arXiv:2601.04486v1 [cs.CR] 8 Jan 2026

面向SOC警报分诊的决策感知信任信号对齐

Israt JahanChowdhury Md Abu Yousuf Tanvir安大略理工大学（加拿大） Ontario Tech University（加拿大）

摘要

基于机器学习的检测系统越来越多地应用于安全运营中心（SOC），以筛查海量安全告警。在实践中，这些系统经常提供概率输出或置信度分数，但这些信号往往校准失准，在实际压力下也难以解读。有关安全运营中心实践的既有定性研究和基于调查的研究结果表明，告警质量和告警过载与分析师的决策负担和难度相关，尤其是在工具输出噪声较大，或难以据此采取一致行动时 [3],[4]。其主要弱点在于，置信度这一概念往往未明确对应不对称的决策成本：与漏掉攻击相比，误报造成的危害要小得多。

本文提出一种用于SOC警报分诊的决策感知信任信号对齐框架。该框架不改变检测模型，而是将 (i) 校准置信度、(ii) 轻量级不确定性线索和 (iii) 成本敏感决策阈值映射到一个协调一致的决策支持层 [6]。校准采用常见的事后校准策略，这些策略已被证明能够提高概率预测的一致性 [5], ；不确定性线索则在模型预测不确定时提供保守保护 [6]。

为衡量我们在 UNSW-NB15 入侵检测基准上评估该方法时的模型无关效应， [1],[2] 我们使用两种不同的分类器：逻辑回归和随机森林。仿真结果表明，未对齐的置信度显示会显著增加假阴性，而决策对齐的信任信号可使成本加权损失降低几个数量级。我们还提供了一套人在回路研究方案，以便未来评估分析师在安全运营中心中的工作，例如分诊任务。

关键词： 安全运营中心（SOC）、警报分类、可信人工智能、概率校准、不确定性估计、成本敏感型决策、人机交互、入侵检测。

1 引言

1.1 背景

当前的安全运营中心（SOC）建立在自动化检测系统之上，用于监测极其复杂的基础设施，并从海量数据中识别恶意活动。机器学习在入侵检测中的应用已逐渐成为一种常见实践，因为机器学习能够对高维网络遥测数据进行建模，并适应各种攻击模式 [7],[8]。工作环境中使用的检测模型从不独立运行：这会持续产生大量警报，而这些警报必须由人工分析师进行筛选、检查和升级处理。不同于传统的入侵检测系统研究，本文不提出新的检测模型，而是关注如何在安全运营中心环境中解读现有模型输出并据此采取行动。

其中一些长期存在的问题包括警报疲劳，它由大量警报、误报频发以及时间压力共同造成，进而导致分析师注意力下降，决策的一致性也随之降低 [4]。根据对安全运营中心工作的实证描述，另一个重要因素是警报质量会显著影响分析师采取的行动和工作负荷；信号质量也会影响分诊质量以及所检测到的安全事件的质量 [3],[1]。在这方面，假阴性（漏报攻击）的概率与假阳性（需要进一步调查）并非成比例相关，因此，SOC分诊实际上是成本敏感的。

1.2 问题陈述

为了支持大多数安全运营中心工具，模型输出通常采用预测标签、排序或置信度分数的形式。尽管这些信号可能被视为可信度的体现，但它们通常并非针对成本意识型决策而设计。缺乏置信度信息时，分析师可能会对高分过度信任、对低分信任不足，这会带来危险。从更广泛的意义上说，目前还不存在可操作的信任，也就是置信度：概率在统计层面上或许准确，但若未与升级处置和关闭所产生的影响相联系，在实际操作中就可能并不准确。

1.3 研究空白

大多数入侵检测系统研究着重于提升模型层面的准确率、检测率或曲线下面积 [7],[8]。与此同时，可信人工智能研究也不断发展，提出了校准技术和不确定性估计技术 [5],[6], ，以及能够以模型无关的方式解释预测结果的可解释人工智能技术 [9],[10]。然而，这些研究方向似乎从未在以决策为中心的安全运营中心场景中得到整合。具体而言，关于信任信号（置信度、校准、不确定性）如何影响分诊决策，研究寥寥无几——

也很少有实证研究探讨现有系统如何将呈现的信任显示与漏报攻击和误报的实际成本模型相匹配。这促使我们采用一种新范式，将信任视为一种决策支持构念，而非粗略的模型输出。

1.4 贡献

本文作出以下贡献：

- 我们提出了一个面向决策的SOC警报分诊信任信号对应系统，明确向运营决策者指示信任信号。
- 我们表明，即使在底层检测模型保持不变的情况下，未对齐的置信度显示也可能导致假阴性率上升。
- 我们提出了一种成本敏感的信任对齐机制，该机制结合事后校准与不确定性感知防护措施，以作出升级处置决策。
- 我们在UNSW-NB15基准数据集上，对逻辑回归和随机森林分类器进行了模型无关的实证验证。
- 为促进未来对分析师或经过训练的代理使用方式的研究，我们设计了一套适用于安全运营中心风格分诊的人机协同评估协议。

2 相关工作

2.1 基于机器学习的入侵检测

基于机器学习的入侵检测系统（IDS）已作为大规模网络中恶意行为识别的方法得到广泛研究 [29]。KDD99、NSL-KDD、CIC-IDS 和 UNSW-NB15 [1],[2] 等公开基准数据集，使得在受控环境下评估检测方法成为可能。其中，UNSW-NB15 凭借最新的攻击场景和更丰富的特征表示，被认为更加贴近现实。

