Indian Journal of Artificial Intelligence and Neural Networking (IJAINN)
ISSN: 2582-7626 (Online), Volume-6 Issue-1, December 2025

Meta Framework for AI-to-AI Interaction (MAI²) in Professional Contexts

Shridhar Marri, Gayatri Menon

Abstract: ***Rapid advances in Artificial Intelligence (AI) have led to autonomous agents that not only respond to humans but also interact directly with other AI agents. They are not just exchanging information but also making decisions, collaborating, and even competing as they transform several business functions. As a result, the emerging field of AI-to-AI interaction poses significant challenges around how agents collaborate and how their decisions impact business outcomes. Most existing AI agents depend on strict, rule-based communication. This approach falls short when context changes dynamically, new situations emerge, or conflicting priorities arise amongst the agents. Our research addresses these critical gaps identified through a systematic review of multi-agent systems, communication models, and interaction design. Building on the insights from our multiple-case study research on Human-AI interaction, we developed the Meta Framework for AI-to-AI Interaction (MAI²). This framework is devised around six interconnected layers that make AI-to-AI interaction reliable and trustworthy. The aspirational layer of the framework establishes the agents' goals and values, the cognitive layer supports reasoning and real-world perception, and the strategic layer focuses on planning and execution. The governance layer ensures the system remains accountable through oversight. The synchronisation layer ensures that different agents work together smoothly. The interactional layer handles the nuts-and-bolts of communication. These layers, together, outline how AI agents collaborate, coordinate, and remain aligned with human values and expectations. MAI² is designed to enable AI agents to learn from each other, evolve together, and adapt over time to collaborate responsibly and effectively. This paper aims to advance AI-to-AI interaction by providing a structured starting point while acknowledging the limitations of its validation across diverse professional contexts.***

Keywords: Agentic AI, AI-to-AI Interaction, Autonomous Agent Ecosystems, Distributed AI, Multi-Agent Systems.

Nomenclature:
AI: Artificial Intelligence
MAS: Multi-Agent Systems
ACID: Adaptive Conversational Interaction Dynamics
CSDF: Conversational Social Dynamics Framework
MAI²: Meta Framework for AI-to-AI Interaction

Manuscript received on 15 October 2025 | First Revised Manuscript received on 31 October 2025 | Second Revised Manuscript received on 06 December 2025 | Manuscript Accepted on 15 December 2025 | Manuscript published on 30 December 2025.

*Correspondence Author(s)
Dr. Shridhar Marri*, 86, 16th C Main, 4th Block, Koramangala, Bangalore 560034, (Karnataka), India. Email ID: shridhar_m@nid.edu, ORCID ID: 0009-0007-6750-9361
Dr. Gayatri Menon, Department of Vice-Chairperson, Education, National Institute of Design, Opposite Tagore Hall, Rajnagar Society, Paldi, Ahmedabad, (Gujarat) India. Email ID: gayatri@nid.edu, ORCID ID: 0009-0003-2970-3076

I. INTRODUCTION

Artificial intelligence has shifted from being a set of isolated tools to a growing ecosystem of interconnected systems. Whether negotiating traffic as autonomous vehicles or responding to market signals as financial systems, disparate AI systems need to interact with one another without constant human oversight. As these interactions become ubiquitous, how they are designed and governed remains loosely defined. Our research examines this need in greater detail. The aim is to identify key principles that make AI-to-AI interactions more effective and reliable. Without clear guidelines and frameworks, these interactions will remain suboptimal and even lead to conflict and negative impact. The proposed Meta Framework for AI-to-AI Interaction (MAI²) brings together ideas from analyses of multi-agent systems, communication theory, and interaction models to establish shared conventions. It also draws on Human-AI interaction frameworks to interpret shared goals and build trust, even when those goals are contradictory. The ultimate aim of this framework is to ensure that AI-to-AI interactions align with human values, supporting better outcomes for the people who use and depend on them. This framework offers structure without imposing rigidity and encourages cooperation without sacrificing autonomy. The framework is not meant to be a finished blueprint, but rather a foundation for systematically thinking about how intelligent systems should communicate with each other, one-on-one or in larger, interdependent networks.

