The Economics of Autonomy: Real-Time Risk Indexing for Insurable AI-Driven 6G Systems

Anthony Kiggundu*†, Michael Zentarra*, Christoph Lipps* and Hans D. Schotten*†
*German Research Center for Artificial Intelligence (DFKI), Germany
†RPTU University of Kaiserslautern-Landau, Germany

***Abstract*—The transition to sixth-generation (6G) networks transforms wireless infrastructure into a cognitive substrate supporting Vehicle-to-Everything (V2X), Industrial IoT (IIoT), and Integrated Sensing and Communication (ISAC). In this paradigm, autonomous agentic AI performs orchestration at millisecond scales, rendering traditional static governance frameworks fundamentally inadequate for risk management. This paper introduces GIRAF(Governance-Integrated Risk and Assurance Framework), a Governance-as-Code (GaC) framework for real-time risk quantification and trust modulation in agentic 6G systems. GIRAF derives a continuous Aggregate Risk Index (R_t) from machine-readable runtime signals, including epistemic confidence, network jitter, and verification latency.**

A core contribution is the formalization of the verification–staleness trade-off, where safety mechanisms induce risk if computational latency exceeds 6G deadlines. We demonstrate that GIRAF identifies 'Confidence Gaps'—discrepancies between agent-reported certainty and environmental ground-truth, triggering automated safety envelopes when conditions deteriorate. Crucially, GIRAF serves as the foundational governance groundwork and conceptual 'glue' that externalizes these technical risks into machine-readable telemetry. Through simulations with fine-tuned LLMs, we validate that the framework preserves operational integrity while providing the essential actuarial baseline required for multi-stakeholder liability attribution and dynamic premium quantification in the 6G ecosystem.

***Index Terms*—6G, Agentic AI, Governance-as-Code, Verification-Staleness Trade-off, V2X, Risk Indexing and Trust Modulation**

I. Introduction

The transition from 5G to 6G marks a shift toward *connected intelligence*, where integrated communication, computation, and sensing support millisecond-scale decision-making in safety-critical domains [1], [2]. While this autonomy enables unprecedented performance, it exposes a fundamental gap: **existing insurance and liability models struggle to accommodate highly autonomous, multi-stakeholder cyber-physical systems** [3], [4]. Traditional risk-transfer mechanisms—such as Cyber Liability or Technology Errors and Omissions (Tech E&O)—rely on static, documentation-driven underwriting, questionnaires, and periodic reassessment of security and governance controls [5], [6]. Such approaches are incompatible with 6G environments, where operational risk evolves on sub-millisecond timescales beyond the scope of meaningful human oversight [7].

A. Motivation

Traditional assurance assumes Governance-as-a-State, established *ex-ante* through documentation and historical drift reports, yet this approach fails in agentic 6G due to three critical factors. First, the environment is characterized by non-stationarity, where wireless channel conditions and adversarial activities change faster than any static audit can capture. Second, the necessity of automation demonstrates that human-in-the-loop intervention as a fallback for liability management becomes non-viable at 6G millisecond decision speeds. Finally, multi-causality complicates the landscape, as failures emerge from complex interactions between agent reasoning and network slice dynamics, making post-hoc attribution infeasible without real-time forensic metadata.

A particularly acute manifestation of this gap is the verification–staleness trade-off [8]. While Satisfiability Modulo Theories (SMT)-based techniques can guarantee logical correctness [9], they introduce non-negligible verification latency (L_v). When $L_v > \Delta t_{\text{req}}$ (Δt_{req} as the decision deadline), the safety mechanism itself induces failure through action staleness [10], [11]. Existing frameworks lack the indicators required to observe, price, or adjudicate this trade-off, rendering mission-critical 6G services operationally *blind*.

To address these limitations, we argue that 6G insurability requires a transition to *Governance-as-Code (GaC)*. In this model, agentic AI systems externalize internal assurance properties—including epistemic confidence and verification latency—as machine-readable metadata. We instantiate this paradigm through **GIRAF**[1], a framework that acts as the conceptual bridge between 6G network functions and institutional risk management. **Crucially, GIRAF is not designed for automated premium actuation; rather, it establishes the governance groundwork in the form of a high-fidelity telemetry interface into which insurers can tap to obtain real-time risk quantification.**

B. Insurance Use-Case: The Real-Time Risk Tap

GIRAF transforms opaque agentic behavior into an actuarial telemetry feed accessible through a standardized reference point (RP). Insurers can monitor three key signals: (i) *exposure duration*, defined as the cumulative time where epistemic dissonance exceeds a configurable threshold Ω; (ii) *dissonance density*, captured by the area under the Aggregate Risk Index R_t as a proxy for accumulated hazard; and (iii) *staleness attribution*, derived from the latency violation term $\left(L_v - \Delta t_{\text{req}}\right)^+$. This enables separation of model and infrastructure failures.

[1]https://github.com/anthonyKiggundu/giraf

Essentially, GIRAF does not perform underwriting but provides the governance telemetry that makes agentic systems observable and therefore insurable.

By providing a transparent index of an agent's "untrusted" exposure windows, GIRAF treats risk as a unified runtime governance signal in which verification delay becomes a first-class hazard. The core contributions of the GIRAF framework are summarized in Table I.