人们提出了许多不同的机器学习和深度学习模型用于入侵检测系统（IDS），其中包括逻辑回归、支持向量机、随机森林、卷积神经网络和循环架构 [7],[8]。相关文献主要关注提升检测准确率、召回率和曲线下面积。尽管这些进展确实显著提升了预测性能，但它们大多假设检测是一项独立的分类任务，很少考虑人工操作员在运营分诊中如何使用模型输出 [11],[29]。

2.2 可信且可解释的人工智能

在提升检测性能的同时，可信人工智能领域也在探索增强机器学习系统可靠性和可解释性的方法。概率校准方法旨在实现这样一种特性：当模型以置信度量表给出分数预测时，预测分数应能反映实际准确率的概率，并尝试纠正当前模型中系统性的过度自信或信心不足[5],[18],[27]。这些方法对于决策支持系统日益重要，因为此类系统会使用概率结果来指导人工判断[21],[28]。

预测性特征解释已被广泛用作可解释人工智能（XAI）方法，例如 LIME 和 SHAP [9],[10],[19],[25]。这些技术通过凸显具有影响力的特征和决策依据来提升透明度。然而，解释并非建立适当信任的唯一因素。先前的一项研究表明，即使预测结果并不可靠，用户也可能误解解释，并对看似可解释的模型产生过度信任 [24]。因此，不能将解释与经过校准或与决策对齐的信任混为一谈。

2.3 安全领域的人机协作

近年来，研究人员开始研究实际安全运营中心（SOC）环境中安全分析师与自动化警报系统之间的交互。定性研究表明，分析师往往会被大量警报、质量参差不齐的警报以及不足的情境线索所淹没。这些都是导致警报疲劳和决策能力受损的主要因素 [3],[4],[18]。这些论文着重指出，分析师的行为会受到模型准确性、警报呈现方式以及置信度信号的影响 [22]。

人机协同安全系统旨在整合自动化判断与人工判断，尤其适用于高影响决策。尽管这些系统并未忽视分析师在监控中的作用，但人们往往认为，只需提供模型输出或解释[17],[18],[19]即可。关于不同信任信号设计对分析师决策的影响的实证研究，尤其是在非对称成本设置下，关于是否升级处理的研究仍然稀少 [26]。

2.4 以决策为导向的成本敏感学习

成本敏感学习处理的是不同分类错误受到不同惩罚的情形。在安全应用中，假阴性的代价通常远高于假阳性。因此，建议采用阈值和损失敏感型优化策略 [14],[15]。以往研究提出了决策理论方法，通过调整分类阈值或学习目标来考虑错误成本。在所有这些进展中，成本

敏感推理几乎没有体现在用户所见运营系统的信任指标中。成本意识型阈值通常隐含在模型中，而不是通过决策支持界面呈现。因此，分析师可能会收到统计上有效、但与运营风险不对齐的置信度分数。

然而，这些方法止步于模型正确性，并未解决信任信号如何影响成本敏感的人类决策。

总体而言，现有文献分别研究了入侵检测准确率、模型敏感度和成本敏感学习。尽管如此，尚未有任何单一框架能够在安全运营中心环境中整合信任信号、决策成本和人在环分诊。具体而言，现有系统无法将升级决策的非对称成本与置信度和不确定性的呈现相匹配。这一空白促成了本文提出的决策感知信任信号对齐框架。

3 提出的框架

本节介绍一种用于SOC警报分诊的决策感知信任信号对齐框架。其核心是确保向分析师呈现的置信度、不确定性等信任指标能够凸显出来，并直接根据安全决策的非对称成本进行调整，而不是将这些指标视为静态的模型输出。

3.1 问题表述

对我们而言，SOC警报分诊是一个二元决策问题。对于每条警报，分析师都需要作出以下选择：

$$d_i \in \{\text{Escalate}, \text{Close}\} \tag{1}$$

升级处置会启动进一步调查，而关闭则会驳回该警报。

3.1.1 成本不对称性

决策错误的成本不对称性在实际操作中确实存在。假阴性（FN）导致恶意警报被关闭，可能造成严重影响；而假阳性（FP）导致良性告警被关闭，主要会增加分析师的工作量。这种不一致与安全领域中现有的成本敏感型学习方法以及风险敏感型决策系统相符[11]。

这通过一个成本模型形式化表示：

$$C_{\text{FN}} \gg C_{\text{FP}} \tag{2}$$

3.1.2 信任信号不匹配

当代检测系统通常会提供一个预测标签和一个分数

$$p(y = 1 \mid x_i) \tag{3}$$

标签的置信度。然而，这种置信度通常会校准失准，且不再与运营决策成本相关。因此，分析师可能会对警报过度信任或信任不足，从而导致他们的决策结果出现不成比例的误判。这种差异正是推动人们建立信任信号的原因，此类信号旨在为非对称成本情境下的决策提供便利。

3.2 信任信号

建议的框架会为每条警报生成一组模型无关的轻量级信任信号。

3.2.1 置信度

基础置信度信号是检测模型根据公式 3 生成的后验概率。常用的原始置信度无法提供足够的信息，因而不足以支持安全运营中心场景下的可靠决策。

3.2.2 校准

系统性的过度自信或信心不足可以通过标准方法对原始概率进行事后校准来纠正 [5]。校准会生成一个修正后的概率（p_{cal}），更准确地估计恶意活动发生的可能性。

之所以需要校准，是因为基于未校准分数设定的决策阈值可能与实际风险不对齐，尤其是在模型输出由人工分析师直接解读时 [5]。

3.2.3 不确定性

除置信度外，该框架还基于校准后的概率生成离散不确定性信号。概率接近决策边界的警报将被归类为高不确定性，而远离决策边界的警报则被归类为低不确定性。

这种不确定性信号在存在疑虑时提供一种保护性缓冲，体现了这样一个事实：不应仅仅因为置信度略有差异，就排除不确定的通知 [6]。该框架可与 LIME 或 SHAP 等事后解释方法配合使用 [9],[10] ，并可提供特征级解释。尽管如此，解释仍被视为补充背景信息，而非主要的信任指标，这承认了可解释性并不能保证获得充分的信任。