II. LITERATURE REVIEW

This literature review examines several overlapping fields of AI-to-AI interaction, highlighting the state of the art and current gaps. Multi-Agent Systems (MAS) research has pioneered the study of how autonomous agents cooperate through structured protocols [1]. More recent studies have also emphasised that an agent must sense and act independently within a given environment [2]. These studies have provided valuable insights into agent coordination where the goals are clearly defined. But real-world settings are rarely well-defined. Distributed decision-making approaches have been designed to enable agents to maintain their autonomy within a larger set of system objectives. While this improves agent effectiveness, it does not resolve the tension between autonomy and control when information is incomplete. Other related approaches, such as federated learning, help resolve collaboration through knowledge sharing while preserving privacy without compromising raw data [3]. However, as AI agents continue to depend on inputs

from other AI agents to make appropriate decisions, there are security lapses and potential grounds for manipulation. Research highlights vulnerabilities to attacks, data poisoning, and model manipulation, emphasising the need for governance [4]. Furthermore, decision autonomy raises difficult questions about accountability, particularly when agents act on behalf of humans or organisations with differing priorities. Interpretability also becomes crucial, especially in situations where safety, ethics, and trade-offs are paramount [5]. Through a systematic study of the literature, we have identified three critical gaps that demand holistic approaches to address them. First, the current protocols cannot assess or adjust the reliability of AI agents, thereby failing to establish the much-needed integrity. Hence, the ethical aspects and trust mechanisms remain underdeveloped. Second, interactional adaptability is still limited in AI-to-AI exchanges. While these systems excel at dividing tasks, they struggle to infer partner intent and shift strategies mid-interaction. Finally, current systems lack safeguards against agent deception and traceability of agent decisions. Together, these gaps underscore the need for a conceptual meta-framework that holistically integrates the ethical, emotional, and interactive dimensions of AI-to-AI interactions.

III. METHODOLOGY

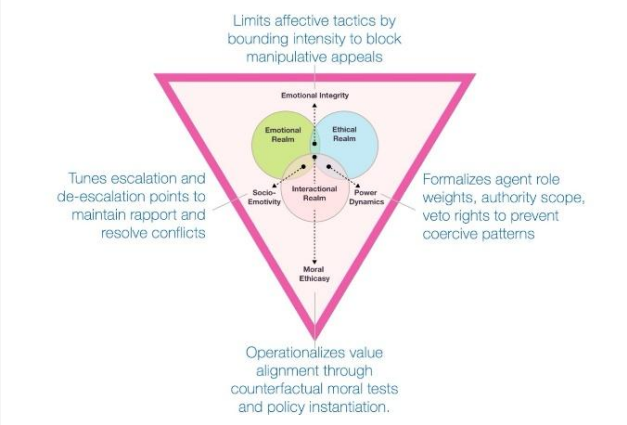
This research is based on a multiple-case study that examined how people interact with AI systems in real-world settings across six cases in India, the US, Singapore, and Indonesia. The cases were deliberately chosen to represent a wide range of sectors, domains, and geographies. The scale of organisations ranged from large enterprises and medium-scale companies to government agencies in the public sector, to capture the wide range of Human-AI interaction practices. The data collection methods included recordings of fundamental life interactions, interview transcripts, and observational notes. The analysis was conducted in three steps. First, each case was treated on its own, paying close attention to its own context, constraints, and the specific ways people and AI systems engaged with one another. Then we compared these cases to understand how similar issues played out in different settings. This cross-case comparison helped us unearth recurring patterns. In the final stage, through a multiple-case study analysis, we identified insights and developed broader themes while acknowledging underlying differences. This led to several contributions, including 12 conversational archetypes [6] and three specific frameworks: the Adaptive Conversational Interaction Dynamics (ACID) framework, the Conversational Social Dynamics Framework (CSDF), and the Conceptual Framework for Conversational Human–AI Interaction in Professional Contexts [7]. These contributions shaped our thinking in collating diverse issues that matter not only in Human-AI interaction but also in AI-to-AI interactions. These insights are now embedded into a meta-framework for AI-to-AI interaction design. We built on observations of how people and AI systems interact and translated those human-centred insights into core principles to guide autonomous AI agents.