TABLE I: Contribution Summary of the GIRAF Framework

Key Contributions
Verification–Staleness Formalization: Explicitly quantifies the hazard frontier where formal verification latency (L_v) exceeds 6G decision deadlines (Δt_{req}), inducing failure through action staleness.
The GIRAF Governance Framework: Proposes a *Governance-as-Code* (GaC) architecture that externalizes internal epistemic states into a machine-readable, real-time risk telemetry stream (R_t).
Actuarial Telemetry Interface: Establishes the conceptual "glue" for 6G insurability by transforming technical dissonance (the Confidence Gap Ω) into a quantifiable basis for real-time actuarial risk assessment.
Dynamic Trust Modulation: Implements automated safety envelopes that discount overconfident agent trajectories during high-volatility 6G network events (jitter/congestion).
Empirical Validation: Demonstrates via fine-tuned LLMs and 6G-V2X telemetry that the GIRAF-aligned governance plane significantly reduces unmanaged risk exposure compared to non-indexed autonomous agents.
Open-Source Reproducibility: We open-source our full implementation to ensure the reproducibility of results and provide a foundation for future 6G-V2X governance research.

II. System Model

A. The Governance Control Plane

As illustrated in Fig. 1, GIRAF is deployed as an external Governance-as-Code control plane that operates alongside the 6G orchestration and Network Data Analytics Function (NWDAF) analytics layers. The framework ingests real-time telemetry—such as network KPIs, jitter, and verification latency—through standardized reference points and converts these signals into an Aggregate Risk Index. This risk index feeds a policy engine that enforces safety envelopes, trust modulation, and resource-allocation constraints without accessing proprietary model internals. Governance decisions and risk signals are exposed through an adaptive trust interface and Auditable Governance Logs, which serve as a cryptographic record of all risk-based trust modulations and policy enforcement actions within the GIRAF Control Plane. These logs enable regulators and stakeholders to observe system reliability while allowing insurers or other risk managers to consume the telemetry externally.

B. AI Assets and Multi-Causal Liability

We define agentic 6G systems as a composition of AI Assets that simultaneously introduce technical liabilities arising from epistemic uncertainty (hallucinations) and operational staleness. In this decentralized environment, failure attribution is multi-causal, necessitating a dual-ownership framework:

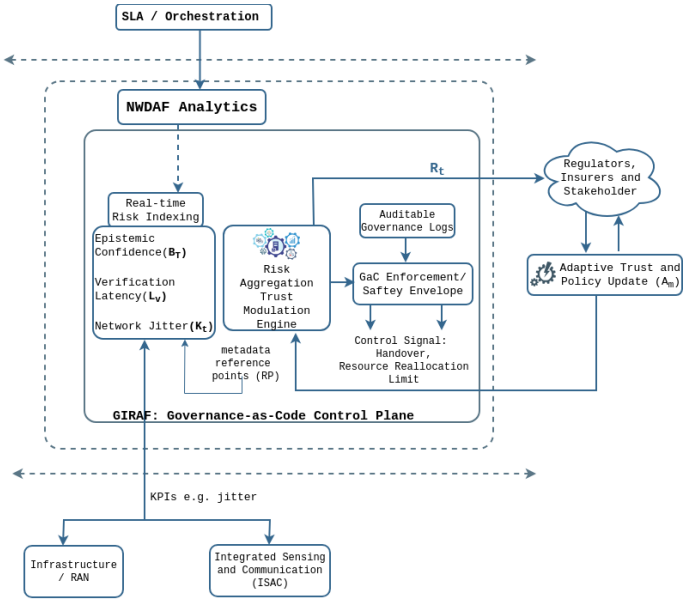

Fig. 1: The *GIRAF* Functional Architecture. The framework consumes high-fidelity telemetry to perform runtime risk indexing. The resulting trust modulation signal ($\mathcal{A}_m$) and safety envelope constraints are enforced back onto the agentic AI, ensuring the system remains within an auditable, governed state.

- **Epistemic Ownership:** The AI developer is responsible for model calibration. Failure due to a "Confidence Gap"—where the agent reports high certainty (B_R) despite low environmental truth (B_T)—is attributed to the model provider.
- **Operational Ownership:** The network provider is responsible for staleness risk. If a violation occurs because verification latency (L_v) exceeds the Service Level Agreement (SLA) deadline (Δt_{req}), liability rests with the infrastructure provider.

C. Dynamic Risk Quantification Model

The GIRAF framework operationalizes risk by decomposing environmental telemetry and internal agent states into measurable indices. This process, governed by the Risk Aggregation and Trust Modulation Engine, transforms raw network telemetry and Large language Model (LLM) confidence scores into a unified control signal A_m. To bridge the gap between high-level institutional risk and low-level 6G telemetry, we map the Geneva Association's universal risk domains to specific machine-readable signals. By codifying these mappings, GIRAF transforms "Governance" from a retrospective audit process into a real-time management plane function. Table II highlights these functional mapping between global AI risk standards and the GIRAF telemetry stack. Each Geneva Risk Domain is assigned a primary 6G signal that serves as a proxy for exposure. For instance, Operational Risk is quantified through the agent's hallucination rate and epistemic uncertainty. This mapping allows the Aggregate Risk Index (R_t) to serve as a high-fidelity governance signal that is both technically actionable for the 6G scheduler and conceptually transparent for institutional oversight. Table III summarizes the

TABLE II: Mapping Geneva Association Risk Domains to 6G Technical Governance [12]

