International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Building Trust in Agentic AI: TRACE Framework for Policy-Driven Multi-Agent System Design

Dr. Nabanita Sinha
Deloitte, India

ABSTRACT: The rapid adoption of multi-agent AI systems— ranging from prescriptive, workflow-driven deployments to fully agentic, autonomous ecosystems—raises urgent challenges for trust, accountability, and regulatory compliance. This paper introduces the TRACE Framework (Trust, Review, Accountability, Critique, Explainability), a governance-first architecture designed to make multi-agent AI systems auditable, policy-aligned, and operationally reliable across varying degrees of agent autonomy. TRACE embeds governance anchors at the agent level, enforces data privacy and policy checks, supplies a dedicated Critic agent for meta-validation, and preserves human-in- the-loop oversight where required. We present a layered architecture that separates Governance & Compliance, Operational Agents, and Oversight & Assurance, and provide a concrete methodology for instrumenting agent behaviour with provenance, explainability outputs, and per-agent metrics. A formal scoring rubric—comprising agent operational metrics, critic checks, and aggregation rules—yields an Overall System Confidence (OSC) that drives automated actions, human escalation, and continuous learning. Finally, we propose a suite of operational KPIs for each layer as Governance and Compliance Indicators (GCI), Agentic Performance Metrics (APM), and Assurance Indicators (AI) that enable financial institutions and other regulated organisations to deploy multi-agent systems that are efficient, auditable, and compliant. TRACE bridges the gap between regulatory expectations and system engineering practice— providing a practical roadmap for trustworthy multi-agent AI deployment in high-stakes domains.

KEYWORDS: Multi-Agent Systems, Agentic AI, AI Governance, Explainable AI, TRACE Framework, Trusted AI

I. INTRODUCTION

Artificial Intelligence (AI) adoption accelerates across data-driven industries including high-stakes sectors such as finance, health care. New challenges are emerging around autonomy control, systemic bias, explainability, and accountability in increasingly complex multi-component AI ecosystems. AI governance has moved decisively from aspiration to obligation. The European Union Artificial Intelligence Act (EU AI Act) [1] entered into force on 1 August 2024, establishing a risk-based regulatory framework with staged implementation — including prohibitions effective from early 2025 and full obligations for general-purpose AI systems commencing from 2 August 2025. In parallel, organizations are adopting internationally recognized management standards such as ISO/IEC 42001:2023 [2] and the NIST AI Risk Management Framework (AI RMF 1.0) [3] to institutionalize governance and operationalize AI risk controls. At the same time, the enterprise use of AI continues to accelerate. Global surveys [4–6] report that a significant majority of organizations now deploy AI in at least one core business function, ranging from predictive analytics and automation to decision support and generative applications. This rapid expansion intensifies the need for system-level trust, accountability, and security mechanisms that can extend beyond individual models or algorithms to encompass multi-agent ecosystems — especially as Agentic AI and multi-agent coordination frameworks become foundational in enterprise automation. While existing AI governance frameworks offer high- level guidance for responsible AI deployment, most remain model-centric and organization-focused, with limited operationalization at the system architecture level. The emergence of multi-agent AI systems (MAS) — characterized by distributed decision-making, autonomous reasoning, and dynamic tool integration — introduces new governance challenges. These include traceability across agent workflows, cross-agent accountability, bias propagation through inter-agent communication, and the need for explainable reasoning chains that remain auditable under regulatory scrutiny.

This paper addresses these challenges by introducing the TRACE Framework — an architecture of trust and accountability for both Agentic and Prescriptive MultiAgent Systems. TRACE embeds governance principles directly into agent workflows, ensuring that each decision, tool invocation, and output is traceable, explainable, and policy compliant. The framework is structured around three

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

interdependent layers — Governance & Compliance, Operational Agents, and Oversight & Assurance — that collectively establish the foundation for trustworthy multiagent AI operations.

Each agent in the TRACE ecosystem is instrumented with provenance tracking, explainability logs, and defined performance metrics, enabling transparent and auditable behavior. A formal scoring methodology aggregates these metrics with critic evaluations and policy conformance checks to compute an Overall System Confidence (OSC) score. This score dynamically governs automation thresholds, human-in-the-loop escalation, and continuous improvement cycles — maintaining a balance between autonomy, control, and compliance.

By aligning with global standards such as the EU AI Act, NIST AI RMF, and ISO/IEC 42001, TRACE translates regulatory principles into measurable system design elements and governance indicators. TRACE structures agent collaboration, monitors reasoning integrity, enforces governance checkpoints, and integrates human oversight — collectively enabling auditable, policy-aligned, and explainable multi-agent AI systems.

The rest of the paper is presented in following manner. Section II reviews related work on multi-agent AI systems and governance frameworks. Section III outlines the conceptual foundations of the TRACE Framework. In Section IV presents the framework architecture and core components. and Section V concludes with implications for industry adoption and future research directions.