IV. FINDINGS AND RECOMMENDATIONS

This section outlines the learnings from the multiple-case study analysis and the synthesis of the various frameworks derived thereafter. The Conversational Social Dynamics Framework (CSDF) helped in creating a foundational structure for AI-to-AI interaction design. This unearthed the "stress zones" that could potentially act as guardrails for social and ethical issues in interactions. Building on this, we conducted alluvial mapping by layering the key elements of the Adaptive Conversational Interaction Dynamics (ACID) framework. This, along with the CSDF guardrails, helped identify six essential layers vital to designing AI-to-AI interactions. Alongside these themes, several stood out that can help AI agents understand context effectively, support each other's shared goals, and function in ways aligned with expected outcomes and ethical standards. Further, by incorporating the 12 Conversational Archetypes across these six layers, we derived the key elements in AI-to-AI interactions. Taken together, these findings provide a foundation for systematically thinking about how autonomous agents should communicate and work together.

V. FOUNDATIONAL AI-TO-AI INTERACTION SCHEMA

The social principles of the CSDF framework provided the basis for how AI agents could interact. These principles, consisting of three realms and their four overlapping intersections, became the hallmarks of AI-to-AI conduct. The interactional realm encompasses turn-taking, repair, and grounding. The emotional realm highlights empathy while the ethical realm promotes fairness, trust and privacy.



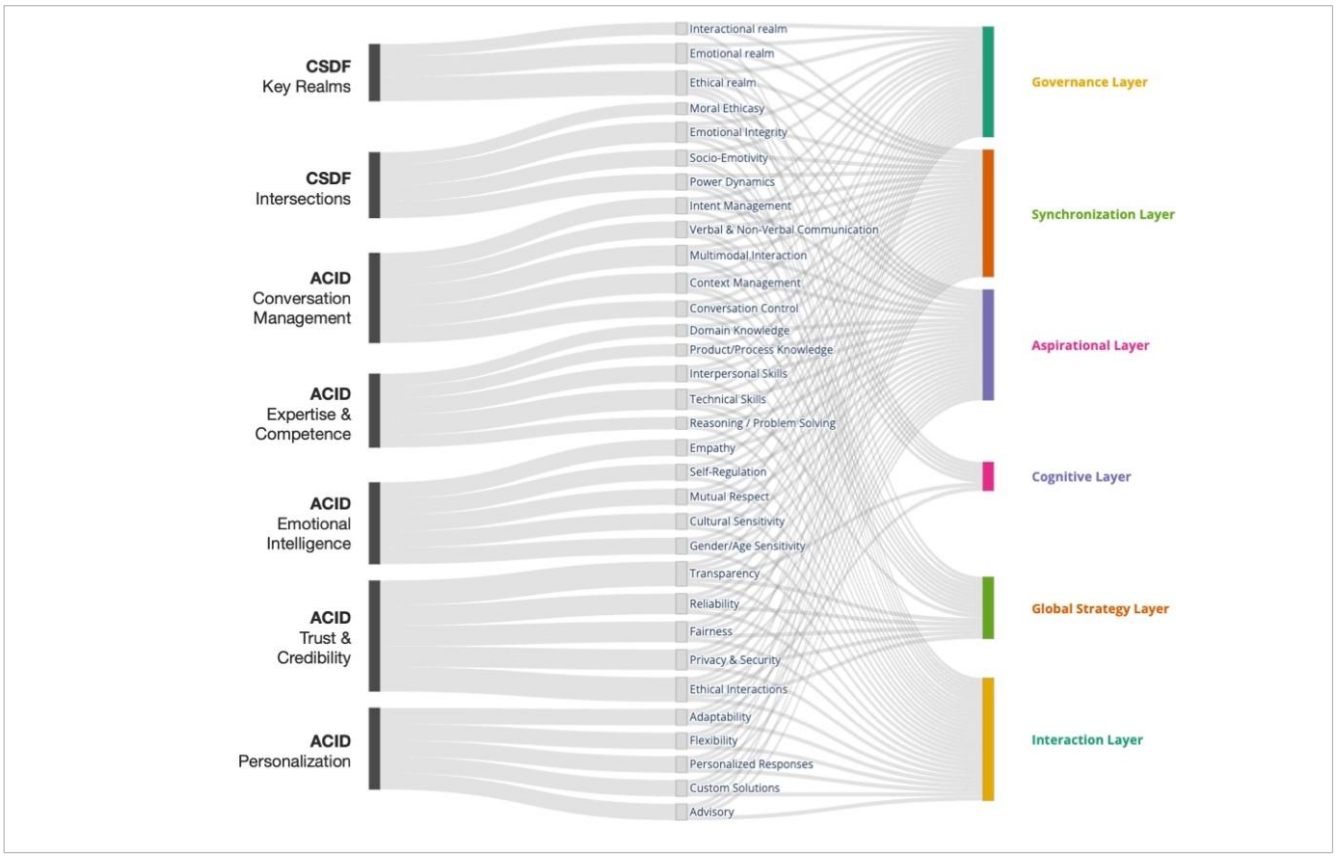
[Fig.1: Stress Zones from CSDF to Provide Guardrails]