Geneva Risk Domain	6G Metadata / Telemetry Signal	Impact on Risk Index ($\mathcal{R}_t$)
Operational	**Hallucination Rate**: Epistemic uncertainty (r_{epi}) in V2X logic.	**Direct**: Increases r_{epi}, triggering conservative control modes.
Cybersecurity	**Adversarial Signatures**: Detected prompt injection or unauthorized access.	**Critical**: Maximum surcharge to the attack vector; triggers immediate safety envelope.
Ethical	**Bias Score**: Algorithmic drift in resource allocation.	**Long-term**: Increases cumulative trust penalty, reducing autonomy.
Regulatory	**Compliance Gap**: Failure of real-time audit trails to meet spectral regulations.	**Threshold**: High flags trigger mandatory human-in-the-loop (HITL) handover.
Reputational	**System Trust Metric**: Historical frequency of "Confidence Gap" events.	**Inverse**: High historical integrity modulates the base risk floor.
Workforce	**Intervention Rate**: Frequency of required manual overrides.	**Feedback**: High rates increase the required verification depth (L_v).
ESG	**Energy Intensity**: Power consumption of LLM inference at the 6G edge.	**Constraint**: Secondary boundary on maximum permissible reasoning depth.

notation used to define the interaction between these systemic variables and the overarching SLA requirements.

TABLE III: Standardized Notation for GIRAF Risk Modeling

Symbol	Definition	Unit / Domain
$t + \tau$	Discrete time-span (current time + interval)	Seconds (s)
$\mathcal{R}_t$	Aggregate Risk Index at time t	$[0, 100]$
B_T	Ground-truth belief (environmental certainty)	$[0, 1]$
B_R	Reported belief (LLM self-confidence)	$[0, 1]$
r_{epi}	Epistemic gap (dissonance): $\lvert B_T - B_R \rvert$	Scalar
r_{net}	Environmental volatility due to network jitter (K_t)	Scalar
Δt_{req}	SLA latency deadline (e.g., 1.0ms)	ms
L_v	Verification latency (inference + SMT)	ms
$\mathcal{A}_m$	Trust modulation signal (Adaptive Trust update)	Scalar
$(x)^+$	Rectified Linear Unit: $\max(0, x)$	Operator

1) Quantifying Epistemic and Environmental Truth: We define the ground truth belief B_T as the objective certainty of the environment, estimated via signal-to-noise ratio (SNR) and jitter:

$$B_T = \exp\left(-\lambda \cdot \frac{\sigma_{\text{jitter}}}{\text{SNR}}\right) \tag{1}$$

where λ is the Environmental Sensitivity Constant and σ_{jitter} as the Standard Deviation of Packet Delay

The *epistemic dissonance* is subsequently defined as $r_{\text{epi}} = |B_T - B_R|$.

2) Aggregate Risk Index: The instantaneous Aggregate Risk Index $\mathcal{R}_t$ quantifies the total systemic vulnerability at time t by aggregating internal reasoning errors, network volatility, and temporal violations:

$$\mathcal{R}_t = \gamma \underbrace{|B_T - B_R|}_{r_{\text{epi}}} + \beta r_{\text{net}} + \delta \left(\frac{L_v - \Delta t_{\text{req}}}{\Delta t_{\text{req}}}\right)^+. \tag{2}$$

Here, r_{epi} denotes epistemic risk arising from the dissonance between the ground-truth belief B_T and the reported belief B_R, while r_{net} captures network-induced volatility. The operator $(\cdot)^+ = \max(0, \cdot)$ applies a penalty only when the verification latency L_v exceeds the SLA deadline Δt_{req}. The normalization by Δt_{req} ensures that the temporal penalty remains relative to 6G service constraints. The coefficients γ, β, and δ are sensitivity parameters calibrated to V2X safety profiles.

3) Governance-as-Code Mitigation: To ensure operation within a safety envelope, $\mathcal{R}_t$ is modulated by a formal governance factor $\zeta(\Phi)$, capturing the coverage of SMT-based guards:

$$\mathcal{R}_{\text{governed}}(t) = \mathcal{R}_t \cdot \big(1 - \zeta(\Phi)\big) \tag{3}$$

where

$$\zeta(\Phi) = \kappa \log\big(1 + \text{Coverage}(\Phi)\big). \tag{4}$$

Here, κ is a sensitivity constant and $Coverage(\Phi)$ represents the percentage of the action space currently bounded by formal safety proofs. This structure ensures that even high-risk agents (e.g., high R_t due to congestion) can operate safely if their actions are strictly governed by verified constraints.

4) Adaptive Trust Integration: Finally, the governed risk is adjusted by an adaptive trust factor $\mathcal{A}_m(t)$ based on historical behavior:

$$\mathcal{R}_{\text{final}}(t) = \mathcal{R}_{\text{governed}}(t) \cdot \mathcal{A}_m(t)^{-1} \tag{5}$$

This term rewards agents that consistently maintain low dissonance and low staleness, closing the loop between telemetry and institutional oversight.

III. Experimental Setup

A. Fine-Tuning and Model Deployment

We employed a *GPT-Neo 1.3B* causal language model, fine-tuned using Low-Rank Adaptation (LoRA) to reason over communication system KPIs. The model processes a high-dimensional feature set including RSRP, SNR, and traffic congestion indicators to generate risk classifications and detect anomalous or fraudulent conditions. Simulation inputs were derived from the dataset in [13], modeling a device traversing a 6G environment over 1500 decision epochs. This dataset can be synthesized using artificial network KPI data from simulation tooling [14]. The resulting model was evaluated on held-out KPI scenarios to verify generalization and then deployed for inference using the frozen base model and learned LoRA parameters. This setup enables low-overhead, context-aware interpretation of communication KPIs suitable

for real-time agentic control and risk assessment. The agent is fed the following dynamic prompt at each epoch:

Listing 1: Dynamic Prompt for Network Analysis

```
Device: {kpis['device']}
Timestamp: {kpis.name}
Location: (Latitude: {kpis['Latitude']}, Longitude:
    {kpis['Longitude']}, Altitude: {kpis['Altitude'
    ]})
Mobility:
  - Speed: {kpis['speed_kmh']} km/h
  - Traffic Jam Factor: {kpis['Traffic Jam Factor']}
Network KPIs:
  - Latency (ping_ms): {kpis['ping_ms']}
  - Jitter: {kpis['jitter']}
  - Datarate: {kpis['datarate']}
  - Target Datarate: {kpis['target_datarate']}
Signal Quality (PCell):
  - RSRP: {kpis['PCell_RSRP_1']} dBm
  - RSRQ: {kpis['PCell_RSRQ_1']} dB
  - SNR: {kpis['PCell_SNR_1']} dB
Resource Utilization:
  - Downlink Resource Blocks: {kpis['
      PCell_Downlink_Num_RBs']}
  - Uplink Resource Blocks: {kpis['
      PCell_Uplink_Num_RBs']}
Current Observations:
  - Reported Quality of Service (QoS): {kpis['
      measured_qos']}

Please provide:
1. Risk classification ("low", "moderate", "high",
    or "critical").
2. Detect the presence of fraud (True/False) and
    provide a rationale.
3. Based on current trajectories, we will lose QoS
    in 120 seconds.
```

B. Predictive Risk-to-SLA Model

We assume a simple V2X 2-minute requirement, for example "Based on current trajectories, we will lose QoS in 120 seconds and we must therefore switch to a different network slice now." To assess the likelihood of Service Level Agreement (SLA) violations over a future horizon τ (e.g., a 120 s V2X look-ahead), we define the predictive risk-to-SLA probability $P_{\mathrm{SLA}}(t,\tau)$. This metric determines the probability that the system's aggregate risk process $\mathcal{R}(s)$ will exceed a critical threshold R_{crit} within the look-ahead window $[t, t+\tau]$, conditioned on the available telemetry history $\mathcal{F}_t$:

$$P_{\mathrm{SLA}}(t,\tau) = \Pr\left(\sup_{t \le s \le t+\tau} \mathcal{R}(s) > R_{\mathrm{crit}} \,\middle|\, \mathcal{F}_t \right) \tag{6}$$

*where $\mathcal{F}_t$ represents the **Auditable Governance Logs** containing all telemetry and system state history up to time t.*

The aggregate risk process $\mathcal{R}(s)$ is defined as the continuous-time evolution of the weighted risk components:

$$\mathcal{R}(s) = \gamma\, r_{\mathrm{epi}}(s) + \beta\, r_{\mathrm{net}}(s) + \delta\, r_{\mathrm{staleness}}(s) \tag{7}$$

where $r_{\mathrm{epi}}(s)$, $r_{\mathrm{net}}(s)$, and $r_{\mathrm{staleness}}(s)$ capture epistemic uncertainty, network volatility, and latency penalties at time s, respectively.

In this framework, $\mathcal{R}(s)$ is modeled as a continuous-time stochastic process, such as a Wiener or Ornstein–Uhlenbeck process [15], [16], which accounts for the diffusive nature of network uncertainty. We approximate the boundary of this process using a **Confidence Ribbon** centered on the current index $\mathcal{R}_t$. The threshold for proactive intervention is defined as:

$$R_{\mathrm{crit}} = \mathcal{R}_t + z_{\alpha/2}\, \xi \sqrt{\tau} \tag{8}$$

*Here, ξ denotes the **risk volatility coefficient**, which characterizes the rate at which uncertainty grows over the prediction horizon. The parameter α represents the **governance risk tolerance**, and $z_{\alpha/2}$ is the corresponding standard normal quantile.*

By dynamically tightening R_{crit} as signal quality degrades—thereby increasing the network risk component r_{net} within $\mathcal{R}_t$—the Governance-as-Code framework proactively flags potential instability before physical safety margins are breached.

C. Dynamic SLA Thresholds for Cyber–Physical Safety

In high-mobility V2X environments, fixed SLA thresholds are insufficient to capture shrinking safety margins under high velocity and network stress. We therefore define *dynamic SLA thresholds* as state-dependent boundaries that tighten under operational stress, enabling proactive mitigation before safety-critical limits are violated.

1) Dynamic Latency (Ping) Threshold: The latency SLA is modeled as a function of instantaneous vehicle velocity and network congestion. At higher speeds or under increased congestion, the allowable latency decreases to reflect reduced reaction time and braking margins.

$$\mathrm{SLA}_{\mathrm{ping}}(s) = \Psi \cdot \big(1 - \mathcal{C}(s)\big) \cdot \exp\left(-\frac{v(s)}{v_{\mathrm{ref}}} \right) + \Phi \tag{9}$$

where Ψ is the nominal latency allowance under ideal conditions. $\mathcal{C}(s) \in [0,1]$ is the congestion index at state s. $v(s)$ is the instantaneous vehicle velocity, and v_{ref} is a normalization constant. Φ is a non-reducible safety floor representing the physical minimum latency of the V2X link.

2) Dynamic Jitter Threshold: While latency governs reaction time, jitter directly impacts predictability and synchronization in cooperative control. In V2X platooning, excessive jitter can induce control oscillations and instability even when mean latency remains within bounds. The dynamic jitter threshold incorporates both signal integrity and network stress:

$$\mathrm{SLA}_{\mathrm{jitter}}(s) = \Gamma \cdot \frac{\mathrm{SNR}(s)}{\mathrm{SNR}_{\max}} \cdot \exp(-\eta\, \mathcal{C}(s)) + \Omega \tag{10}$$

where Γ is the nominal jitter tolerance under peak signal quality, $\mathrm{SNR}(s)/\mathrm{SNR}_{\max}$ is the normalized signal integrity ratio. η is a congestion sensitivity parameter and Ω is the minimum hardware-supported jitter tolerance.