II. BACKGROUND & RELATED WORK

A. Governance Frameworks:

As AI moves into critical decision-making roles across industries, regulatory and management frameworks have become central to operational governance. One of the most widely discussed enterprise governance models is AI TRiSM (Trust, Risk, and Security Management), introduced by Gartner [6]. AI TRiSM focuses on building and maintaining trustworthy AI systems by embedding governance, fairness, robustness, and lifecycle management controls across the AI pipeline. It underscores the importance of model monitoring, data lineage, access control, and runtime inspection to ensure ongoing compliance and risk mitigation.

Complementing this, the NIST AI Risk Management Framework (AI RMF 1.0), released by the U.S. National Institute of Standards and Technology (NIST) in 2023 [3], provides a structured, voluntary framework for designing, developing, deploying, and managing trustworthy AI systems. The NIST AI RMF defines four core functions— Govern, Map, Measure, and Manage— that help organizations identify, assess, and mitigate AI risks across technical, societal, and ethical dimensions. It aligns with ISO/IEC 42001:2023 [2], the world's first management system standard for AI governance, which specifies organizational processes and documentation requirements for responsible AI deployment.

Together, AI TRiSM, ISO/IEC 42001, and the NIST AI RMF represent a maturing governance landscape where trustworthiness, traceability, and risk accountability are no longer aspirational goals but regulatory imperatives. However, these frameworks remain largely model-centric, lacking mechanisms to extend governance to multi-agent systems (MAS)—where distributed agents interact autonomously and dynamically.

B. Evolution of Multi-Agent and Agentic AI:

An AI agent is typically defined as a computational entity that perceives its environment and acts to achieve designated goals [7]. Early software agents were deterministic, narrow, and rule based [8]. In contrast, agentic systems emerging since 2023 leverage large language models (LLMs), external tool use, and persistent memory, enabling sophisticated reasoning, collaboration, and adaptive planning. Traditional agents performed single-step tasks such as information retrieval, summarization, or scripted dialogue generation [9]. These systems lacked the multi-step reasoning, autonomy, and adaptability necessary for complex real-world applications. Recent LLM-driven architectures, however, allow specialized agents—such as planners, analysts, and coders—to dynamically decompose tasks, share contextual memory, and coordinate actions over extended time horizons [10]. This evolution transforms the nature of AI systems from isolated decision units to interacting agent collectives capable of emergent, decentralized behavior. Yet, with increased autonomy come magnified governance challenges—particularly concerning transparency, coordination, and accountability across agent networks [11].

C. Governance Challenges in Multi-Agent AI:

The rise of multi-agent ecosystems introduces new governance complexities:

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

- Traceability and Provenance: Each agent's autonomy complicates end-to-end traceability of reasoning chains, data usage, and decision flow [11].
- Bias Propagation: Bias in one agent's reasoning can propagate through inter-agent communication, compounding fairness risks [12].
- Transparency and Explainability: Explaining reasoning across multi-step, multi-agent workflows remain technically demanding [13].
- Accountability in Autonomous Coordination: Dynamic tool-calling and goal-chaining can produce unpredictable interactions that defy static policy control.
- Regulatory Alignment: Current frameworks such as NIST AI RMF or ISO 42001 primarily address centralized AI systems, not distributed agentic architectures [3].

Recent research has begun to address these issues. The MAESTRO Framework proposed by the Cloud Security Alliance [15] introduces a threat-modeling methodology for agentic AI. Similarly, the AAGATE Platform [16] aligns agentic system governance with NIST RMF principles, emphasizing provenance, policy controls, and risk propagation analysis. Nonetheless, both approaches remain early-stage and lack integrated metrics, scoring mechanisms, and human-in-the-loop oversight for continuous assurance.

III. CONCEPTUAL FOUNDATIONS OF THE TRACE FRAMEWORK

A. Motivation and Theoretical Context:

The development of the TRACE Framework is grounded in the convergence of two distinct research and regulatory trajectories: (1) the institutionalization of AI governance through standards such as ISO/IEC 42001:2023 [2], the NIST AI Risk Management Framework (AI RMF 1.0) [3], and Gartner's AI TRiSM [6]; and
(2) the rapid emergence of multi-agent systems (MAS) driven by large language models (LLMs) and autonomous reasoning architectures [11], [17], [18].

While these governance frameworks define macro-level principles for trustworthy AI—fairness, transparency, privacy, reliability—they do not prescribe mechanisms to enforce such principles at the agent-to-agent interaction level. Conversely, the design of Agentic AI systems prioritises autonomy, adaptability, and performance, often without built-in accountability or explainability [10], [12]. TRACE was conceived to bridge this gap by embedding governance controls within the operational fabric of multiagent ecosystems. Beyond governance, TRACE also serves as a structured methodology for system development, prescribing measurable metrics and evaluation checkpoints that guide the design, validation, and monitoring of each agent and inter-agent process throughout the AI lifecycle.

Conceptually, TRACE aligns governance functions with the risk-based approach of the NIST AI RMF [3]— specifically its four phases: Govern, Map, Measure, and Manage—while mapping trust and security attributes from AI TRiSM [19] into the agent lifecycle. This combination produces a governance-anchored system architecture that operationalizes regulatory principles through continuous monitoring, critique, and human oversight.

B. Core Principles and Dimensions

TRACE is founded on five interdependent pillars—Trust, Review, Accountability, Critique, and Explainability— each corresponding to a governance function and measurable performance dimension.