3.3 感知决策的信任对齐

本文的主要贡献在于建立了信任线索与决策成本之间的对应关系。

3.3.1 成本敏感阈值化

给定成本模型 C_{FN} 和 C_{FP}，我们得到一个决策感知阈值：

令 C_{FN} 和 C_{FP} 分别表示与假阴性和假阳性相关的运营成本。在此成本模型下，我们推导出如下决策感知分类阈值：

$$t^* = \frac{C_{\mathrm{FP}}}{C_{\mathrm{FP}} + C_{\mathrm{FN}}} \tag{4}$$

该阈值直接体现了错误的不对称成本，并取代通常采用的固定阈值（如 0.5）；后者在运营上对风险保持中立。信任信号到决策映射。该阈值与预期决策成本的低估有关，其中 0-1损失是根据非对称运营风险加权的。

需要升级处置时，建议在以下情况下发出警报： $p_{\mathrm{cal}} \geq t^*$

这种映射确保置信度值是在决策成本的背景下理解的，而不是被视为正确性的绝对值。不确定性安全升级。

为了进一步减少假阴性，该框架实施了一条安全规则：即使警报的校准置信度略低于阈值 [11]，对于高不确定性的警报仍应升级处理。这是一种保守做法，体现了安全运营中心实践：模糊的警报应得到更多调查。

3.4 界面条件

为了考察信任信号对齐的影响，我们将区分三种界面条件，它们在信任信号的呈现和使用方式上有所不同。

3.4.1 基线界面 -C0

基线条件或原始条件仅显示预测类别标签。决策隐含地与一个设定阈值相关联，且不会显示置信度、不确定性和成本数据。

3.4.2 不匹配的信任界面-C1

匹配不佳的情况表现为未校准的置信度，且与实际情况不符。我们假设，分析师过度依赖置信度值，并采用更为僵化的升级率；当置信度的解读未结合决策情境时，实际操作中常会出现这种情况。

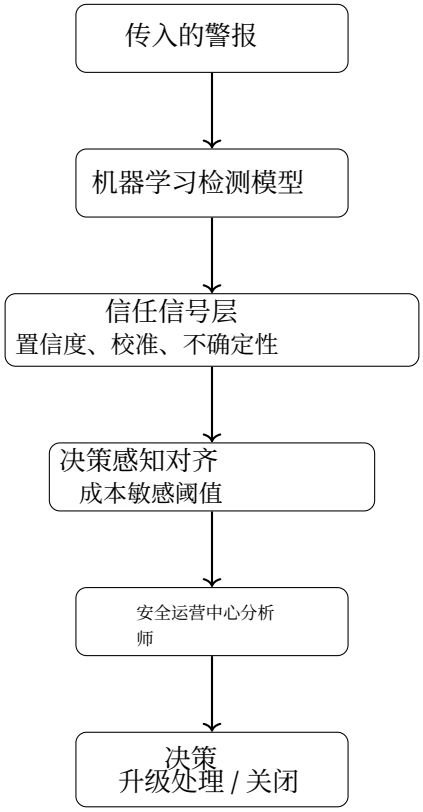


图1：用于SOC警报分诊的所提出的决策感知信任信号对齐框架概览。

3.4.3 所提出的信任接口-C2（对齐）

所提供的条件具有校准置信度，将不确定性明确呈现，并基于成本敏感阈值给出决策建议。信任信号与运营风险直接对齐，引导分析师做出风险较低的升级决策。

3.5 系统概述

图1展示了建议的决策感知信任信号对齐模型示例。通过校准和不确定性估计，模型预测结果随后被转换为信任信号。接着，这些信号与成本敏感决策策略相匹配，并呈现给分析师。该框架的优势在于明确地将检测与决策支持分离，从而使信任信号能够从运营风险的角度加以理解。

4 实验评估

本节通过实证研究探讨，将信任信号与运营决策成本对齐是否能够提升SOC警报分诊效果。与第3节以抽象方式阐述框架不同，本节完全聚焦于该框架在真实实验环境中的具体化、实现与测试。

4.1 数据集与警报流构建

这将通过构建数据集和警报流来完成。

UNSW-NB15 是一组入侵检测实验，基于 UNSW-NB15 入侵检测数据集执行。该数据集旨在捕获当前的攻击模式，同时模拟真实的背景流量[1],[2],[12]。每条记录都与一个网络流相关联，并被视为单个 SOC警报，具有良性或恶意活动这一二元真实标签。

我们遵循最初的基准划分，并使用预先安排的测试部分来仿真警报流。为了尽量减少混杂变量的数量，并能够比较不同模型的结果，我们仅保留数值特征，这与此前关于 UNSW-NB 的实证研究采用的方法类似15 [7],[12]。警报会被单独处理，这类似于典型的 SOC 分诊流程：分析师一次处理一条警报，且不进行模型重新训练或反馈。

