A Moral Hazard Detection Framework: Reinforcing Trust in ORAN

Khalid Ibrahim*, Mohaned Chraiti*, and Ali Ghrayeb**
* Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkiye.
** College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar.
Emails: khalid.ibrahim@sabanciuniv.edu; mohaned.chraiti@sabanciuniv.edu; and aghrayeb@hbku.edu.qa

***Abstract*—With the emergence of the Open Radio Access Network (ORAN) concept and related standardization efforts, future radio access networks are anticipated to feature elements from diverse vendors. Although the ORAN elements can authenticate as legitimate, the system may fail to meet service requirements if some network components do not adhere to their respective agreements, i.e., moral hazard. This issue raises concerns about the network's end-to-end performance, complicating fault attribution and conflict resolution. Therefore, there is a need for an automated zero-trust framework capable of continuously detecting instances of moral hazard. The complexity is exacerbated by the dynamic nature of network elements or artificial intelligence (AI) model performance, which may degrade over time intentionally (e.g., malicious tampering) or unintentionally (e.g., model obsolescence or device performance decline), limiting the effectiveness of offline testing. To address this, we develop a mechanism based on subjective logic principles, incorporating a logic-based argumentation framework that explicitly accommodates argument schemes, argument accrual, and burden of proof. Building upon this framework, we apply contract theory to incentivize compliant devices to participate truthfully in the ORAN ecosystem, thereby enhancing system performance. The simulation results show improved system efficiency and reduced operational costs.**

***Index Terms*—ORAN, subjective logic, contract theory, and zero-trust architecture.**

I. Introduction

The integration of components from various vendors in a Radio Access Network (RAN) became feasible and promising with the introduction of Open-RAN (ORAN) [1]. The primary objective of ORAN goes beyond fostering a dynamic and thriving supplier ecosystem; it aims to enhance the RAN by providing greater flexibility, interoperability, and automation, boosting the overall performance [2]. ORAN offers the advantage of a wide range of selection options for network elements along with unconventional features, such as temporarily leasing underused radio resources from private access nodes and utilizing third-party computational resources or trained models, leading to cost savings and enhanced network performance. Additionally, ORAN creates profit opportunities for private resource owners [3]. Recognized by the 3GPP standardization community, ORAN has been supported by numerous operators globally, including Rakuten Mobile, Dish Network, Verizon, Telefónica, Huawei, ZTE, and Vodafone [1], [4]. Its adoption reduces reliance on traditional infrastructure vendors, enhancing network performance, scalability, and the seamless integration of new technologies and services.

In an ORAN system, trust has emerged as a critical issue, requiring continuous proactive measures against internal and external threats [5]. While authentication mechanisms can verify the legitimacy of ORAN components, they cannot ensure end-to-end performance, a fair cost-benefit ratio or, in general, compliance with contractual terms, leaving the system vulnerable to moral hazards [6], [7]. Specifically, an ORAN element may deliver content that does not comply with the initial agreement, either intentionally (as with malicious devices) or unintentionally (due to outdated models or defective elements). Although pre-selection mechanisms can be established to choose the suitable devices under the most favorable contract, the moral hazard issue persists, often becoming apparent only after an ORAN element has been selected and the agreement established [2], [6].

Failure to address the moral hazard can compromise the system's end-to-end performance and negatively impact various factors with real-world implications, especially in mission-critical services [6]. In accordance with the Zero Trust Architecture (ZTA), trust goes beyond privacy and asserts that successful authentication within a network does not imply implicit trust. Thus, an automated supervision mechanism is crucial. In the context of ORAN, trust is assessed from two perspectives: security and performance compliance. While the former has been thoroughly investigated, the latter is still emerging, with limited research available. The National Institute of Standards and Technology (NIST) emphasizes essential procedures such as authentication, authorization, monitoring, and detection to safeguard critical infrastructure and mitigate risks [8], [9]. Security concerns and moral hazard extend to risks from open interfaces, multi-vendor integration, and misconfiguration, necessitating intelligent threat detection and proactive measures. Numerous studies have implemented ZTA to automate the continuous detection of security threats using the RAN Intelligent Controller (RIC). For instance, [10]

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 101108094. This work is supported by Tübitak under grant 122E497.

proposed a game-theoretic ZTA for automated detection of intelligent jammers in 6G RAN. Liu et al. [11] introduced multi-device anonymous authentication within ORAN and proposed a ZTA-based local and roaming identity authentication protocol.