1. ***Trust:*** Trust represents both the objective quality of system performance and the subjective assurance of reliability perceived by users and regulators. In the TRACE model, trust emerges from consistent evidence of system integrity: accurate outputs, privacy compliance, bias mitigation, and transparent reasoning [11], [14]. Trust is quantified through per-agent metrics— accuracy, tool reliability, bias detection rate—and aggregated into the Overall System Confidence (OSC) score that governs automation thresholds.
2. ***Review:*** The review dimension operationalizes humanin-the-loop (HITL) oversight, ensuring that automated decisions remain aligned with legal and ethical norms [16]. TRACE formalizes review checkpoints where human auditors validate critic outputs and approve or reject system recommendations. This structure mirrors the Govern and Manage functions of NIST AI RMF [3], which emphasize stakeholder oversight and accountability documentation.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

3. ***Accountability*:** Accountability in TRACE is achieved through explicit provenance tracking and agent-level auditability. Each agent maintains cryptographically verifiable logs of inputs, reasoning steps, tool invocations, and outputs. These logs align with ISO/IEC 42001 requirements for audit traceability [2] and support responsibility attribution when errors occur. This principle ensures that governance obligations are not abstract policy goals but measurable, enforceable system properties.
4. ***Critique:*** The Critic Agent—a supervisory metaagent—embodies the TRACE critique function. It performs cross- validation of agent reasoning, consistency checking, and policy compliance analysis. By applying rule-based and LLM- driven evaluations, the critic quantifies confidence for each agent's output, detects hallucinations, identifies missing evidence, and flags noncompliant actions [17], [18]. This automated "second-layer reasoning" aligns with AI TRiSM's emphasis on runtime inspection and resilience management [19].
5. ***Explainability*:** Explainability ensures that both agent and human reviewers can interpret decisions in a structured, verifiable manner [11]. Each agent in TRACE generates a reasoning summary that links conclusions to evidence and policies. The system assesses Explainability a quantitative measure of how well generated explanations match underlying reasoning paths. This supports regulatory interpretability mandates under the EU AI Act [1] and NIST's Measure function [20].

Fig 1. Pillars of TRACE

C. Governance Mapping and Policy Alignment

TRACE translates abstract governance requirements into operational mechanisms as shown in Table I. This mapping formalizes TRACE as a policy-aligned operationalization framework—translating governance intent into measurable, auditable behavior at the system level and report generation.

Agentic and Prescriptive Multi-Agent Systems:

The TRACE Framework is intentionally designed to operate across both Agentic and Prescriptive Multi-Agent Systems, ensuring consistent governance, accountability, and trust irrespective of the system's autonomy level. While Agentic Multi- Agent Systems (AMAS) exhibit adaptive, self-organizing behaviors through autonomous reasoning and decision-making cycles, Prescriptive MultiAgent Systems (PMAS) rely on structured, predefined workflows governed by external control logic. TRACE provides a

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

unifying architectural layer that embeds monitoring, critique, and explainability across both paradigms.

In Agentic Systems, agents independently plan, reason, and act, often invoking tools or sub-agents based on emergent goals [10]. TRACE introduces governance anchors at key decision points to ensure transparency, ethical alignment, and traceable accountability without constraining the system's autonomy. Through built-in audit trails and the Critic Agent, TRACE enables evaluation of agent reasoning, decision paths, and outcome justification. Each autonomous decision is logged with associated rationale, tool calls, and confidence metrics, supporting post verification and regulatory audit readiness.

In Prescriptive Systems, where agent workflows and tool calls are predefined, TRACE acts as a compliance and assurance layer, enforcing consistency, privacy, and performance monitoring. Every agent action is validated against rule-based governance constraints, ensuring policy alignment and minimizing risk propagation across the agent chain. The framework's critic and review components verify that outputs remain logically consistent, contextually relevant, and ethically sound, while the accountability mechanism ensures traceability from task initiation to completion.

Conceptually, TRACE extends established AI governance frameworks into the systemic domain of multi-agent AI, integrating risk management, oversight, and transparency within the operational logic of autonomous agents [10, 21]. It treats each agent—autonomous or prescriptive—as a governable entity within a traceable ecosystem. This enables unified monitoring of agent behaviour, tool-call reliability, reasoning transparency, and human oversight, promoting trust and resilience across diverse AI deployments.

Table 1. Policy Allignment

Governance Source	Principle / Requirement	TRACE Implementation Mechanism
NIST AI RMF 1.0 [3]	Govern Establish risk policies and oversight Map: Contextualize AI use cases Measure: Assess trustworthiness Manage: Mitigate risk and monitor outcomes	Governance & Compliance Layer with configurable risk thresholds Per-agent metadata and contextual task descriptors Agent-level metrics, Critic Score and OSC aggregation Dynamic escalation to human review and retraining triggers
AI TRiSM [19]	Runtime risk and security monitoring Lifecycle trust and fairness management	Critic Agent validation and anomaly detection Bias detection, redaction coverage, and fairness reporting
ISO/IEC 42001 [2]	Governance documentation and audit trails	Provenance logging, policy anchoring, and traceable evidence chains
EU AI Act [1]	Transparency and human oversight mandates	HITL checkpoints, explainability reports, and confidence-based escalation

IV. TRACE FRAMEWORK ARCHITECTURE

The TRACE Framework operationalizes its conceptual foundations through a three-layer architecture that integrates governance, execution, and assurance in a unified control system. Figure 2 illustrates the TRACE Framework architecture, comprising three primary layers. These layers—Governance & Compliance, Operational Agents, and Oversight & Assurance—ensure that every action, reasoning step, and outcome produced by an agent is measurable, traceable, and explainable in alignment with ISO/IEC 42001 [2], NIST AI RMF [20].