4.2 预测输出与检测模型

本文采用两种分类器，在两组模型中研究所提方法的韧性：

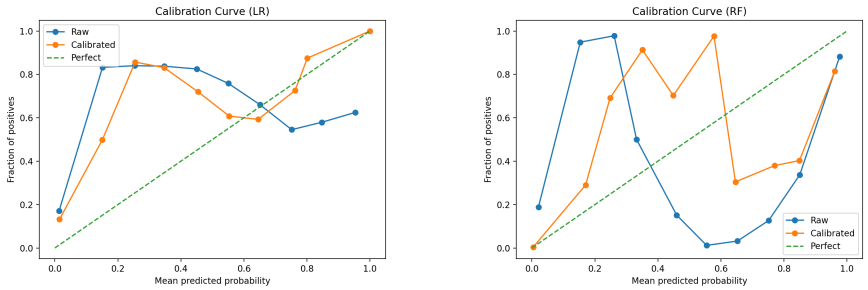
- 在线性概率模型中，采用逻辑回归（LR）。
- 随机森林（RF）是一种非线性集成分类器。

两个模型还使用相同的特征集进行训练，并为每条警报生成后验概率。 $p(y = 1 \mid x_i)$。值得注意的是，这些概率被视为评估的固定输入：测试时不进行重新训练、阈值调整或损失重加权。此设计旨在隔离信任信号与决策逻辑对齐的影响，而非提升检测能力。

4.3 信任信号构建

- 校准：由于机器学习系统在现实场景中的概率输出往往校准不佳，因此我们采用事后校准，将原始概率转换为校准后的概率 [5],[13],[14]。

具体而言，在我们的流程中：



(a) 逻辑回归（LR）。 (b) 随机森林（RF）。

图2：对比原始概率与校准后概率的可靠性图。虚线对角线表示完美校准。

– 对逻辑回归的输出进行 S 形校准（Platt 风格校准）。– 使用保序回归（一种非参数单调校准器）对随机森林的输出进行校准。

- 不确定性：通过使用固定区间进行演示，我们得到了一个不确定性类别：

– 高不确定性： $p_{\text{cal}} \in [0.45, 0.55]$**– 中等不确定性：** $p_{\text{cal}} \in [0.35, 0.45) \cup (0.55, 0.65]$**– 低不确定性：**其他情况

这些不确定性类别是界面上刻意设定的粗略决策提示，而非概率保证；再次强调，它们反映了分析师所面对的工作流程。

我们采用事后概率校准，使模型置信度能够被解读为恶意活动近似可能性的衡量。图2报告了逻辑回归和随机森林的可靠性图，将原始概率输出与校准后的概率同理想对角线进行比较。校准后的概率 p^{cal}被用作对齐信任条件（C2）中的置信度信号，而未对齐条件（C1）则呈现原始概率 [16]。

这些区间被有意设置得较低，以便作为界面指示器易于读取，并在模型几乎无法做出明确判断时提供保守决策 [6],[9],[13]。

4.4 界面条件

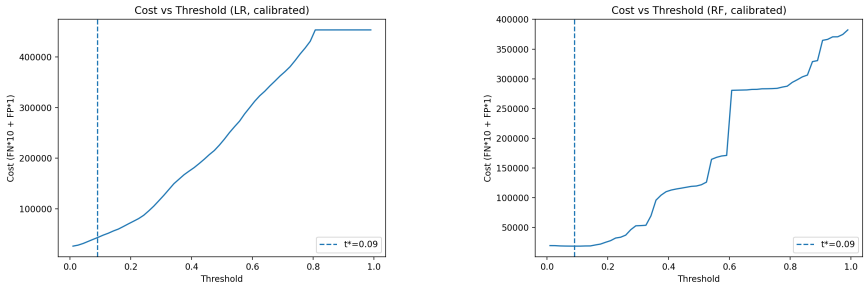
我们考虑三种界面条件，用于生成每条警报的升级处置或关闭决策。

- C0，默认阈值（基线）：决策规则：若出现 $p \geq 0.5$ ，则升级处理；否则关闭。不进行校准，也不提供不确定性提示。

- C1，信任失配（过度信任置信度）：界面揭示未校准的置信度 [28]。决策规则：如果 $p \geq 0.7$，则升级处理；否则关闭。这是该模型反映的一种失效模式（随着不确定性增加，保守的分析师对故障会变得更加谨慎，除非他们非常确信故障即将发生）：在实际运行情境中，低置信度下代价高昂的漏报不可接受，但高置信度下代价高昂的漏报却可以容忍。0.7 的阈值代表了在 SOC 环境中观察到的保守升级处理行为：分析师只有在认为置信度较高时才会采取保留态度。

- C2，对齐信任（拟议方案）：界面展示校准置信度及不确定性类别。同时，它会根据成本模型给出决策建议。决策规则：如果 $p_{\text{校准}} \geq t^*$安全覆盖：如果不确定性为高，则即使在 $p_{\text{校准}} < t^*$的情况下也提高阈值。这是对决策感知对齐的明确应用，并对存疑案例采取保留处理。

4.5 决策感知成本模型与阈值



(a) 逻辑回归 (LR)。　　(b) 随机森林 (RF)。

图 3：使用校准后的概率时，成本加权损失随升级阈值变化的情况。虚线表示根据非对称成本（$C_{FN} = 10, C_{FP} = 1$）得出的决策感知阈值 $t^* = 0.0909$ 。