Despite the implementation of ZTA, the issue of non-compliance due to device-related moral hazard remains unaddressed [2, Sec. VII-F]. To this end, we build on the subjective logic principle to develop a mechanism to detect the moral hazards [12]. We establish a logic-based argumentation framework that explicitly caters to argument schemes, accrual of arguments, and burden of proof. Subsequently, we use contract theory to incentivize highly compliant devices to participate in the ORAN system truthfully, enhancing system performance. Our main contributions are as follows:

- To the best of our knowledge, we are the first to address the moral hazard problem in the multi-vendor environment of ORAN.
- We introduce a novel approach to identify devices with consistently poor performance using a subjective logic-based reputation mechanism. This history-based reputation mechanism gives more weight to users with higher reputations to remain in the system.
- We apply contract theory to offer lucrative contracts to highly reputed devices, encouraging their participation in system operations while isolating poorly reputed devices that consistently perform poorly. Simulation results demonstrate the effectiveness of our proposed approach.

The remainder of this paper is organized as follows: Section I introduces the ORAN and the associated moral hazard problem. The system model and its flow are detailed in Section II. In Section III, we present the problem formulation and discuss reputation-based subjective logic. Section IV explains the application of contract theory to the previously formulated problem. Simulation results are presented in Section V. Finally, Section VI provides the conclusion of the paper.

II. System Model

A. System Model Description

We consider a generic model of an ORAN system comprising M devices from diverse vendors, denoted as $\mathcal{D} = \{D_1, D_2, \ldots, D_j, \ldots, D_M\}$ illustrated in Fig. 1. In the figure, the devices from different vendors are indicated by different colors. The ORAN elements are consistently treated as untrusted, i.e., Devices Under Test (DUTs). Accordingly, the RIC, specifically the Quality-of-Service (QoS) management block, is responsible for assessing trust by sending test signals (pilot signals) to the DUTs. The test signals are sent at random instances without prior notice to the device being tested. We assumed that for each device, D_j the RIC has the preknowledge of a set of possible inputs $\{\phi_i\}_{i=1}^{\mathcal{N}}$ and their corresponding outputs $\{\theta_{c,i}\}_{i=1}^{\mathcal{N}}$. The RIC randomly selects

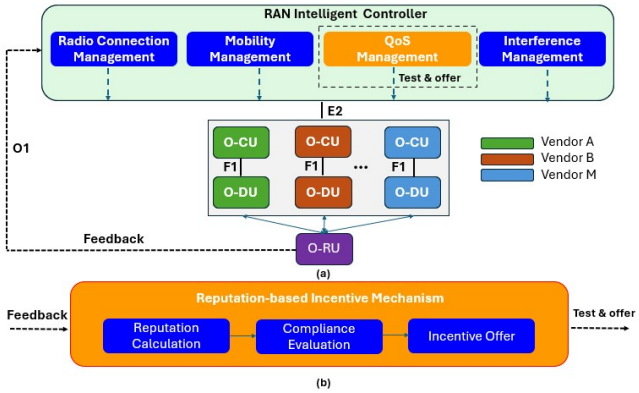

Fig. 1: a) ORAN system overview b) Reputation-based incentive mechanism as a part of the functions of the QoS management block.

a set of test inputs and records their corresponding outputs to build trust assessments for each DUT.

B. Maverick Modules

Maverick modules may deviate from system specifications in two forms: unintentional divergence and deliberate non-compliance. Some devices exhibit deviations due to inherent probabilistic variations in their outputs, such as signal classifiers at the receiver, where deviations are a natural result of stochastic behavior. Conversely, other maverick modules with moral hazards intentionally do not adhere to the pre-agreed performance, potentially degrading the system's overall functionality. In an ORAN environment, where devices from multiple vendors operate interdependently, a single non-compliant device can degrade system performance, as functionality relies on each device's adherence. Technical incompatibilities—such as proprietary hardware, incompatible software, or substandard performance—as well as operational issues like outdated firmware, insufficient testing, and security vulnerabilities, further drive non-compliance. Vendor-specific factors, including proprietary implementations and customization conflicts, also hinder adherence to ORAN standards, collectively compromising system reliability. This study centers on identifying non-compliance in DUTs.