We have designated their intersections as socio-emotivity, power dynamics, emotional Integrity and moral ethicacy. Socio-emotivity resolves conflicts by regulating emotional escalation and de-escalation. Power dynamics outlines the scope and roles of authority to avoid intimidating patterns in interactions. Emotional integrity blocks manipulative approaches by limiting the dubious or untrustworthy tactics. Finally, the moral ethics principle aligns the values of AI agents through moral testing and policy administration. These intersections, together, improve coordination, facilitate negotiation, and resolve conflicts amongst AI agents.

Thus, CSDF converts social norms into programmable, real-time protocols.

VI. SYNTHESIS OF CSDF AND ACID FRAMEWORKS

The stress zones outlined in the CSDF provided the foundation for our investigation. The zones and their interpretations were combined with the Adaptive Conversational Interaction Dynamics (ACID) framework to produce a coherent, layered interpretation. We first coded every conversational feature into CSDF's three realms and four intersections. In parallel, the ACID framework's five primary dimensions—Conversation Management, Expertise & Competence, Emotional Intelligence, Trust & Credibility, and Personalisation, along with their 25 operational elements — were harnessed. Plotting these 32 items in a Sankey plot from source groups to outcomes revealed recurring motifs. The overarching goals and motivations of an agent were organised into an Aspirational Layer. Expertise and emotional intelligence were fused into a Cognitive Layer. Collaborative planning and execution, with resource-allocation capabilities, were embedded in a Global Strategy Layer. Overseeing security and trust compliance was merged into a Governance Layer. Real-time coordination, autonomy, and knowledge sharing were integrated into a Synchronisation Layer. And finally, multimodal communication and messaging capabilities with context persistence were merged into an Interactional Layer.



[Fig.2: Derivation of Six Functional Layers for AI-to-AI Interactions]

Thus, the layered alluvial mapping helped us not only visualise the data but also derive six functional layers that are precisely aligned with the demands of autonomous AI-to-AI interaction. This analysis is plotted in a Sankey diagram above.