图 3 展示了使用校准后的

概率时，逻辑回归和随机森林的成本加权损失随升级阈值变化的情况。在这两种情况下，随着阈值提高，期望成本迅速增加，这反映出在保守的升级策略下，假阴性会急剧增多。

我们设定：($C_{FN} = 10, C_{FP} = 1$)；在此成本模型下，决策感知升级阈值为

$$t^* = \frac{C_{FP}}{C_{FP} + C_{FN}} = \frac{1}{1 + 10} = 0.0909.$$

图3中的垂直虚线表示该阈值，它与两个模型期望成本最低的区域非常接近。这是 C2（对齐）中的唯一阈值。

安全运营中心的分诊具有不平衡性：误报通常远不如漏报攻击严重。我们采用与成本敏感决策理论一致的明确成本模型 [11],[15],[23]。

4.6 基于仿真的结果

非对称成本比 ($C_{\mathrm{FN}} \gg C_{\mathrm{FP}}$) 体现了常见的安全运营中心实践：漏报入侵造成的下游成本通常显著高于误警。我们采用一个具有代表性的比率来展示决策的影响，但所提出的框架可扩展至任意结构的成本体系，并可根据组织的风险承受度进行调整。

表1展示了所有界面条件和模型下基于仿真的分诊结果。报告假阴性（FN）、假阳性（FP）以及由此产生的成本加权损失（$10 \cdot FN + 1 \cdot FP$），以突出信任信号对齐对成本敏感决策性能的影响。

表1：基线、未对齐和对齐的信任界面下的 SOC分诊结果。成本 $= 10 \cdot FN + 1 \cdot FP$。

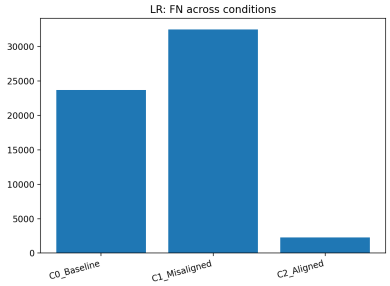
模型	条件	FN	FP	Cost
LR	C0 基线	23693	12959	249889
LR	C1 未对齐	32490	9285	334185
LR	C2 对齐	2286	20396	43256
RF	C0 基线	27400	12034	286034
RF	C1 未对齐	27509	7681	282771
RF	C2 对齐	77	18007	18777

如表1所示，与基线相比，未对齐信任条件（C1）在减少假阳性的同时，大幅增加了假阴性。相比之下，对齐信任条件（C2）通过积极减少假阴性，持续降低了两个模型的成本加权损失。

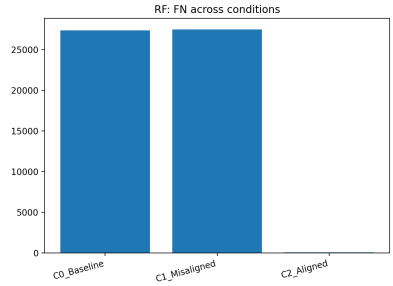
我们的结果根据假阴性（FN）、假阳性（FP）和成本加权损失计算：

成本 =10.假阴性 + 1.假阳性

逻辑回归（LR）：



(a) 逻辑回归 (LR)。



(b) 随机森林 (RF)。

- C0 基线：假阴性 $= 23{,}693$，假阳性 $= 12{,}959$，成本 $= 249{,}889$
- C1 未对齐 (0.7)：假阴性 $= 32{,}490$，假阳性 $= 9{,}285$，成本 $= 334{,}185$
- C2 对齐 (.0909 + 不确定性覆盖)：假阴性 $= 2286$ 假阳性 $= 20396$ 成本$= 43256$

当原始置信度被公开并采用严格阈值 (C1) 时，与基线相比，假阴性值显著增加；而决策感知对齐 (C2) 则使假阴性值显著降低，并使成本加权损失降至最低。

随机森林 (RF)：

- C0 基线：假阴性 $= 27{,}400$，假阳性 $= 12{,}034$，成本 $= 286{,}034$
- C1 未对齐（0.7）：假阴性 $= 27{,}509$，假阳性 $= 7{,}681$，成本 $= 282{,}771$
- C2 对齐（0.0909 + 不确定性覆盖）：假阴性 $= 77$，假阳性 $= 18{,}007$，成本 $=18{,}777$

图4：逻辑回归和随机森林在不同界面条件下的假阴性（FN）。与基线（C0）和未对齐（C1）条件相比，对齐的信任界面（C2）显著减少了漏报攻击。

我们确认，决策感知阈值化的定性表现对于多种非对称成本比（C_FN/C_FP 介于 5:1 和 20:1 之间）仍然成立，在这些比值下，对齐的信任界面始终能够优化期望成本。这表明，与仅适用于一种成本配置相比，所提出的对齐机制具有很强的稳健性。

该图表展示了两种检测模型在三种界面条件下的假阴性数量。在逻辑回归和随机森林中，对齐信任界面（C2）相比基线（C0）和未对齐（C1）条件，均显著减少了漏报攻击。这一效果与校准置信度和决策感知的升级阈值的使用相一致，二者共同促使系统以更为谨慎的方式处理不确定警报。

在不同的模型家族中也会出现相同的模式，这证明了这样一种观点：对齐是一种模型无关的决策支持层。

4.7 人在回路研究方案

为了补充仿真结果，并评估决策意识型信任信号一致性是否能提升人工分诊效果，我们引入了一套受控的人机协同协议，以反映安全运营中心的两种现实情况：（i）警报量大且存在警报疲劳；（ii）漏掉攻击所带来的运营风险。以往的实证研究记录表明，安全运营中心工具可能造成沉重的误报负担，而且分析师必须投入大量精力验证警报 [3],[4]。警报疲劳是安全运营中心持续面临的另一项挑战，它影响了决策质量和工作负荷管理 [3],[17]。

