasan *et al.*, Survey on leveraging uncertainty estimation towards trustworthy deep neural networks: The case of reject option and post-training processing. *arXiv arXiv:2304.04906* (2023). <https://arxiv.org/abs/2304.04906>
- [11] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *arXiv arXiv:1506.02142* (2016). <https://arxiv.org/abs/1506.02142>
- [12] J. Pawlick, E. Colbert, Q. Zhu, A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. *ACM Comput. Surv.* **52**(4), 1–28 (2019). doi: 10.1145/3337772
- [13] A. Schlenker *et al.*, Deceiving cyber adversaries: A game theoretic approach. In *Proc. Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS)* (2018). <https://par.nsf.gov/biblio/10050303-deceiving-cyber-adversaries-game-theoretic-approach>
- [14] S. Sengupta, A. Chowdhary, A. Sabur, A. Alshamrani, D. Huang, S. Kambhampati, A survey of moving target defenses for network security. *arXiv arXiv:1905.00964* (2020). <https://arxiv.org/abs/1905.00964>
- [15] J.-H. Cho *et al.*, Toward proactive, adaptive defense: A survey on moving target defense. *arXiv arXiv:1909.08092* (2019). <https://arxiv.org/abs/1909.08092>
- [16] S. Tariq, M. B. Chhetri, S. Nepal, C. Paris, Alert fatigue in security operations centres: Research challenges and opportunities. *ACM Comput. Surv.* **57**(9) (2025). doi: 10.1145/3723158
- [17] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, M. Colajanni, Modeling realistic adversarial attacks against network intrusion detection systems. *Digit. Threat. Res. Pract.* (2021). doi: 10.1145/3469659

- [18] A. Łukasik, Quantum models of cognition and decision. *Int. J. Parallel Emerg. Distrib. Syst.* **33**(3), 336–345 (2018). doi: 10.1080/17445760.2017.1410547
- [19] J. M. Yearsley, J. R. Busemeyer, Quantum cognition and decision theories: A tutorial. *J. Math. Psychol.* **74**, 99–116 (2016). doi: 10.1016/j.jmp.2015.11.005
- [20] J. M. Arrazola, A. Delgado, B. R. Bardhan, S. Lloyd, Quantum-inspired algorithms in practice. *Quantum* **4**, 307 (2020). doi: 10.22331/q-2020-08-13-307

9 Acknowledgments

The author thanks the research community whose open-access publications and software tooling made this conceptual study possible. This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The views expressed are solely those of the author.

A Research Methods

This paper follows a qualitative, conceptual research design. Rather than proposing a new dataset or reporting controlled experiments, I develop a theory-grounded model of *abstention as a first-class cyber defense action* and validate its plausibility through structured, literature-supported scenarios. The method is designed to (i) anchor the concept in established theory (reject-option / selective prediction), (ii) reflect realistic operational and adversarial constraints (SOC operations, adaptive attackers), and (iii) introduce a quantum-inspired abstraction used as a modelling lens rather than a claim about quantum hardware.

A.1 Part One: Theory synthesis and conceptual model construction

I conduct a structured synthesis of three literature streams and use them to construct an integrated conceptual model.

Stream 1 (Reject option / selective prediction). I extract core principles that justify abstention as a rational decision under uncertainty (e.g., error–reject trade-offs, risk–coverage reasoning, and calibrated selective decision-making). These foundations support the argument that abstention is not an operational failure, but a deliberate action that can reduce expected risk when evidence is weak or costs are asymmetric [1, 2, 5–7, 10, 11].

Stream 2 (Cybersecurity operations and adaptive adversaries). I incorporate cybersecurity perspectives where defensive actions are not costless and may be strategically observed. Specifically, I draw on deception and strategic-interaction viewpoints to motivate *leakage-driven attacker inference* from defender responses [12, 13], use operational evidence on analyst overload to motivate an explicit operational cost term [16], and ground the adaptive-attacker assumption in work modelling realistic adversarial behaviours against intrusion detection settings [17].

Stream 3 (Quantum-inspired decision framing). I adopt a quantum decision/cognition lens to formalise the distinction between *commitment* and *non-commitment* as an abstract modelling device: commitment is treated as a measurement-like operation that collapses uncertainty into a concrete (and potentially observable) outcome [18, 19]. I further justify the use of a *quantum-inspired* framing as a legitimate modelling approach that can yield practical insight without quantum hardware [20].

Model construction procedure. Using these streams, I define (i) an action set that includes abstention, (ii) an attacker feedback channel representing what the adversary can

observe, and (iii) a minimal loss decomposition separating security loss, operational cost, and leakage (observability) loss. This produces the quantum-inspired abstention model presented in Section 4.

A.2 Part Two: Scenario-based qualitative validation

Because the contribution is conceptual, I validate the model through a scenario-based qualitative method. I develop three illustrative scenarios (SOC triage, NIDS probing, and MTD vs. abstention) chosen for their relevance to (i) operational constraints, (ii) adaptive attacker behaviour, and (iii) proactive defense trade-offs. For each scenario, I apply a consistent analysis template:

1. **Observation:** specify the signal context and uncertainty drivers.
2. **Decision:** compare action-biased response (intervene/signal) against abstention.
3. **Attacker learning:** describe what information the adversary can infer from observable responses.
4. **Benefit of abstention:** explain how abstention reduces operational burden and/or leakage-driven learning, and under what assumptions this holds.

This approach does not claim empirical optimality. Instead, it provides *plausibility and coherence checks*: the scenarios demonstrate how abstention can be justified by existing theory (reject option), operational reality (alert fatigue), and strategic interaction (attacker inference), while remaining compatible with proactive defenses such as moving target defense [14–17].

A.3 Scope, assumptions, and limitations of the method

The method is intentionally minimal: it is designed to introduce abstention as a first-class security action and to clarify when it may be rational. The paper does not evaluate a deployed system or benchmark performance across datasets. As a result, the claims are bounded to conceptual validity and literature-grounded plausibility. Empirical evaluation, red-team simulations, and formal game-theoretic extensions are treated as future work rather than requirements for this initial conceptual contribution.

B Online Resources

All referenced materials are accessible through stable online identifiers. For peer-reviewed articles, I prioritise DOI links, and for preprints I cite canonical arXiv landing pages. When an item is not hosted on a publisher platform (e.g., workshop or conference copies), I cite an official archive or institutional landing page to ensure long-term accessibility. The URLs listed in the bibliography therefore serve as the primary online resource index for this paper.

In addition, the conceptual diagram and tables included in this manuscript are fully reproducible from the \LaTeX source (TikZ and tabular environments) and do not depend on external image assets. This design choice minimises link rot and ensures that the paper remains portable across submission systems. Where web pages were used to access a preprint or supplementary copy, an access date is recorded in the reference entry to indicate the retrieval time for the online resource.