C. System Overflow

The RIC initiates a structured protocol by issuing a randomized request to the ORAN DUT for task-specific responses, which are then transmitted to the designated testing apparatus. The testing apparatus subjects the DUT to an exhaustive suite of performance evaluations, validating the compliance of the device against predefined task-oriented metrics. Upon completing the verification process, the testing equipment communicates binary feedback (positive or negative) to the RIC based on the DUTs' adherence to performance benchmarks.

The RIC aggregates the feedback over a statistically increasing sample size, reducing the uncertainty over time.

Accordingly, the RIC computes a reputation score for the DUT. This score is derived based on the feedback from the testing equipment that reflects the device's consistency in meeting performance expectations, with special attention given to any deviations from the claimed operational specifications. The computed reputation score is then archived in the RIC's local storage, serving as a historical record for ongoing and future evaluations.

The RIC invokes a subjective logic-based algorithm, which applies a reputation-based decision-making framework to determine whether the DUT satisfies the reliability threshold required for operational trust to thwart moral hazards by the non-compliant devices. This algorithm processes the accumulated reputation score stored in the RIC's local storage, giving more preference to the recent responses over the obsolete responses, and cross-referencing it with predefined compliance criteria to make an informed decision about the device's trustworthiness. Finally, contract theory is used to attract the potentially better devices/services by reliable vendors to participate in the system contract to elevate the overall system-level experience.

The RIC's compliance decision is made in alignment with the DUT's performance history, dynamically updated based on empirical testing data, and ensuring that only devices with sufficiently robust reputations are integrated into the network ecosystem as shown in Fig. 1. In Fig. 1(a), the reputation-based incentive framework within the RAN Intelligent Controller (RIC) is pivotal for critical decision-making and incentive distribution, as further detailed in Fig. 1(b). Upon receiving feedback from various devices, the RIC's framework calculates each device's reputation for compliance evaluation. Subsequently, it formulates incentive offers to highly reputed devices, encouraging their participation in system contracts to optimize overall performance. In essence, this process encapsulates an automated, data-driven compliance framework that continuously monitors and updates the status of each device within the ORAN architecture and attracts the compliant device by offering lucrative offers to avoid moral hazard.

To assess compliance in uncertain contexts, this analysis will utilize subjective logic, examining the beliefs, disbeliefs, and uncertainties associated with each device's compliance, followed by contract-theory-based incentive mechanisms to engage compliant devices in enhancing system efficiency.

III. Problem Formulation and Proposed Solutions

A. Subjective Logic

Subjective logic is probabilistic in nature and explicitly incorporates source trust and epistemic uncertainty. It is particularly useful for representing and analyzing scenarios where information is uncertain or comes from unreliable sources [12]. In subjective logic, an opinion is defined in the form of an ordered quadruple ω, formed over a proposition—in our case, "the device D_j is compliant"—and can be written as follows:

$$\omega_{D_j} = (b_{D_j}, d_{D_j}, u_{D_j}, \alpha_{D_j}) \tag{1a}$$

$$0 \leq b_{D_j} \leq 1 \quad D_j \in \mathcal{D} \tag{1b}$$

$$0 \leq d_{D_j} \leq 1 \quad D_j \in \mathcal{D} \tag{1c}$$

$$0 \leq u_{D_j} \leq 1 \quad D_j \in \mathcal{D} \tag{1d}$$

$$b_{D_j} + d_{D_j} + u_{D_j} = 1 \quad D_j \in \mathcal{D} \tag{1e}$$

where $b_{D_j} \in [0,1]$ is belief in the compliance of the device, $d_{D_j} \in [0,1]$ is disbelief in the compliance, $u_{D_j} \in [0,1]$ is uncertainty due to lack of information, and $\alpha_{D_j} \in [0,1]$ is relative atomicity [13]. The uncertainty $u \in [0,1]$ captures the lack of evidence or the ignorance regarding the truth of the proposition "the device D_j is compliant." It represents the proportion of the total evidence that is unknown or missing, mathematically expressed as $u_{D_j} = 1-(b_{D_j}+d_{D_j})$. Conversely, the following relationship is used to relate b_{D_j}, d_{D_j}, and u_{D_j}: $b_{D_j}+d_{D_j}+u_{D_j} = 1$, as in (1e). Relative atomicity or the base rate α_{D_j} represents a prior knowledge, a prior probability, or an assumption of compliance without specific observations, i.e. $u_{D_j} = 1$. For example, if we assume most devices from reputable vendors are compliant, we could set α_{D_j} based on historical data or vendor reputation (e.g., $\alpha_{D_j} = 0.8$). The overall opinion is summed up as:

$$r_{D_j} = b_{D_j} + \alpha_{D_j} \cdot u_{D_j} \tag{2}$$

B. Reputation Calculation based on Subjective Logic

In this section, we map the subjective logic parameters to the parameters of the considered system to evaluate each device's reputation, thereby confirming its reliability within the ORAN [14]. The belief corresponds to the proportion of observed output performance $\theta_{r,i}$ that is in line with the claimed performance θ_c^j . Belief is high if most observations match the expected output.

$$b_{D_j}^{\kappa_z} = \frac{\eta \cdot g_{D_j}^{\kappa_z}}{\rho + \eta \cdot g_{D_j}^{\kappa_z} + \zeta \cdot h_{D_j}^{\kappa_z}} \tag{3}$$

where $g_{D_j}^{\kappa_z} = \sum_{i=1}^{\mathcal{N}} 1(\theta_{r,i} \approx \theta_c^j)$ is the number of positive outcomes of the device interaction during κ_z the time period with the system. This compares each observed output $\theta_{r,i}$ to the claimed performance level θ_c^j, counting the number of matches. η and ζ are the weights of the positive and negative feedback from the testing equipment [15]. η and ζ are summed to one, i.e. $\eta + \zeta = 1$. Finally, ρ is the constant quantity. Disbelief is the proportion of observations that deviate significantly from the claimed performance θ_c^j, indicating non-compliance.

$$d_{D_j}^{\kappa_z} = \frac{\zeta \cdot h_{D_j}^{\kappa_z}}{\rho + \eta \cdot g_{D_j}^{\kappa_z} + \zeta \cdot h_{D_j}^{\kappa_z}} \tag{4}$$

where $h_{D_j}^{\kappa_z} = \sum_{i=1}^{\mathcal{N}} 1(\theta_{r,i} \not\approx \theta_c^j)$ is the number of negative outcomes of the device interaction with the system. Uncertainty represents the lack of information or observations about the device's performance.

$$u_{D_j}^{\kappa_z} = \frac{\rho}{\rho + \eta \cdot g_{D_j}^{\kappa_z} + \zeta \cdot h_{D_j}^{\kappa_z}} \tag{5}$$

Recent interaction events possess elevated importance owing to their recency, thereby exerting a more substantial influence compared to those from earlier periods. This behavior is encapsulated in the concept of temporal decay, mathematically formalized as:

$$\nabla(t_z) = \nabla_z = y^{Z-z}. \tag{6}$$

Here $y \in (0, 1)$ represents a decay coefficient that governs the rate of temporal degradation, with values closer to 1 indicating a slower decay. The variable Z denotes the current time slot and z refers to the most recent time period, establishing a framework where interaction significance diminishes exponentially as the time difference $Z - z$ increases. Consequently, the temporal decay captures the exponential attenuation of reputation scores over time, ensuring that more recent interactions are assigned a disproportionate weight in the computation of the overall reputation metric, while older interactions progressively lose influence.

$$b_{D_j}^{ter} = \frac{\sum_{z=1}^{Z} \nabla_z b_{D_j}^{\kappa_z}}{\sum_{z=1}^{Z} \nabla_z}, \tag{7a}$$

$$d_{D_j}^{ter} = \frac{\sum_{z=1}^{Z} \nabla_z d_{D_j}^{\kappa_z}}{\sum_{z=1}^{Z} \nabla_z}, \tag{7b}$$

$$u_{D_j}^{ter} = \frac{\sum_{z=1}^{Z} \nabla_z u_{D_j}^{\kappa_z}}{\sum_{z=1}^{Z} \nabla_z}. \tag{7c}$$