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

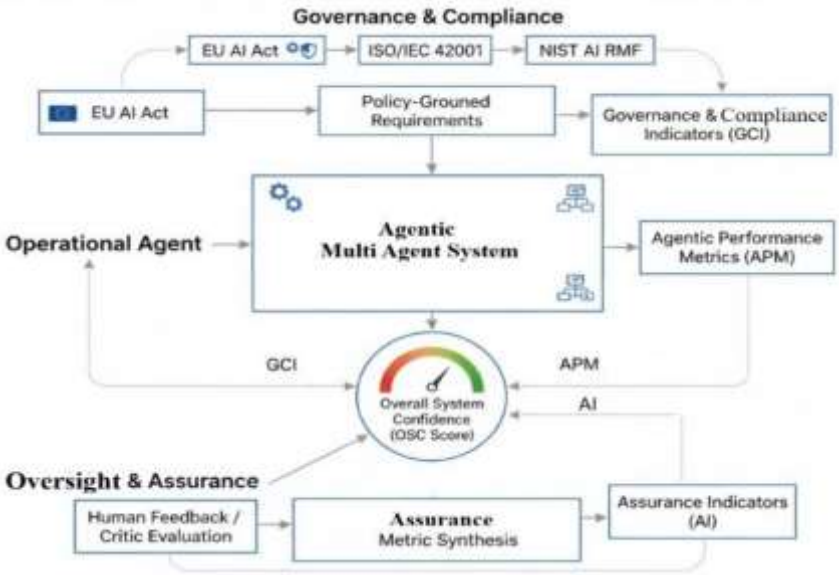

Fig 2. TRACE Framework

A. Governance & Compliance Layer:

This layer defines the rules, safeguards, and constraints that govern the TRACE ecosystem. It establishes what must be done and controlled to ensure legal, ethical, and operational integrity. This layer does not execute domain tasks; instead, it regulates how data, prompts, and agents behave, functioning as the policy brain and shield of TRACE. Its focus is on defining what is allowed, what is prohibited, and what must be logged and applying defensive mechanisms before agents act ensuring trust-by-design before any reasoning or decision-making begins. This layer is prescriptive and preventative — it sets boundaries, prepares inputs, and configures the risk and control environment.

- ***Input Data Sanitization:*** Before any agent interaction, incoming data and user inputs pass through a multi-stage validation and control pipeline that enforces system hygiene and resilience. Structural and semantic checks validate the integrity of input data. Unstructured content is normalized, missing fields are imputed or flagged, and anomalous records are isolated for human review, consistent with ISO/IEC 42001 secure-data-handling guidelines [2].
- ***Prompt-Injection Detection****:* Natural-language inputs are automatically scanned for adversarial or promptinjection patterns [23] capable of manipulating agent behaviour or circumventing policies. A hybrid defense model—combining rule-based pattern filters with LLMdriven classifiers—flags and quarantines malicious fragments before they propagate through the network [17].
- ***Bias and Sensitivity Screening****:* Inputs in agentic system are evaluated for demographic or contextual bias using pre- trained bias detection models [22]. Detected bias signals trigger mitigation measures—either blocking the biased input, requesting user clarification, or routing the record for supervised review before further processing. This preventive gate ensures that unfair or discriminatory data do not enter the agentic workflow, maintaining equity across decision paths for every agent in the multi-agent system.
- ***Privacy Protection****:* Each input is automatically scanned for personally identifiable information (PII) [24] using entity-recognition and pattern-matching algorithms. Detected identifiers are either masked, tokenized, or redacted based on sensitivity level and access privilege.
- ***Controlled Data Exposure****:* Each agent in the multiagent system processes only the data strictly required for its assigned task thar limits the potential damage from data breaches or unauthorized sharing between agents[20].
- *Access and Privilege Controls:* Role-based privilege maps restrict which agents, tools, or APIs can access specific data types or invoke certain actions, preventing unauthorized tool calls or inter-agent leakage [19].
- ***Risk Context Mapping:*** Each agentic task is dynamically assigned a risk profile—low, medium, or high—based on autonomy level, domain criticality, and data sensitivity. These risk scores determine monitoring frequency, control intensity, and the degree of required human oversight.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

- ***Cross-Agent Fairness Checks:*** Evaluates outcomes across agents and user cohorts to detect potential bias or disparate treatment. Identified disparities are recorded in a Fairness Impact Report, supporting transparency and equitable decision-making [18].
- ***Audit Logging and Provenance Tracking:*** Every event within TRACE—from input arrival to final output—is captured in a Provenance Ledger, which constitutes the foundation of system traceability. Entries satisfy ISO/IEC 42001 audit requirements [2] and facilitating post-hoc accountability and external audit compliance.