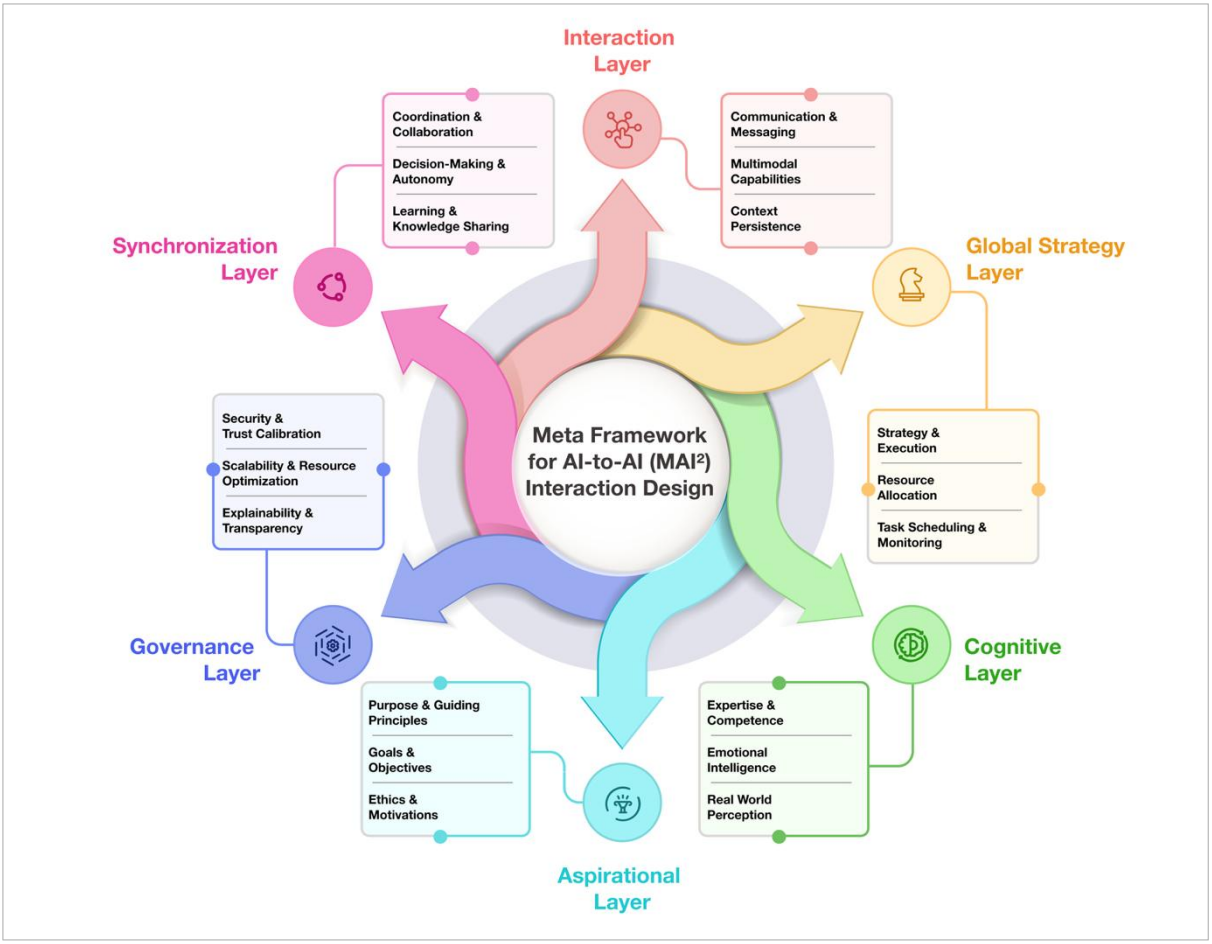
VII. INTEGRATION OF 12 CONVERSATIONAL ARCHETYPES

Further integration of the twelve Conversational Archetypes with their dialogue blueprints into six functional nodes of AI-to-AI interactions enriched each layer with purpose-specific dialogue competencies, turning them into a library of reusable interaction models. At the Interaction Layer, Informational, Casual, Awareness, and educational archetypes shape concise signals, chatter, attention cues, and explanations. They also encompass multimodal utterances and metadata tags and preserve context, ensuring seamless messaging across agent interactions. At the Synchronisation Layer, negotiation, transactional, and resolution archetypes provide structured handshakes, fair concessions, autonomy powers, and feedback loops. This enables coordinated collaboration, joint decisions, continual learning, and knowledge exchange among agents. At the Cognitive Layer, the analytical, intellectual, and educational archetypes provide reasoning abilities and intellectual faculties, ensuring that interactions occur with perceptive emotional intelligence. Resolution archetypes further allow agents to resolve conflicts between AI agents. The Aspirational Layer emphasises the purpose and values of AI agents. It uses the motivational, persuasion, and advisory archetypes to help them align not only with their stated objectives but also with their intrinsic motivations. The Global Strategy Layer sequences archetypes such as awareness and persuasion to address strategic needs. In contrast, transactional and

analytical archetypes provide the resources needed to schedule tasks and monitor AI agents' progress. Finally, the Governance Layer oversees enforcement and uses several archetypes to ensure agents adhere to objectives. The transactional archetype creates audit trails, and the analytical archetype provides oversight. Thus, the 12 conversational archetypes transform the six layers from a static architecture into a blueprint for socially intelligent, task-optimised AI-to-AI interactions.