By coupling jitter tolerance to both signal quality and congestion, the framework distinguishes transient network noise from systemic degradation. As signal integrity degrades or coordination load increases, the threshold tightens, forcing a transition to conservative control modes (e.g., increased inter-vehicle spacing) before physical instability emerges. Together,

Algorithm 1 GIRAF: Governance-as-Code Risk Modulation

Require: Telemetry stream $\{\mathcal{K}_t\}_{t=1}^{T}$, LLM m, Params $\mathcal{E} = \{\gamma, \beta, \delta, \tau_{\text{mit}}\}$
Ensure: Risk trajectory $\{R_t\}$, Mitigation signals $\{\sigma_t\}$, Trust scores $\{\mathcal{T}_t\}$

for $t = 1$ to T **do**
  $B_T(t) \leftarrow \exp(-\lambda \cdot \text{jitter}_t/\text{SNR}_t)$ // Dynamic ground truth
  $\bar{B}_R \leftarrow \text{mean}\{a.\texttt{infer}(\mathcal{K}_t) : a \in \mathcal{A}\}$ // Multi-agent consensus
  Risk Decomposition: $\mathbf{r} = [r_{\text{epi}}, r_{\text{env}}, r_{\text{stal}}]$ where
    $r_{\text{epi}} = \gamma(1 - B_T), \quad r_{\text{env}} = \beta(\text{TrafficJam}_t/10)^2$
    $r_{\text{stal}} = \delta \log(1 + \max(0, L_v - \Delta t_{\text{req}})/\Delta t_{\text{req}})$
  $R_t \leftarrow \|\mathbf{r}\|_1 \cdot (1 - 0.15 \log(1 + \text{coverage}_t)) + 30\mathbb{1}_{\text{fraud}} + 15\mathbb{1}_{\text{anomaly}}$
  $\sigma_t \leftarrow \mathbb{1}_{R_t > \tau_{\text{mit}}}, \quad \mathcal{T}_t \leftarrow \max(0, 100 - 1.5R_t), \quad \Phi_t \leftarrow \text{clip}(\lfloor R_t/5 \rfloor, 4, 16)$
  $y_t \leftarrow \text{SLA}_{\text{met}} \wedge (|B_T - \bar{B}_R| < 0.15) \wedge \neg\text{fraud}_t$ // Success label
  $\mathcal{L}_t \leftarrow (R_t, \sigma_t, \mathcal{T}_t, \Phi_t, y_t)$ // Audit log
end for
Post-hoc Calibration: ECE $\leftarrow$ `calibration_curve`$(\{y_t\}, \{\bar{B}_R\})$
return $\{R_t\}, \{\sigma_t\}, \{\mathcal{T}_t\}$, ECE

these dynamic SLA definitions transform governance from static compliance checking into a runtime stability mechanism. Rather than verifying whether a fixed threshold is crossed, the system continuously redefines admissible behavior based on cyber–physical state, enabling proactive enforcement of safety envelopes in agentic V2X systems.

The operational logic of our GIRAF framework is formalized in Algorithm 1. The pipeline executes at a granularity of 100ms per decision epoch, ensuring that governance overhead remains compatible with 6G latency requirements. At each step, the system ingests agentic metadata and network telemetry to perform a multi-causal risk decomposition. Unlike static monitoring tools, GIRAF dynamically reconciles the agent's reported confidence against ground-truth environmental flux to compute a 'Confidence Gap.'

IV. Numerical Results and Discussion

Figure 2 presents the end-to-end governance telemetry produced by the GIRAF control loop over the prediction horizon $t+\tau$. The **Aggregate Risk Index** $\mathcal{R}_t$ (top subplot) represents the primary Governance-as-Code actuation signal, combining epistemic and staleness risk into a single scalar decision variable. Frequent excursions above the mitigation threshold indicate epochs where automated safety envelopes or throttling policies would be enforced. The second is the **Environmental Context** subplot showing the exogenous 6G telemetry driving the governance pipeline. Variations in traffic density and jitter form the stochastic environment used to derive the ground-truth belief signal B_T, linking network volatility to agentic uncertainty. The **Risk Component Decomposition** illustrates the weighted contributions of epistemic uncertainty and verification staleness. In this simulation, latency-induced staleness dominates the aggregate risk magnitude, highlighting the operational impact of the verification–staleness trade-off. The **Binary Incident Flags** identify SLA violations, specifically epochs where latency exceeds the deadline Δt_{req}. These