Collectively, these mechanisms ensure trust-by-design before any agent reasoning & task begins, blocking unsafe prompts, sanitizing untrusted content, and establishing an auditable baseline for downstream operations.

Governance and Compliance Merics: To measure the effectiveness of the governance controls embedded within TRACE, this layer defines a structured set of Governance and Compliance Indicators (GCI), listed in Table 2. These indicators quantify the degree to which policies are enforced, inputs are sanitized, bias is mitigated, and audit integrity is maintained. The resulting values provide a continuous assurance signal to the Oversight Layer, enabling proactive risk management, real-time compliance tracking, and adaptive governance refinement.

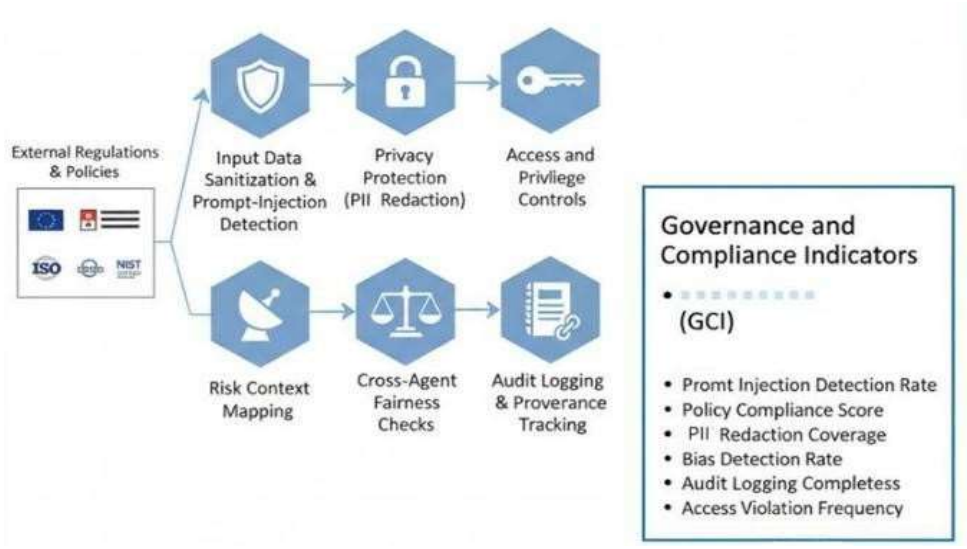

Fig 3. Governance and Compliance Layer

Table 2. Governance and Compliance Merics

GCI	Description / Purpose	Governance Objective
Prompt Injection Detection Rate	Percentage of adversarial inputs correctly intercepted and quarantined.	Input integrity and system resilience.
Input Validation Accuracy	Accuracy of syntactic and semantic input checks.	Data quality and reliability.
Bias Detection Rate	Proportion of biased inputs identified and blocked before processing.	Fairness and bias control.
PII Redaction Coverage	Extent of personal-data removal prior to agent access.	Privacy and regulatory compliance.
Policy Compliance Score	Conformance of agent operations to defined policies and risk thresholds.	Policy enforcement assurance.
Access Violation Frequency	Frequency of unauthorized data or tool access attempts detected and contained within the system.	Privacy and regulatory compliance
Audit Logging Completeness	Percentage of events, inputs, and decisions successfully captured in the immutable provenance ledger.	Traceability, accountability, and audit readiness

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

B. Operational Agent Layer

The Operational Agent Layer constitutes the functional core of the TRACE framework.It comprises a coordinated set of specialized agents responsible for executing analytical, decision-making, and tool-invocation tasks within pre-defined governance boundaries. While the Governance & Compliance Layer (Layer 1) establishes what is allowed and how inputs must be controlled, this layer performs how work is executed—applying reasoning, collaboration, and contextual intelligence under active monitoring.Every activity in this layer is recorded, explainable, and auditable, ensuring that autonomy never operates outside traceable oversight.The output is designed to be clinician-friendly, supporting decision- making and early intervention.

Agent Roles and Functional Responsibilities*:*

TRACE supports a modular and scalable multi-agent design [21], allowing each agent to operate independently while adhering to shared governance protocols.The primary agent types and their responsibilities include:

Extractor Agent: Acquires and preprocesses structured and unstructured data, applying validation and metadata tagging defined by the Governance & Compliance Layer. It ensures that every dataset entering the system is verified, cleansed, and privacy compliant.

- ***Reasoning Agent:*** Performs contextual reasoning, inference, and hypothesis generation. It leverages both internal knowledge and external context sources, maintaining a structured reasoning log for explainability and post-hoc validation.
- ***Tool-Adapter Agent:*** Interfaces with APIs, analytical engines, or visualization components. Each external call (e.g., computational model, knowledge graph, or charting function) is logged, version-tracked, and validated to prevent unapproved tool use or data leakage [10].
- ***Knowledge Agent:*** Retrieves domain-specific information or historical cases using vectorized retrieval and contextual matching. It supports continuity and reasoning depth, ensuring decisions are informed by prior validated knowledge [19].