Combining (6) and (7) results in

$$b_{D_j}^{ter} = \frac{\sum_{z=1}^{Z} y^{Z-z} \cdot \frac{\eta \cdot g_{D_j}^{\kappa_z}}{\eta \cdot g_{D_j}^{\kappa_z} + \zeta \cdot h_{D_j}^{\kappa_z} + \rho}}{\sum_{z=1}^{Z} y^{Z-z}}, \tag{8a}$$

$$d_{D_j}^{ter} = \frac{\sum_{z=1}^{Z} y^{Z-z} \frac{\zeta \cdot h_{D_j}^{\kappa_z}}{\eta \cdot g_{D_j}^{\kappa_z} + \zeta \cdot h_{D_j}^{\kappa_z} + \rho}}{\sum_{z=1}^{Z} y^{Z-z}}, \tag{8b}$$

$$u_{D_j}^{ter} = \frac{\sum_{z=1}^{Z} y^{Z-z} \frac{\rho}{\eta \cdot g_{D_j}^{\kappa_z} + \zeta \cdot h_{D_j}^{\kappa_z} + \rho}}{\sum_{z=1}^{Z} y^{Z-z}}. \tag{8c}$$

The time-decayed average terminal reputation of the DUT is calculated as

$$R_{D_j}^{ter} = b_{D_j}^{ter} + \alpha_{D_j} \cdot u_{D_j}^{ter} \tag{9}$$

Putting equations together, we get (10). The algorithmic representation of the subjective logic-based reputation calculation, followed by the device compliance mechanism, is provided in Algorithm 1.

Algorithm 1 Subjective Logic-Based Reputation Algorithm for Device Compliance

Input: $D_j \in \mathcal{D}$: Devices under test.
$\mathcal{T}$: Compliance threshold
Output: Reputation score R_{D_j} and compliance decision (Trusted/Flagged)
Initialization:
    Set $R_{D_j} \leftarrow 0$
    Set $Z \leftarrow$ Number of previous interactions.
for each device $D_j \in \mathcal{D}$ **do**
    Extract subjective opinion components from equations (3), (4), and (5) : $\{b_{D_j}^{\kappa_z}, d_{D_j}^{\kappa_z}, u_{D_j}^{\kappa_z}, a_{D_j}^{\kappa_z}\}$.
    Apply time decay function from equation (8)
    Compute time averaged terminal reputation using (10).
end for
Compliance Decision:
if $R_{D_j} \geq \mathcal{T}$ **then**
    Mark D_j as Compliant device.
else
    Mark D_j as Flagged for Review
end if

IV. Contract Theory-Based Incentive Mechanism

Contract theory is crucial for designing incentives that align agent behavior with system goals, enhancing efficiency and mitigating risks [6]. The contract between the RIC and individual devices/services is directly linked to their reputation, i.e. it is a reputation-based contract. Devices with higher reputations are deemed to be more reliable and receive greater incentives from the central unit, enhancing overall ORAN system performance. The RIC acts as the principal, and each ORAN device D_j with an observable reputation score $R_{D_j} \in [0, 1]$ is the agent. Emphasizing compliance, the utility function can be formulated as:

$$U_{RIC^{\mathbb{I}_{\{\eta_j=1\}}}} = \sum_{j \in M} \left(\theta_{D_j}(R_{D_j}) - C_{D_j}(R_{D_j})\right) \cdot \mathbb{I}_{\{\eta_j=1\}} - e_{RIC} \tag{11}$$

where $\theta_{D_j}(R_{D_j})$ is the gain of the system by the device D_j based on their reputation R_{D_j}, $C_{D_j}(R_{D_j})$ is the compensation paid to the devices expressed in (13), and $\mathbb{I}_{\{\eta_j=1\}}$ is an indicator function, $\eta_j = 1$ if device D_j is compliant. The agent's utility depends on compensation and effort:

$$U_{D_j}(R_{D_j}) = C_{D_j}(R_{D_j}) - e_{D_j}(R_{D_j}) \tag{12}$$

where $e_{D_j}(R_{D_j})$ is the effort made by the device to maintain or improve its reputation, modeled in (14). The compensation $C_{D_j}(R_{D_j})$ to the device D_j is given by

$$C_{D_j}(R_{D_j}) = \alpha \cdot RB_{D_j} + \beta \cdot P_{D_j} + \gamma \cdot BW_{D_j} + \delta \cdot S_{D_j} \tag{13}$$