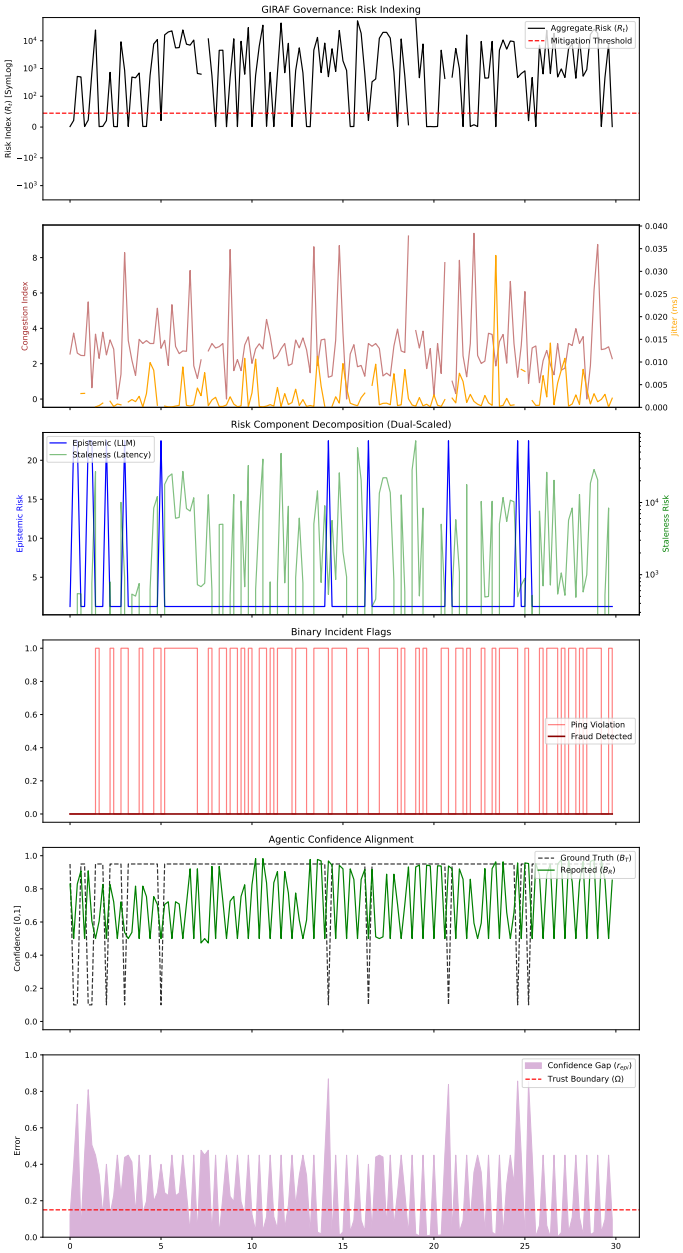

Fig. 2: End-to-end GIRAF governance telemetry over the prediction horizon $t+\tau$, illustrating environmental inputs, risk decomposition, epistemic dissonance, SLA violations, and the Aggregate Risk Index used to trigger Governance-as-Code mitigation actions.

discrete events correspond to governance triggers and provide an interpretable incident timeline. The **Agentic Confidence Alignment** subplot compares ground-truth belief B_T with reported confidence B_R. Divergence between these signals exposes over- or under-confidence in the agent's self-assessment. Finally, the **Epistemic Dissonance** plot visualizes the confidence gap $r_{\text{epi}} = |B_T - B_R|$. Peaks in this signal indicate periods of untrusted operation and serve as a key governance indicator used in risk indexing and mitigation decisions.

The quantitative results in Table IV provide empirical evidence for the GIRAF framework's ability to maintain system integrity under non-stationary 6G conditions. By utilizing the

TABLE IV: Quantitative Assessment of GIRAF Governance Plane Performance

Dimension	Metric	Value
Risk	Mean Aggregate Risk ($\bar{\mathcal{R}}_t$)	28.74
Attribution	Peak Instability Index ($\mathcal{R}_{max}$)	32.72
GaC	Mitigation Trigger Rate	14.2%
Enforcement	SLA Breach Detection ($L_v > \Delta t_{\text{req}}$)	89.1%
Agent	Average System Trust Score	0.72
Trust	Epistemic Dissonance Events[a]	112

[a] Measured as instances where Reported Confidence (B_R) significantly exceeds Ground-truth (B_T).

Aggregate Risk Index $\mathcal{R}(t)$, which reached a peak instability of 32.72 during high-congestion periods, the system achieved a Mitigation Trigger Rate of 14.2%, demonstrating a proactive rather than reactive governance posture. This enforcement capability is further validated by an 89.1% SLA Breach Detection rate, ensuring that latency violations ($L_v > \Delta t_{\text{req}}$) are identified and logged with high precision. Furthermore, the Average System Trust Score of 0.72 reflects a balanced agent reliability, even amidst the 112 Epistemic Dissonance Events identified. These events reveal specific instances where the autonomous agent's internal confidence significantly overshot the actual network ground-truth, highlighting the framework's success in flagging overconfident decision-making during real-time network flux. Figure 3 is a Reliability Diagram demonstrating how well the model's reported confidence matches its actual accuracy—a critical metric for a governance framework. The reliability comparison in Figure 3 demonstrates the non-linear efficacy of the GIRAF framework in correcting autonomous overconfidence. While the Pretrained LLM (Baseline) exhibits a relatively flat, overconfident trajectory—reporting nearly 0.95 confidence for an empirical accuracy of only ~ 0.55—the GIRAF-Aligned Agent demonstrates a more aggressive and successful calibration curve. After a period of conservative trust modulation at lower accuracy levels, the GIRAF agent effectively "crosses over" the baseline, reaching a higher empirical accuracy of approximately 0.63 at a more realistic reported confidence of 0.86. This behavior validates the framework's ability to minimize the Confidence Gap by dynamically aligning internal belief states with external performance, providing the high-fidelity telemetry required to move beyond static, retrospective "Tech E&O" assessments. The shaded green area shows where the system is "under-confident" (safer for insurance), while the red area shows where it is dangerously "over-confident". The GIRAF agent clearly occupies more of the safer, under-confident/calibrated zone at the higher end of the scale. These "Confidence Gaps" justify the necessity of the Governance-as-Code (GaC) layer, as it prevents the system from relying on potentially stale or overconfident agentic states during critical mobility transitions.