为能够对信任对齐的效果进行受控验证，我们已拟定人在回路研究方案；受资源和机构审查委员会（IRB）限制，这项优化计划留待未来开展。

4.7.1 参与者

我们招募了 40 名参与者，并分为两组：

- 25 代理分析师根据安大略理工大学的情况招募（研究生阶段的网络安全课程或密切相关课程）。
- 15从业人员受聘于一家外部行业公司（从事安全运营、信息技术安全或相关职位）。

4.7.2 材料与界面条件

试次条件：向参与者施行三种界面条件下的分诊：

- C0（基线）：不加修饰的界面（不显示任何明确的置信度/不确定性信息）。
- C1（未对齐）：未经修饰地直接显示置信度。
- C2（对齐，拟议方案）：显示置信度、不确定性和成本考量型决策建议；对于高不确定性情形，采取保守升级处置。

所有条件下的内容和呈现特征均相同，唯一的区别在于信任信号的呈现方式。

4.7.3 任务与流程

在被试内设计中，每位参与者都会接触一组固定的警报，并需要在全部三种条件下对其进行分诊：

- 熟悉环节（5-8分钟）：一组简短的说明，其中描述了分诊的目标（升级处理或关闭）以及界面上各元素的含义。

- 条件区块：参与者按照平衡化顺序（拉丁方轮换）对 C0/C1/C2 进行分诊，以消除学习和疲劳对各条件的影响。

4.7.4 每个试次的行为

参与者选择升级处理或关闭。

- 可选择给出置信度评分（例如，使用 1-5 李克特量表）
- 进入下一条警报（不反馈其是否正确）
- 为了尽量减少记忆的影响，每个条件区块中的警报都会随机打乱，但各条件下使用相同的一组警报。

4.7.5 日志记录

每个试次记录：

- 参与者标识符和参与者组（安大略理工大学代理分析师或行业从业者），
- 界面条件（C0、C1 或 C2），
- 警报标识符及其对应的真实标签，
- 参与者决定将警报升级处理 或关闭 ，
- 每条警报记录的决策时间。

4.7.6 主要和次要结果指标

成本敏感的安全运营中心风险与主要结果：

- 假阳性（FP）和假阴性（FN），
- 使用仿真评估中采用的同一非对称成本模型计算的成本加权损失。

次要结果评估效率以及主观信任行为：

- 每条警报的决策时间，以及按界面条件汇总的决策时间，
- 自我报告的信心与决策正确性之间的校准，通过比较信心评分与实证准确率计算得出。

4.7.7 分析计划

由于所有参与者都会接触所有界面条件，因此采用被试内（配对）分析设计。主要分析聚焦于基线条件（C0）与对齐信任条件（C2）之间的差异。由于不对行为测量指标的分布作任何假设，因此采用非参数配对统计检验。在标准置信水平下评估统计显著性，并结合描述性统计报告结果，以便于解读 [30]。

5 讨论

本节对实验结果进行解读，并阐释为何基于决策感知的信任信号对齐能够提升SOC警报分诊绩效。本讨论并非重复陈述数据，而是着重探讨不同界面条件下成功与失败的机制，并对SOC系统设计得出一般性结论 [6],[24]。

5.1 未对齐信任为何失效

与基线条件和信任对齐条件相比，未对齐信任条件（C1）的表现都更差，这首先可归因于过度自信的加剧以及对阈值的误用。当用户看到未经校准的原始置信度分数，且没有任何情境线索时，他们更可能将高置信度视为答案正确性的良好预测指标，尽管这些分数与实际风险的相关性很低 [6],[22],[26]。此前已有证据表明，人们往往会过度信任显然经过量化的概率输出，尤其是在时间紧迫时，从而导致自动化偏见和决策自满[14],[19]。

在 SOC 分诊中，采用过高的严重升级阈值也会加剧这一效应。正如模拟结果所示，提高置信度阈值（例如从 0.5 提高到 0.7）会减少假阳性，却大幅增加假阴性。这种做法体现了一种普遍的心理捷径：人们会识别某个认知置信度水平，并认为它对应着某种重要程度，同时完全忽视漏报攻击所带来的成本不对称性。对阈值的这种误用最终非但没有减少人为偏见，反而加剧了偏见，导致决策质量低下，尽管还有其他信息可供利用。

5.2 为什么信任对齐有效

相比之下，当信任信号直接关联到决策成本，而不是原始模型输出时，信任对齐条件（C2）会得到增强。校准置信度用于使概率值更接近经验似然，从而减少系统性过度自信。更重要的是，将置信度重新界定为决策相对度量的成本敏感决策阈值，可用于指导升级行为，使其与运营风险 [20],[25]保持一致。此外，还加入了明确的不确定性信号，这也有助于提升性能。接近决策边界的警报会被标记为不确定，并以保守方式提高其风险估计，以免边际置信度的波动导致过早驳回告警。目前的不确定性感知升级方法与先前的研究结果一致；这些结果表明，在高风险领域，谨慎处理不确定情况具有优势 [14]。成本意识与不确定性管理相结合，使决策目标

转向尽量减少运营危害，而非假警报，因此在信任对齐条件下，成本加权损失和假阴性都显著降低。

5.3 对安全运营中心设计的启示

这些结果与安全运营中心工具和界面设计直接相关。首先，这些结果表明，就运营结果而言，用户界面设计与模型性能同样重要。尽管检测模型可能保持不变，但改变信任信号的呈现方式或解读方式，可能会使成本敏感性能产生数量级的变化。

其次，研究结果表明，信任不能被视为模型的绝对特征，而应视为相对特征。置信度只有结合决策成本和不确定性考量时才有意义。脱离语境呈现粗略置信度的界面，可能会诱使用户产生不当依赖或拒绝 [15]。最后，结果表明，增量模型的升级不足以解决安全运营中心的决策失误 [27]。如果信任对齐和决策支持方面没有相应进展，即使是更出色的模型，仍可能导致较差的人机交互。