where $RB_{D_j}$is number of resource blocks allocated, P_{D_j} is transmission power granted to the device, BW_{D_j} is

$$R_{D_j}^{ter} = \frac{\sum_{z=1}^{Z} y^{Z-z} \cdot \frac{\eta \cdot g_{D_j}^{\kappa_z}}{\eta \cdot g_{D_j}^{\kappa_z} + \zeta \cdot h_{D_j}^{\kappa_z} + \rho}}{\sum_{z=1}^{Z} y^{Z-z}} + \alpha_{D_j} \cdot \frac{\sum_{z=1}^{Z} y^{Z-z} \frac{\rho}{\eta \cdot g_{D_j}^{\kappa_z} + \zeta \cdot h_{D_j}^{\kappa_z} + \rho}}{\sum_{z=1}^{Z} y^{Z-z}}. \tag{10}$$

bandwidth allocated to the device, S_{D_j} is scheduling priority, and $\alpha, \beta, \gamma, \delta$ are the task-oriented binary coefficients determining the significance of each resource for the device $\mathcal{D}_j$. The compensation $C_{D_j}(R_{D_j})$ for individual devices is structured to maximize the RIC's utility while satisfying the constraints of individual rationality (IR) and incentive compatibility (IC). IC ensures truthful reporting and aligns incentives, addressing the challenges of private information disclosure. To enforce incentive compatibility: $U_{D_j}(R_{D_j}) \geq U_{D_j}(\hat{R}_{D_j}), \quad \forall \hat{R}_{D_j} \neq R_{D_j}, \quad \forall D_j$ where $U_{D_j}(R_{D_j})$ is the utility of device D_j reporting its true reputation, and $U_{D_j}(\hat{R}_{D_j})$ is the utility when misreporting. IR guarantees that each device D_j achieves non-negative utility from the contract: $U_{D_j}(R_{D_j}) \geq 0$ This implies that compensation must exceed effort costs: $C_{D_j}(R_{D_j}) - e_{D_j}(R_{D_j}) \geq 0$. The reward provided by devices is their contribution to overall system performance, and can be in the form of spectral efficiency, reduced interference, and low latency by device.

In the ORAN system, the costs incurred by the devices to provide the reward/services to the RIC can be defined as the effort or resources that the device must expend to meet the system's performance expectations. In reputation-based systems, lower reputation typically correlates with higher costs due to the increased risks and reduced trust associated with these entities. This principle is applied to incentivize improved performance and participation. In our framework, devices with lower reputations incur higher operational costs, which serves to enhance overall system reliability. This mechanism encourages devices to improve their reputations, thereby reducing their costs and increasing their incentives to participate, ultimately leading to a more efficient and trustworthy system. Therefore, the total cost $e_{D_j}(R_{D_j})$ incurred by device D_j can be the combination of the costs of optimizing the power, bandwith, and interference mitigation, weighted by complement of their reputations, expressed as:

$$e_{D_j}(R_{D_j}) = \{\alpha_E \cdot E_{D_j} + \alpha_B \cdot BW_{D_j} + \alpha_I \cdot I_{D_j}\} \cdot (1 - R_{Dj}) \tag{14}$$

where α's are binary variables associated with distinct devices. The total cost to RIC include compensation to the individual devices, cost of network monitoring, reputation calculations, and resource allocations, which can be expressed as:

$$e_{RIC} = \sum_{j \in M} C_{D_j}(D_j) + C_{mon} + C_{rep} + C_{ra} \tag{15}$$

To maximize overall system performance while adhering to these constraints, the optimization problem can be formulated as the sum of all elements within the ORAN system as follows:

$$\max_{\mathbb{I}_{\eta_j=1}} \mathcal{P}_{\text{system}} = U_{RIC}\mathbb{I}_{\{\eta_j=1\}} + \sum_{j=1}^{M} U_{D_j}\mathbb{I}_{\{\eta_j=1\}}(R_{D_j}) \tag{16a}$$

$$U_{D_j}(R_{D_j}) \geq U_{D_j}(\hat{R}_{D_j}), \quad \forall \hat{R}_{D_j} \neq R_{D_j}, \quad \forall D_j \tag{16b}$$

$$C_{D_j}(R_{D_j}) \geq e_{D_j}(R_{D_j}). \tag{16c}$$

$$\mathbb{I}_{\{\eta_j=1\}} \in \{0, 1\}, \forall j \in M \tag{16d}$$

We propose a contract that evaluates the efficiency-to-cost ratio $\mathcal{E}_{system}$ for engaging devices, asserting that increased costs correspond to improved efficiency through the incentivization of reputable devices.