VIII. META FRAMEWORK FOR AI-TO-AI INTERACTION (MAI²) DESIGN

The resultant Meta Framework for AI-to-AI Interaction (MAI²) presents a comprehensive, multi-layered architecture that enables seamless, adaptive AI-to-AI interactions. This six-layered framework integrates components for aspiration, cognition, strategy, governance, synchronisation, and interaction, ensuring a structured yet flexible approach to AI-to-AI Interaction, decision-making, and coordination.



[Fig.3: Meta Framework for AI-to-AI Interaction (MAI²) Design]

- **Aspirational Layer:** Defines intrinsic objectives and motivations of AI agents as guideposts
- **Cognitive Layer:** Combines expertise and competence with emotional intelligence
- **Global Strategy Layer:** Overseas planning and execution with resource allocation
- **Governance Layer:** Governs compliance with transparency and oversees security and trust
- **Synchronisation Layer:** Orchestrates real-time coordination, autonomy and learning approaches
- **Interactional Layer:** Outlines multimodal communication and messaging with context persistence

These layers form the quintessential architecture for AI-to-AI interactions, enabling fluid, collaborative and yet ethically guided exchanges. Further operationalised into 18 core components, these layers create a comprehensive blueprint for AI-to-AI interactions. The core components provide the foundation for AI agents to function independently and collaborate without compromising security or transparency. They help agents communicate and coordinate their actions while continuously learning and adapting to dynamic situations. Overall, they ensure that the agent's behaviour aligns with human needs and values. Thus, the MAI² framework is designed to provide a robust starting point for building AI systems that can learn from and interact with one another without constant human oversight. The primary objective of the MAI² framework is to design and develop autonomous, interoperable AI agents that share a sense of purpose, with room for growth and ethical alignment with human values and expectations.

A. Layer I: Aspirational Layer

This foundational layer sets the primary goals of each AI agent. It provides clarity on what the agent is expected to achieve within the boundaries of its role. It also presents a clear sense of purpose, and the following three aspects collectively form the core components of this layer.

i. Purpose & Guiding Principles

Purpose and guiding principles describe the mission and core values that shape an AI agent's behaviour. They set the direction that an agent should follow and ensure that the agent acts in line with the intended role. In real-world settings, this alignment becomes crucial. In an autonomous supply chain, when a logistics agent is tasked with on-time deliveries, reducing emissions is also a critical responsibility. The agent might need to choose low-carbon footprint routes while optimising delivery schedules. This larger responsibility is driven by the global purpose and guiding principles component.

ii. Goals & Objectives

Goals and objectives define the specific targets that an AI agent is supposed to meet. While they operate strictly to achieve the targets set out for them, they also need to be cognizant of the dynamic situations with varied outcomes. For example, a trading agent in financial markets might be working to improve its client's returns, but it must remain within the client's risk profile. As markets shift, the agent could adjust its tactics to achieve its goals in response to dynamic market conditions.

iii. Ethics & Motivations

The ethics and motivations layer defines how AI agents would make decisions in a given scenario. It sets norms for AI agents to be fair and accountable for their choices. It aims to reduce biases and prevent agents' actions that can cause unintended harm. For instance, accuracy and patient safety must be primary considerations in healthcare settings. Diagnostic agents that assess radiology scans should clearly flag ambiguities and uncertainties and seek human or other AI validation before making critical decisions or recommendations.