Together, these agents perform distributed reasoning within a controlled communication framework, where inter-agent dialogue follows schema-based messaging validated by the Governance Layer to ensure security and policy alignment.

Agent Collaboration and Context Management*:* Agent collaboration within the Operational Agent Layer is structured, coordinated, and transparent rather than freeform. Interactions occur through a Controlled Workflow Bus (CWB) — a managed communication channel that governs message exchange, context propagation, and synchronization across agents.Each inter-agent message carries standardized metadata — including task identifiers, reasoning type, confidence score, timestamp, and datalineage hash — enabling full traceability of decision flow.

The Operational Orchestrator Agent (OOA), a coordination module within this layer, sequences agent actions, manages dependencies, and ensures balanced resource utilization and fair tool access across concurrent workflows. TRACE supports both Agentic and Prescriptive modes of multi-agent collaboration.

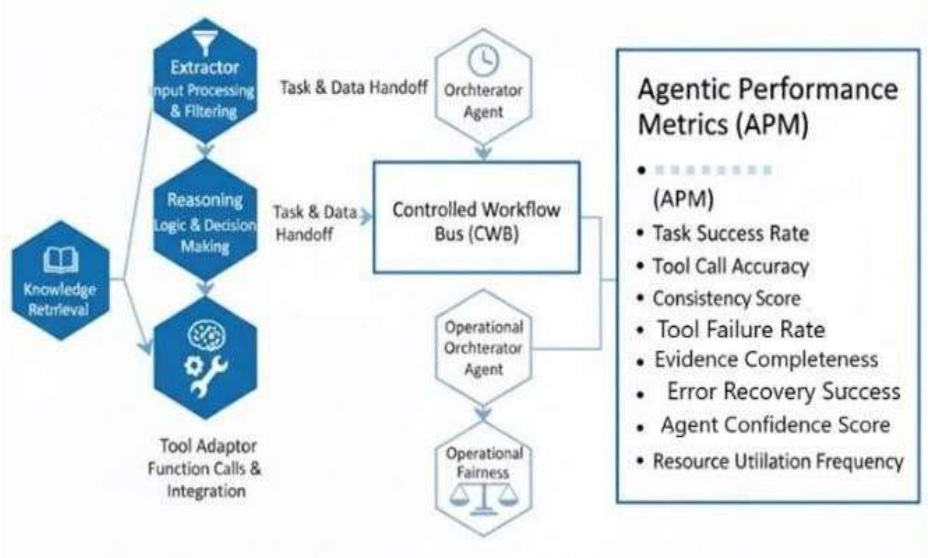

Fig 4. Operational Agent Layer

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

• ***AMAS - Dynamic, Context-Aware Collaboration:*** In AMAS, collaboration emerges dynamically as autonomous agents plan, reason, and adapt to evolving task contexts. The CWB serves primarily as a context synchronization and monitoring layer, allowing agents to exchange reasoning states, confidence values, and evidence chains in real time without compromising traceability. Agents determine what to share and when based on their internal goals, confidence thresholds, and reasoning logic.
The OOA observes but does not prescribe sequence— intervening only to resolve conflicts, prevent circular reasoning, or rebalance computational load. This mode of collaboration supports creativity, adaptability, and emergent problem- solving while preserving deterministic accountability through metadata logging and reasoning trace capture.

• ***PMAS- Structured, Rule-Based Coordination:*** In Prescriptive MAS, collaboration is governed by predefined workflows and fixed task dependencies. The CWB functions as a workflow controller, routing outputs and context strictly according to predefined schemas. The OOA enforces deterministic sequencing (e.g., Agent A → Agent B → Agent C), ensures compliance with timing and data integrity constraints, and prevents unauthorized task deviation. Inter- agent communication follows standardized formats, and validation checks to guarantee that each agent performs its assigned role consistently within the governed process.
This approach provides predictability, repeatability, and compliance assurance—making it ideal for regulatory or safety- critical environments where transparency and process consistency outweigh autonomy.

Agent Performance Measure:
Each agent in the Operational Agent Layer continuously emits telemetry and operational statistics referred in Table 3, as Agentic Performance Metrics (APM). These metrics provide a quantitative foundation for both internal optimization and external assurance, enabling developers, auditors, and governance controllers to monitor performance, interpretability, and reliability in real time.