5.4 模型无关的见解

评估的另一个显著结果是，逻辑回归和随机森林模型都呈现出相同的定性模式 [3]。尽管这两种模型在表征能力和错误分布方面有所不同，但在信任失配的情况下，它们的假阴性更多；而在信任对齐的情况下，它们的成本最低。这种一致性表明，观察到的效果并不取决于某一种特定的分类器，而是源于信任信号与决策逻辑之间的相互作用 [14]。

这些结果与模型无关，从而增强了所提出框架的适用性。由于信任对齐以后处理和界面层框架的形式实现，因此无需重新训练或进行架构变更，即可将其应用于各种检测模型。这种灵活性使该方法对实际的安全运营中心部署尤为重要，因为在这些环境中，通常无法替换现有模型。

6 局限性与未来工作

6.1 局限性

尽管取得了令人鼓舞的成果，但仍需承认本研究存在一些局限性。

首先，主要评估基于分析师的仿真行为，而非对安全运营中心分析师实际工作情况的观察。尽管

仿真分析是大规模测试决策政策和系统行为的一种常用工具，但它无法充分反映现实安全运营情境中影响认知策略、情境推理和组织限制的因素。以往研究表明，经验、工作负荷和态势感知会影响分析师的决策，而要在明确构建的仿真模型中体现这些因素颇具挑战性 [18]。

其次，实验测试仅基于一个基准数据集（UNSW-NB15）开展。该数据集广泛应用于入侵检测研究，且涵盖多种攻击类型。尽管它是最受欢迎的数据集之一，但未必能代表不断变化的威胁环境以及企业网络环境的多样性。类别不平衡、特征分布和标注实践等数据集特有的特征，可能会影响绝对性能结果。这些数据集特有的特征包括类别不平衡、特征分布和标注实践，可能会影响绝对性能结果。数据集特有的特征可能是影响绝对性能结果的问题之一，例如类别不平衡、特征分布和标注实践。类别不平衡、特征分布和标注实践等问题，都是可能影响绝对性能结果的数据集特有特征。

第三，本文关注的是二元选择（升级处置或关闭）。尽管这种抽象符合典型的 SOC 分诊流程，但在实践中，实际告警管理流程中的决策空间更为细致，还包括优先级排序、延迟调查和部分升级等方面。因此，所报告的收益主要归因于保守升级决策的改进，而非完整的事件响应流程。

6.2 未来工作

这些局限性带来了若干颇具前景的研究方向。

一项重要的后续举措是对专业的安全运营中心分析师开展人在回路测试，以便直接测定其在一致和不一致的信任界面中的信任校准、认知负荷以及决策置信度。受控用户研究将有助于更深入地研究校准置信度与不确定性信号在较长工作时段内对分析师行为的影响。

未来还应将该框架扩展至多类别、多阶段告警环境，在此类环境中，分析师必须区分攻击类型、严重程度以及应对攻击的紧迫程度 [23]。层级式或序贯决策过程应纳入决策感知型信任对齐，使其更贴近 SOC 工作流及安全事件升级处置路径。

另一个重要方向是实施不断变化且因情境而异的成本模型。实际而言，假阴性和假阳性的成本可能会随时间变化，具体取决于

资产的重要性、威胁情报，或运营环境的具体情境。采用能够随时间根据不同风险概况调整决策阈值的自适应成本模型，以及更能适应风险概况变化、并在未来遇到新风险与新机遇时重新设定原始决策阈值的成本模型，可以进一步增强运营稳健性[21]

最后，未来可以研究自适应信任界面，根据分析师专业能力、工作负荷或过往交互模式，调整置信度、不确定性和解释的呈现方式。这些界面有望避免长期过度信任或信任不足，并促进安全活动中的长期人机协作。

7 结论

在当代安全运营中心（SOC）中，基于机器学习的告警已成为核心组成部分，而如何呈现模型输出并由人工分析师据此采取行动，也逐渐成为一项重大问题。具体而言，置信度分数往往不考虑决策相关成本或不确定性便直接呈现，因此容易导致系统性的过度信任或信任不足，进而造成代价高昂的分诊错误。这在安全运营中心环境中尤为突出，因为攻击的假阴性会带来严重的实际运营影响 [15],[21]。

本文提出了一种决策导向的信任信号对齐框架，将模型确定性、校准和不确定性直接关联到成本敏感型成本升级决策。该方法不调整检测模型，而是在界面层面，利用校准后的概率和基于理论依据设定的决策阈值，根据非对称错误成本匹配信任信号。该框架与模型无关，无需重新训练或进行任何架构修改，因此可用于现有的安全运营中心流水线。

与以往专注于检测准确率或独立信任机制的研究不同，本文通过建立模型置信度、运营成本与分析师行为之间的联系，将信任重新界定为与决策一致的控制问题。

对 UNSW-NB15 数据进行分析，并使用逻辑回归和随机森林模型，结果显示决策结果发生了显著变化。在每种模型中，对齐的界面条件（C2）使假阴性率较基线和未对齐条件降低了一个数量级以上；此外，在 $C_{\mathrm{FN}} = 10$, $C_{\mathrm{FP}} = 1$ 模型下，成本加权损失也大幅降低。

这些收益无需采用更复杂的决策即可实现，展现了决策感知型信任对齐的实用价值。

综合来看，这些发现表明，大多数由模型约束（可能是动脉粥样硬化性约束）导致的安全运营中心故障，实际上源于不匹配的信任界面。在传达置信度与不确定性的过程中明确纳入决策成本，