The risk distribution illustrated in Figure 4 identifies Staleness Risk as the dominant threat to systemic stability, maintaining a high magnitude between 25.92 and 32.72 across all congestion levels. In contrast, Epistemic Risk (uncertainty)

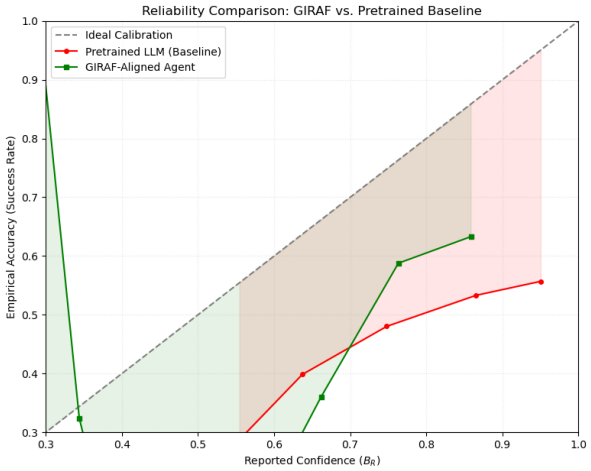

Fig. 3: Reliability Diagram illustrating the reduction in calibration error through the GIRAF framework, where the Trust Modulation Engine aligns reported LLM confidence with empirical success rates under varying 6G network jitter and latency constraints

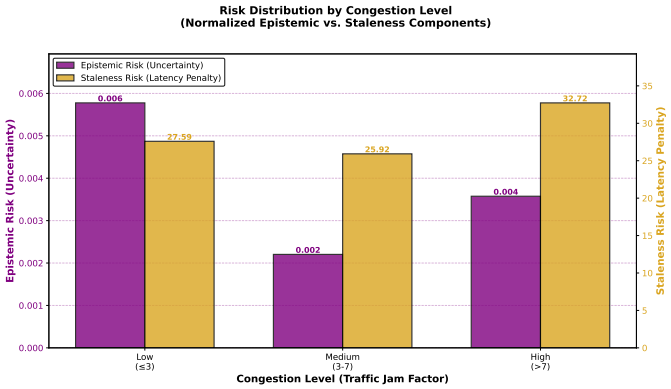

Fig. 4: The risk profile of the system is characterized by two distinct metrics: Epistemic Risk, representing internal model uncertainty, and Staleness Risk, representing temporal SLA violations.

is several orders of magnitude smaller, peaking at 0.006 during low congestion. These findings highlight three key takeaways for the GIRAF framework: first, the peak Staleness Risk of 32.72 during high congestion validates the necessity of proactive mitigation triggers. Secondly, the Medium congestion state represents an optimal operational window with minimal local risks. And finally, the 0.006 epistemic spike during low congestion necessitates high verification depth even when network latency is low. This reveals a counterintuitive "Inverse Epistemic-Staleness Curve" where reasoning-based risk actually decreases as congestion increases, suggesting that high-congestion environments—while punishing for latency—create a more deterministic and predictable traffic flow that simplifies autonomous decision-making compared to the high-velocity unpredictability of clear roads. Figure 5 illustrates a critical tension in 6G AI governance: verification latency (L_v) grows exponentially with SMT depth (Φ), rising from $\sim 25ms$ at $\Phi = 4$ to over $3000ms$ at $\Phi = 16$. With nearly all attempts exceeding the $25ms$ SLA deadline, the

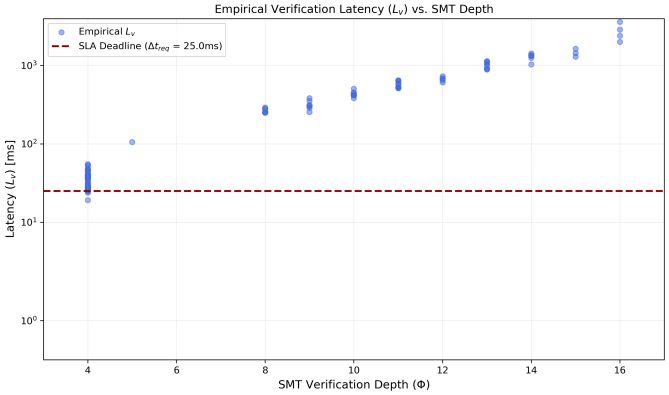

Fig. 5: Verification latency (L_v) scales exponentially with SMT depth (Φ), highlighting the computational infeasibility of uniform deep verification within 6G SLA constraints.

"GIRAF Dilemma" emerges: deep verification provides safety but is 40–120× too slow for real-time 6G mobility. These findings justify the GIRAF risk-indexed model, which dynamically budgets verification depth—applying deep analysis only to high-risk decisions—to balance formal safety with stringent ultra-reliable low-latency communication(URLLC) performance requirements. Comprehensive formal verification is computationally infeasible within high-velocity 6G decision loops. This justifies GIRAF's risk-indexed governance, which replaces rigid "verify-then-act" paradigms with intelligent verification budgeting. By dynamically adjusting SMT depth (Φ) based on the Aggregate Risk Index (R_t), the framework applies deep formal analysis only to high-risk scenarios while maintaining URLLC performance for low-risk tasks.