Table 3. Agentic Performance Metrics

APM	Description	Purpose / Developer Insight
Task Success Rate	Ratio of successfully completed agent tasks to total attempted tasks.	Measures task completion accuracy.
Tool Call Efficiency	Indicator combining the success rate of tool invocations,	Measures resource efficiency and tool use optimization.
Tool Failure Rate	Percentage of tool or API calls that return errors, timeouts, or invalid responses.	Identifies integration and reliability bottlenecks.
Tool Call Accuracy	Measures how accurately the agent selects the correct tool for the assigned reasoning goal, validated through critic or rule mapping.	Ensures reasoning to-action consistency and task–tool alignment.
Error Recovery Success Rate	Share of failed or partial tasks successfully retried or auto corrected by the agent.	Indicates resilience and fault- tolerance efficiency.
Data Accuracy	Proportion of correct inferences or calculations validated against ground truth or deterministic benchmarks.	Ensures correctness of analytical reasoning.
Evidence Completeness	gree to which reasoning outputs reference relevant data or supporting evidence.	Validates reasoning transparency and internal traceability.
Consistency Score	Statistical variance in agent outputs for semantically equivalent inputs.	Evaluates reasoning reproducibility and logical stability.
Agent Confidence Score	Self-assessed confidence level emitted by the agent based on reasoning uncertainty, tool success rate, and internal evaluation.	Indicates perceived reliability and reasoning certainty

C. Oversight & Assurance Layer:
This layer provides supervisory intelligence, independent verification, and human-aligned control over multi-agent operations. While the Governance & Compliance Layer establishes policy boundaries, and the Operational Agent Layer executes tasks within those boundaries, the Oversight Layer performs evaluation, critique, escalation, and continuous improvement. This layer ensures

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

that agent autonomy does not compromise trust, safety, or regulatory alignment, regardless of whether the system operates in AMAS or PMAS mode. It integrates three core functions: Automated Critic Evaluation, Human-in-the-Loop (HITL) and System-level Assurance Scoring and Escalation.

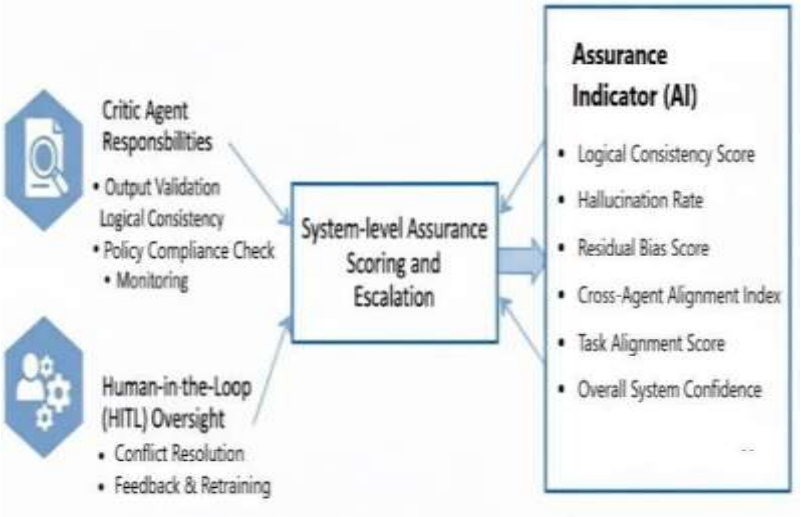

Fig 5. Oversight & Assurance Layer

- ***Critic Agent Responsibilities:*** The Critic Agent serves as the analytical core of the Oversight & Assurance Layer, ensuring that agentic reasoning and outputs remain logically coherent, factually grounded, and unbiased across the multi- agent ecosystem [24]. It operates as an independent validator, examining agent outputs, reasoning traces, inter-agent interactions and provides measurable metrics as Assurance Indicator (AI), listed in Table 4.
- ***Human-in-the-Loop:*** **Human** oversight remains a core assurance pillar in TRACE. The system escalates outputs based on risk, confidence, and regulatory context. Human reviewers provide qualitative feedback on reasoning quality, coherence, domain correctness, and policy adherence, corrective labels for critic fine-tuning, override power for final decision execution. This implements safe autonomy, aligned with ISO 42001 human-oversight requirements [20].
- ***System-level Assurance Scoring and Escalation:*** TRACE quantifies end-to-end trust through a unified Overall System Confidence (OSC) score that integrates assurance signals from all three architectural layers. Each layer contributes a normalized composite score derived from its respective metric set—Governance & Compliance Indicators (GCI), Agentic Performance Metrics (APM), and Assurance Indicators (AI)— representing policy adherence, operational reliability, and reasoning integrity, respectively. The OSC is computed as a weighted aggregation of the three layer-level composites:

$$OSC = wGGCIavg + wAAPMavg + wCAIavg \qquad (1)$$

Where , wg, wA, wCdenote non-negative layer total weights equal to 1. The resulting OSC provides a single quantitative indicator of system assurance, used to automate escalation: high scores authorize autonomous execution, mid-range scores trigger human validation, and low scores mandate audit or remediation

Table 4. Assurance Indicator

AI	Measurement Focus / Purpose
Logical Consistency Score	Evaluates the internal coherence and logical soundness of agent reasoning and outputs.
Hallucination Rate	Measures the proportion of unsupported or fabricated statements identified in agent outputs.
Residual Bias Score	Quantifies remaining demographic or contextual bias after agent- level mitigation.
Cross-Agent Alignment Index	Assesses semantic and contextual alignment among agents collaborating on shared tasks.
Task Alignment Score	Evaluates how closely Agentic layer output aligned with predefined task objective and organizational policy