$$\mathcal{E}_{system} = \frac{\mathcal{P}_{system}}{e_{total}} \tag{17}$$

where $e_{total} = e_{RIC} + \sum_{j=1}^{M} e_{D_j(R_{Dj}) \cdot \mathbb{I}_{\eta_j=1}}$ is the total cost incurred to the principal and all the agents participating in the contract. This reputation-driven mechanism followed by compensation-based lucrative contracts allows the ORAN system to establish attractive contracts that reward high-reputation devices, resulting in superior system performance, albeit at a higher cost. A higher $\mathcal{E}_{system}$ indicates a more optimized system, ensuring that investments are oriented toward reputable devices to enhance the efficacy of ORAN.

V. Simulation Results

In this section, we discuss the simulation results to assess the efficacy of the proposed mechanism. We considered a total number of $M = 20$ devices (unless otherwise stated). For fair analysis, the parameters η, ζ, and ρ are all equal to 0.5. The reputation threshold is set to $\mathcal{T} = 0.7$.

In Fig. 2, ORAN system management without a moral hazard detection mechanism serves as a performance benchmark, against which we compare the proposed reputation-based incentive mechanism. Devices with poor reputations introduce moral hazards, impacting system performance if not excluded. The figure demonstrates that the proposed approach ensures high efficiency compared to assuming all elements are trustworthy. This performance advantage persists even with low acceptance rates or high incidences of maverick devices, supporting the importance of a zero-trust architecture in ORAN. For instance, in the third simulation run, the rate of reliable devices is less than 30%; nonetheless, the system achieves high performance, more than three times greater than the benchmark case when all devices are considered trusted. The efficiency of the proposed subjective logic based detection mechanism stems partly from its reduced false detection. Reputation scores are weighted to favor recent interactions. This benefits devices with recent

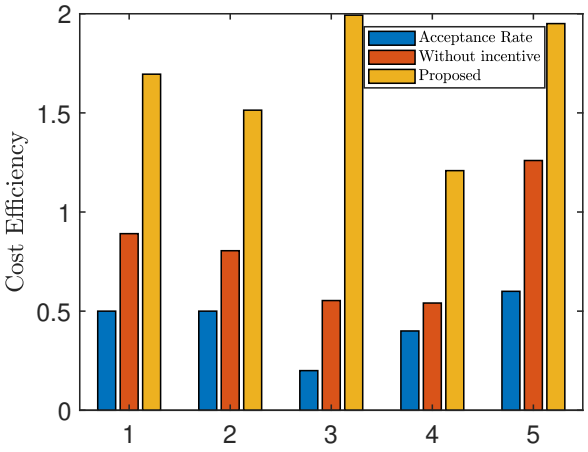

Fig. 2: Reputation-based incentive mechanism: cost efficiency over five simulation runs.

performance improvements, prevents unfair isolation, and increases their incentive to participate. Consequently, the system considers the recently proven reputable devices, resulting in enhanced system efficacy.

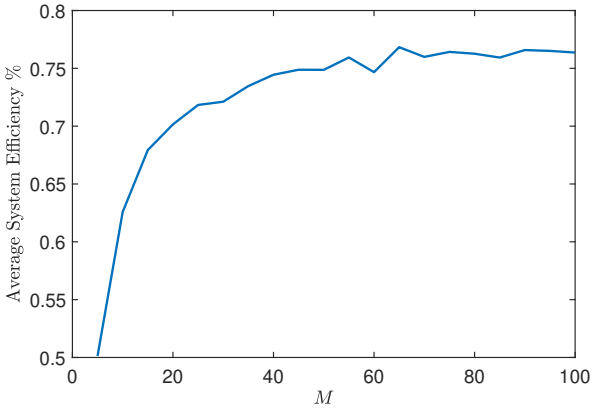

Fig. 3: Effect of increasing number of devices M.