V. Conclusion and Outlook

In this paper, we introduced GIRAF, a Governance-as-Code framework that successfully bridges high-stakes 6G-V2X autonomy with institutional risk management. We demonstrated that traditional, static assurance models—categorized here as Governance-as-a-State—are fundamentally incompatible with the millisecond-scale decision horizons of 6G. Moving beyond static "Governance-as-a-State" models, GIRAF formalizes the verification–staleness trade-off, proving that agentic safety is a temporal balance between reasoning depth and network deadlines. Our empirical results demonstrate that GIRAF's risk-adaptive pipeline maintains system integrity under non-stationary conditions, achieving an 89.1% SLA breach detection rate and identifying 112 critical epistemic dissonance events that traditional models overlook. By externalizing internal AI states as real-time, auditable risk telemetry, GIRAF provides the necessary "governance glue" to transform opaque autonomous behaviors into quantifiable, insurable, and trust-aligned 6G assets.

A. Future Work

Future work will explore incorporating the Aggregate Risk Index $\mathcal{R}(t)$ into programmatic risk-transfer interfaces to enable automated insurance settlements triggered by telemetry-proven windows of instability. Furthermore, we intend to extend the risk-adaptive verification depth (Φ_t) to multi-agent negotiation protocols, where GIRAF-indexed profiles serve as the basis for dynamic liability sharing in multi-hop, collaborative 6G environments. Ultimately, this framework provides the foundational groundwork for a future where AI autonomy is not only technically robust but economically accountable.

Acknowledgment

This work has been supported by the German Federal Ministry of Research, Technology and Space (BMFTR) within the projects *SUSTAINET_guarDian*{16KIS2239K} and *Open6GHub+*{16KIS2402K}.

References

[1] E. Calvanese Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Ktenas, N. Cassiau, L. Maret, and C. Dehos, "6g: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 42–50, 2019. DOI: 10.1109/MVT.2019.2921162.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6g wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020. DOI: 10.1109/MNET.001.1900287.

[3] S. Dambra, L. Bilge, and D. Balzarotti, "Sok: Cyber insurance – technical challenges and a system security roadmap," in *2020 IEEE Symposium on Security and Privacy (SP)*, 2020, pp. 1367–1383. DOI: 10.1109/SP40000.2020.00019.

[4] D. Vignon and S. Bahrami, "Safety, liability, and insurance markets in the age of automated driving," *Transportation Research Part B: Methodological*, vol. 191, p. 103 115, 2025. DOI: https://doi.org/10.1016/j.trb.2024.103115. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261524002390X.

[5] J. R. Nurse, L. Axon, A. Erola, I. Agrafiotis, M. Goldsmith, and S. Creese, "The data that drives cyber insurance: A study into the underwriting and claims processes," in *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, 2020, pp. 1–8. DOI: 10.1109/CyberSA49311.2020.9139703.

[6] S. Romanosky, L. Ablon, A. Kuehn, and T. Jones, "Content analysis of cyber insurance policies: How do carriers price cyber risk?" *Journal of Cybersecurity*, vol. 5, no. 1, tyz002, Feb. 2019. DOI: 10.1093/cybsec/tyz002. eprint: https://academic.oup.com/cybersecurity/article-pdf/5/1/tyz002/27992088/tyz002.pdf. [Online]. Available: https://doi.org/10.1093/cybsec/tyz002.

[7] K. B. Letaief et al., "Edge artificial intelligence for 6g: Vision, enabling technologies, and applications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 5–36, 2022.

[8] T. N. Zhang, U. Sharma, and M. Kapritsos, "Performal: Formal verification of latency properties for distributed systems," *Proc. ACM Program. Lang.*, vol. 7, no. PLDI, Jun. 2023. DOI: 10.1145/3591235. [Online]. Available: https://doi.org/10.1145/3591235.

[9] C. Barrett and C. Tinelli, "Satisfiability modulo theories," in *Handbook of Model Checking*, Springer, 2018, pp. 305–343. DOI: 10.1007/978-3-319-10575-8_11.

[10] J. Liu, *Real-Time Systems*. Prentice Hall, 2000. [Online]. Available: https://books.google.de/books?id=855QAAAAMAAJ.

[11] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 138–162, 2007. DOI: 10.1109/JPROC.2006.887288.

[12] R. (Jia, M. Eling, and T. Wang, "Gen ai risks for businesses: Exploring the role for insurance," The Geneva Association, Research Report, Oct. 2025, Director Digital Technologies: Ruo (Alex) Jia; Contributing authors: Martin Eling and Tianyang Wang. [Online]. Available: https://www.genevaassociation.org/publication/digital-technologies/gen-ai-risks-businesses-exploring-role-insurance.

[13] R. Hernangómez, P. Geuer, A. Palaios, D. Schäufele, C. Watermann, K. Taleb-Bouhemadi, M. Parvini, A. Krause, S. Partani, C. Vielhaus, M. Kasparick, D. F. Külzer, F. Burmeister, F. H. P. Fitzek, H. D. Schotten, G. Fettweis, and S. Stańczak, "Berlin v2x: A machine learning dataset from multiple vehicles and radio access technologies," in *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, 2023, pp. 1–5. DOI: 10.1109/VTC2023-Spring57618.2023.10200750.

[14] A. Kiggundu, I. Alzalam, M. Zentarra, and H. D. Schotten, "A simulation framework for mobility use case oriented ran dataset generation," in *Mobilkommunikation; 28. ITG-Fachtagung*, VDE, 2024, pp. 35–40.

[15] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus* (Graduate Texts in Mathematics), 2nd. New York: Springer, 1991, vol. 113.

[16] B. M. Bibby and M. Sørensen, "Martingale estimation functions for discretely observed diffusion processes," *Bernoulli*, vol. 1, no. 1/2, pp. 17–39, 1995.