B. Layer II: Synchronisation Layer

The Synchronisation layer orchestrates real-time interactions among agents. It defines how different agents coordinate and collaborate to achieve their individual and collective goals. It ensures that the agents do not follow divergent paths that are misaligned with the overall objectives. It also outlines how agents share knowledge as they learn independently and collectively over time.

i. Coordination & Collaboration

Coordination and collaboration are essential for agent teams to work together and ensure there is no duplication or redundancy. In many scenarios, some agents might lead certain efforts, while others follow orders within a clear hierarchy. And some agents would operate as peers. This would call for clearly laid-out arbitration rules and consensus-led settlements on the go. In these varied settings, agent work spanning tasks needs to be synchronised in an orderly manner. In swarm robotics, drones would have to undertake their functions with clear delegation and collaboration.

ii. Decision-Making & Autonomy

AI agents often need to make decisions independently while working towards collective goals. The autonomy at each agent level makes them efficient and responsive in achieving their core objectives. Autonomy within the boundaries of broad governance frameworks ensures that the tasks are performed with speed and scale. In emergency response situations, agents need to direct resources where they are most needed through an independent, rapid assessment that avoids wasted time. This requires autonomy and decision-making powers, with accountability.

iii. Learning & Knowledge Sharing

AI agents learn and acquire knowledge from one another in many different ways. They need to build on each other's strengths and adapt their behaviour to day-to-day interactions. Federated learning helps them train on shared objectives without transferring raw data, thereby keeping sensitive information private. Transfer learning allows each agent to benefit from the teaching of other agents. Together, these processes contribute to continuous learning within the team. Learning from these experiences helps the agents make better decisions over time so that they do not need to start from scratch every time they face a new situation. For example, in medical diagnostics, when an agent detects an anomaly in a patient's health data, it can pull in other agents to weigh in based on their knowledge to guide better treatment decisions.

C. Layer III: Cognitive Layer

The cognitive layer defines an agent's thinking, understanding and reasoning abilities. It addresses three critical aspects. Expertise and competence encompass an agent's skills, domain knowledge and proficiency. Emotional intelligence shapes the agent's outlook by enabling it to identify emotional cues and respond appropriately. Real-world perception abilities help an agent decipher the environment in which it operates and maintain awareness of what is happening around.

i. Expertise and Competence

AI agents need to primarily rely on their expertise and competence to achieve their key tasks. Their roles and areas of expertise in specialised domains equip them to handle tasks competently with no shortcomings. Task allocation to agents needs to be based on their core competencies and aligned with the domain and task at hand. In finance, for example, agents need to be adept at risk analysis and able to spot unusual transactions. In healthcare, accurately interpreting test reports and making treatment recommendations is a critical skill. AI agents would also need to undergo periodic certification to demonstrate that they possess the required skills and are continually upgraded.

ii. Emotional Intelligence

AI agents that operate in collective environments need more than just speed in task completion and accuracy in outcomes. They need to understand the needs and intentions of other agents in the loop and adjust their behaviour when warranted. While this is similar to emotional intelligence in people, it does not mean emotions in the true sense. In the context of autonomous agents, they need to pick up signals from other agents' intent, interpret them, and respond in ways that support collective goal achievement. It is about understanding the nuances of the context, reading shifts in intent, and reacting appropriately.

iii. Real-world Perception

Real-world perception helps AI agents understand what is happening around them. When they can observe their own environment, they become context-aware and can make appropriate decisions for the situation at hand. Agents would need to observe how people react to them, identify patterns, and respond more accurately to dynamic demands. In self-driving vehicles, one agent might track lane markings, while another might look out for obstacles, pedestrians, or changes in traffic. These agents need to work in tandem to prevent untoward incidents. These agents need to be flexible rather than rigidly follow rules. Constant environmental tracking helps agents make safer, more responsible navigation decisions.

D. Layer IV: Global Strategy Layer

The global strategy layer defines how agents plan and carry out their tasks. It manages and sets the overall direction and clearly articulates the steps to achieve defined goals. Task scheduling and monitoring determine the order in which tasks are performed while tracking overall progress to detect any issues. Resource allocation ensures that tasks are assigned to the right agents and that the required data or resources are available as and when needed. This helps the AI agents stay focused, organised and efficient.

i. Strategy & Execution

The strategy & execution component provides clear direction for AI agents to achieve their goals. It helps agents break tasks into clear, manageable steps. It also synchronizes the high-level strategy with concrete, reliable action plans. Take supply-chain automation as an example. Logistics agents who are required to manage supply and demand need to break down tasks into multiple activities while coordinating with warehouse agents, transport scheduling agents, and fulfilment agents. This helps them execute their plans effectively and run operations smoothly.