可以让安全运营中心采取更安全、更保守的分诊行为，同时不会降低模型的通用性或可部署性。在网络安全活动中实现更安全人机交互的一种切实可行且与模型无关的方法，是使信任信号与决策过程的成本保持一致 [19],[22]。

这项工作为部署人工智能系统提供了一条可行途径：当分析师面临运营压力时，他们可以依赖这些系统。其核心理念很简单：将信任重新定义为决策一致信号，而非模型属性。

参考文献

[1] N. Moustafa 和 J. Slay，“UNSW-NB15：用于网络入侵检测系统的综合数据集”，载于军事通信与信息系统会议（*MilCIS*）论文集，2015年。[2] N. Moustafa 和 J. Slay，“网络异常检测系统的评估：UNSW-NB15数据集的统计分析及其与KDD99数据集的比较”，《信息安全杂志：全球视角》，第25卷，第1–3期，第18–31页，2016年。[3] B. A. Alahmadi、L. Axon 和 I. Martinovic，“99%的假阳性：安全运营中心分析师对安全警报看法的定性研究”，载于*USENIX*安全研讨会论文集，2022年。[4] S. Tariq 等，“安全运营中心中的警报疲劳：研究挑战与机遇”，《*ACM*计算调查》，2025年。[5] C. Guo、G. Pleiss、Y. Sun 和 K. Q. Weinberger，“关于现代神经网络的校准”，载于国际机器学习会议（*ICML*）论文集，2017年。[6] A. Kendall 和 Y. Gal，“计算机视觉中的贝叶斯深度学习需要哪些不确定性？”，载于《神经信息处理系统进展》论文集（*NeurIPS*），2017年。[7] A. Pinto 等，“基于机器学习的入侵检测系统综述”，《传感器》，第23卷，第5期，2023年。[8] A. Kocher 和 G. Kumar，“用于入侵检测系统的机器学习与深度学习方法：最新进展与挑战”，《软计算》，第25卷，第15期，第9731–9763页，2021年。[9] M. T. Ribeiro、S. Singh 和 C. Guestrin，“我为什么应该信任你？解释任意分类器的预测结果”，载于*ACM SIGKDD*论文集，2016年。[10] S. M. Lundberg 和 S.-I. Lee，“解释模型预测结果的统一方法”，载于*NeurIPS*论文集，2017年。

[11] V. S. Sheng 和 C. X. Ling，“用于实现成本敏感分类的阈值化”，载于*AAAI*人工智能会议论文集，2006年。[12] I. Sharafaldin、A. H. Lashkari 和 A. A. Ghorbani，“面向生成新的入侵检测数据集与入侵流量特征描述”，载于*ICISSP*论文集，2018年。[13] A. Niculescu-Mizil 和 R. Caruana，“利用监督学习预测良好的概率”，载于*ICML*论文集，2005年。[14] P. Domingos，“MetaCost：实现成本敏感分类的一般方法”，载于*ACM SIGKDD*论文集，1999年。[15] C. Elkan，“成本敏感学习的基础”，载于国际人工智能联合会议（*IJCAI*）论文集，2001年。[16] B. Zadrozny 和 C. Elkan，“将分类器得分转换为准确的多类概率估计”，载于*ACM SIGKDD*论文集，2002年。[17] R. Werlinger、K. Hawkey 和 K. Beznosov，“人、组织与技术层面的信息技术安全管理挑战的综合视角”，《信息管理与计算机安全》，第18卷，第1期，第4–19页，2010年。[18] V. Dutt、A. A. Aponte 和 J. M. Gonzalez，“网络防御中的人类决策”，《人为因素》，第61卷，第7期，第1055–1071页，2019年。[19] K. A. Hoff 和 M. Bashir，“自动化中的信任：整合影响信任因素的实证证据”，《人为因素》，第57卷，第3期，第407–434页，2017年。[20] S. Amershi 等，“人机交互指南”，载于*ACM CHI*计算系统中的人为因素会议论文集，2019年。[21] A. Holzinger，“以人为中心的人工智能：可靠、安全且值得信赖”，*Springer*，2019年。[22] Y. Zhang、Q. V. Liao 和 R. Bellamy，“置信度和解释对人工智能系统信任的影响”，载于*ACM CHI*论文集，2020年。[23] A. Dal Pozzolo 等，“面向不平衡数据集的校准分类器”，载于*IEEE ICDM*研讨会论文集，2015年。[24] G. Bansal 等，“整体是否大于部分之和？人工智能解释对人类决策的影响”，载于*ACM CHI*论文集，2021年。[25] T. Miller，“人工智能中的解释：来自社会科学的洞见”，《人工智能》，第267卷，第1–38页，2019年。[26] D. Sacha 等，“可视分析的知识生成模型”，*IEEE*可视化与计算机图形学汇刊，第22卷，第1期，第160–169页，2016年。

[27] R. Sommer 和 V. Paxson，“封闭世界之外：将机器学习用于网络入侵检测”，载于*IEEE* 安全与隐私研讨会论文集，2010年。[28] Y. Liu 等，“深度学习中的不确定性估计”，*IEEE* 信号处理杂志，第 38 卷，第 1 期，第 16–28 页，2021年。[29] I. Goodfellow、Y. Bengio 和 A. Courville，深度学习。麻省理工学院出版社，2016年。[30] F. Wilcoxon，“基于排序方法的个体比较”，《生物统计学通报》，第 1 卷，第 6 期，第 80–83 页，1945年。