Fig. 3 depicts the system efficiency as a function of the number of available devices (choices). The figure shows an increased efficiency but also a steady phase. As the number of devices increases, the management and testing cost (e.g., signaling overhead) also increases. However, the increase in the number of potential devices expands the set of choices, ultimately outweighing the associated costs. This positive trend continues until a saturation point, beyond which the additional devices contribute only marginally to further enhance the gain. At the saturation point when $M = 65$, the system efficiency stabilizes, meaning that adding more devices does not significantly enhance performance. This steady-state behavior, as depicted in Fig. 3, underscores the balance between the cost of adding more devices and the diminishing returns in system efficiency.

VI. Conclusion and Future Works

We developed the foundation of a moral hazard detection mechanism in ORAN, leveraging subjective logic to quantify device reputation and contract theory to incentivize compliance among highly reputed devices. Subjective logic is employed to dynamically evaluate each device's reputation, integrating both positive and negative interactions with the system. This approach accommodates the inherent uncertainty in ascertaining belief and disbelief regarding device compliance. Our simulation results substantiate the efficacy of this mechanism, demonstrating significant improvements in system performance when adopting our proposed approach.

While this approach assesses the reliability of the DUT, the confidence in compliance decisions is not quantified. In our future work, we will implement a hypothesis testing-based compliance evaluation mechanism for DUT in ORAN. This will include confidence intervals to quantify the certainty of compliance decisions.

References

[1] L. M. Larsen, H. L. Christiansen, S. Ruepp, and M. S. Berger, "The evolution of mobile network operations: A comprehensive analysis of open RAN adoption," *Computer Networks*, p. 110292, 2024.

[2] A. Arnaz, J. Lipman, M. Abolhasan, and M. Hiltunen, "Toward integrating intelligence and programmability in Open Radio Access Networks: A comprehensive survey," *IEEE Access*, vol. 10, pp. 67 747–67 770, 2022.

[3] J. Kirby, "Open RAN stakeholders should collaborate to ensure take-up in private networks," May 2024. [Online]. Available: https://www.analysysmason.com/research/content/articles/open-ran-stakeholders-rma18/

[4] J. T. Penttinen, M. Zarri, and D. Kim, *Open RAN Explained: The New Era of Radio Networks*. John Wiley & Sons, 2024.

[5] Spirent, *How to Test Open RAN*. Spirent eBook Series, 2023. [Online]. Available: https://www.spirent.com

[6] Y. Zhang, L. Song, M. Pan, Z. Dawy, and Z. Han, "Non-cash auction for spectrum trading in cognitive radio networks: Contract theoretical model with joint adverse selection and moral hazard," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 3, pp. 643–653, 2017.

[7] Y. Wu, S. Tang, P. Xu, and X.-Y. Li, "Dealing with selfishness and moral hazard in noncooperative wireless networks," *IEEE Trans. Mobile Comput.*, vol. 9, no. 3, pp. 420–434, 2009.

[8] V. Stafford, "Zero trust architecture," *NIST special publication*, vol. 800, p. 207, 2020.

[9] N. M. Yungaicela-Naula, V. Sharma, and S. Scott-Hayward, "Misconfiguration in O-RAN: Analysis of the impact of AI/ML," *Computer Networks*, p. 110455, 2024.

[10] H. Sedjelmaci, N. Kaaniche, and K. Tourki, "Secure and resilient 6G RAN networks: A decentralized approach with zero trust architecture," *Journal of Network and Systems Management*, vol. 32, no. 2, pp. 1–23, 2024.

[11] H. Liu, M. Ai, R. Huang, R. Qiu, and Y. Li, "Identity authentication for edge devices based on zero-trust architecture," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 23, p. e7198, 2022.

[12] N. Oren, T. J. Norman, and A. Preece, "Subjective logic and arguing with evidence," *Artificial Intelligence*, vol. 171, no. 10-15, pp. 838–854, 2007.

[13] Y. Liu, K. Li, Y. Jin, Y. Zhang, and W. Qu, "A novel reputation computation model based on subjective logic for mobile ad hoc networks," *Future Generation Computer Systems*, vol. 27, no. 5, pp. 547–554, 2011.

[14] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman, "Reputation systems," *Communications of the ACM*, vol. 43, no. 12, pp. 45–48, 2000.

[15] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 700–10 714, 2019.