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

V. CONCLUSION

This paper introduced the TRACE Framework, a unified governance and assurance model that integrates Governance & Compliance, Operational Agent, and Oversight & Assurance layers to create measurable, policyaligned multi-agent ecosystems. TRACE operationalizes international AI governance principles—drawing from the EU AI Act, ISO/IEC 42001, and NIST AI RMF—through quantifiable metrics: Governance and Compliance Indicators (GCI), Agentic Performance Metrics (APM), and Assurance Indicators (AI). By combining these into a single Overall System Confidence (OSC) score, the framework transforms trust from an abstract ideal into a verifiable, system-level property—linking input integrity, operational reliability, and critic evaluation within one continuous assurance loop. While TRACE offers a structured basis for trustworthy multi-agent AI, it assumes uniform data normalization and metric calibration across layers, which may vary under diverse architectures and regulatory settings. Reliance on qualitative in critic evaluation also introduces subjectivity. Future work will pursue automated calibration, expanded metrics for model drift and adversarial risks, benchmarking against sectoral governance standards, and large-scale validation across high-risk domains.

REFERENCES

1. European Commission. (2025). AI act. https://digitalstrategy.ec.europa.eu/en/policies/regulatory-framework-ai
2. International Organization for Standardization. (2023). ISO/IEC 42001:2023 – Artificial intelligence management system (AIMS) – Requirements (Tech. Rep. ISO/IEC 42001:2023). https://www.iso.org/standard/81230.html
3. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
4. Singla, A., Sukharevsky, A., Yee, L., Chui, M., & Hall, B. (2025). The state of AI: How organizations are rewiring to capture value. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
5. Stanford Institute for Human-Centered Artificial Intelligence (HAI). (2025). 2025 AI index report. https://hai.stanford.edu/ai-index/2025-ai-index-report
6. Gartner. (2024). Tackling trust, risk and security in AI models (AI TRiSM). https://www.gartner.com/en/articles/ai-trust-and-ai-risk
7. Wang, T., et al. (2023). AutoGPT: Autonomous GPT-based agents for task automation. arXiv. https://arxiv.org/abs/2304.03442
8. Zhou, L., et al. (2023). AgentBench: Evaluating LLMs as agents. arXiv. https://arxiv.org/abs/2308.03688
9. Zheng, Y., et al. (2024). Emergent cooperation in LLM-based multi-agent systems. arXiv. https://arxiv.org/abs/2310.01985
10. Biran, O., & Cotton, C. (2022). Explainability and trust in autonomous systems. IEEE Transactions on Human-Machine Systems, 52(5), 801–813.
11. Biran, O., & Cotton, C. (2022). Explainability and trust in autonomous systems. IEEE Transactions on Human-Machine Systems, 52(5), 801–813.
12. Mehrabi, S., et al. (2022). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.
13. Barbera, M. (2025). AI privacy risks & mitigations—large language models (LLMs). European Data Protection Board. https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf
14. Cloud Security Alliance (CSA). (2025, February 6). MAESTRO: Agentic AI threat modeling framework. https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro
15. Feng, J., et al. (2025). AAGATE: Aligning agentic AI governance with NIST RMF principles. arXiv. https://arxiv.org/abs/2510.25863
16. Tian, Y., Luo, A., Du, J., Xian, X., Specht, R., Wang, G., Bi, X., Zhou, J., Srinivasa, J., Kundu, A., et al. (2025). An outlook on the opportunities and challenges of multi-agent AI systems. arXiv. https://arxiv.org/abs/2505.18397
17. Yang, Y., Peng, Q., Wang, J., & Zhang, W. (2024). Multi-LLM-agent systems: Techniques and business perspectives. arXiv. https://arxiv.org/abs/2411.14033
18. Acharya, B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals–a comprehensive survey. IEEE Access.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 02 February 2026
DOI: 10.47191/ijcsrr/V9-i2-46, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

19. Raza, S., Sapkota, R., Karkee, M., & Emmanouilidis, C. (2025). TRiSM for agentic AI: A review of trust, risk, and security management in LLM-based agentic multi-agent systems. arXiv. https://arxiv.org/abs/2506.04133
20. National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile. NIST Trustworthy and Responsible AI.
21. Chen, S., Liu, Y., Han, W., Zhang, W., & Liu, T. (2025). A survey on LLM-based multi-agent system: Recent advances and new frontiers in application. arXiv. https://arxiv.org/abs/2412.17481
22. European Union. (2016). General Data Protection Regulation (GDPR) – Article 25: Data protection by design and by default. https://gdpr-info.eu/art-25-gdpr/
23. Lee, M., & Tiwari, M. (2024). Prompt infection: LLM-to-LLM prompt injection within multi-agent systems. arXiv. https://arxiv.org/abs/2410.07283
24. Hannebauer, M. (1999). From formal workflow models to intelligent agents. In Proceedings of the AAAI-99 Workshop on Agent Based Systems in the Business Context (pp. 19–24).

Cite this Article: Sinha, N. (2026). Building Trust in Agentic AI: TRACE Framework for Policy-Driven Multi-Agent System Design. International Journal of Current Science Research and Review, 9(2), pp. 1024-1035. DOI: https://doi.org/10.47191/ijcsrr/V9-i2-46