ii. Task Scheduling and Monitoring

Task scheduling and monitoring enable agents to focus on how work is planned, tracked, and evaluated. It ensures that every agent knows exactly what they need to do and when to carry out the specified tasks while keeping an eye on performance and completion. Even when conditions change, the agents must stay on course to ensure the intended task is completed satisfactorily. In AI-driven manufacturing scenarios, different agents handle specific aspects of the workflow. One agent schedule and monitors robotic assembly, and another handles material procurement. Yet another agent might control quality. Scheduling and monitoring allow all these agents to work in tandem, track their progress, and address any bottlenecks in the process.

iii. Resource Allocation

The resource allocation component defines the array of resources the agents need to execute their tasks. Most often, resources such as computing power, data access, and infrastructure need to be shared across a team of agents to prevent usage spillover. It also provides basic load balancing with required backups. Using a priority matrix helps coordinate resource use and maintain a well-balanced overall performance. In cloud-based systems, optimised resource allocation helps agents adjust their power, storage, or compute needs based on the workload. More resources are deployed based on agents' specific needs and scaled back when demand drops. This helps us use resources more efficiently without wasting capacity.

E. Layer V: Governance Layer

The governance layer outlines the basic rules that make the agents accountable. Security and trust focus on how agents monitor each other. The scalability and resource-optimisation component address the need to deploy more agents when required. The governance layer also deals with explainability. It tracks agents' activities and provides a detailed account of why they behaved as they did or how they made their decisions.

i. Security & Trust Calibration

Security and trust calibration layer sets the guidelines for AI agents to authenticate and validate one another. They would need to perform various checks, such as authorisation, access rights, and verification, to address potential threats and risks. Each agent needs to evaluate the trustworthiness of every agent they encounter in conducting their business and decide when and what kind of information to provide, while remaining vigilant. In scenarios such as financial transactions, agents need to continuously monitor interactions with other agents and ensure legitimacy to avoid manipulation.

ii. Scalability & Resource Optimization

In professional settings, workloads tend to either increase or decrease based on various conditions. Scalability and resource optimisation are critical components for making AI agents perform at their best. This component ensures resources are allocated where they are most needed. It also helps distribute tasks in a balanced manner, so that agents are neither overloaded nor underutilised due to resource constraints. This keeps performance steady while handling downtime or heavy demand. In a networking scenario, when a group of agents faces a surge in activity, more processing power and bandwidth are shifted to them. Once the load reduces, the same resources are reassigned to other agents that need them most.

iii. Explainability & Transparency

AI agents must always be able to explain their actions, decisions and outcomes. Every agent should provide detailed logs of their actions and how they interpreted various scenarios before making specific decisions. This makes it easier to audit and trace their choices and actions, while improving agent performance. People can trust agents more when they can see their decision-making trail. In a clinical diagnosis setting, when an agent identifies a specific condition in a patient, it should also provide a step-by-step process for arriving at that conclusion. This gives other agents, doctors and even patients a clear picture of how the diagnosis was conducted.

F. Layer VI: Interaction Layer

The interaction layer defines how the AI agents communicate with each other and how the information is exchanged across steps. Communication and

messaging capabilities address this, while multimodal expertise supports the use of text, images, audio, and video in the interaction. The context persistence component tracks interactions over time to ensure workflow continuity.

i.Communication & Messaging Capabilities

AI agents need a communication framework to exchange information and interact with one another. This requires shared data formats such as JSON, XML, Protobuf, and WebSockets to enable real-time information flow. Messaging models require two-way encryption and reliable authentication. And they need to support both synchronous and asynchronous exchanges. Semantic layers, when embedded in these protocols, help arrive at a shared understanding of what the messages mean, preventing misinterpretation. Hierarchical information dissemination also helps trace the origin of information and how it is followed through the workflow. Together, these elements provide a robust foundation for seamless communications and messaging.

ii. Multimodal Expertise

AI agents need to process information from different forms, such as text, speech, images, video, and even structured data. They should be able to process these different modes in a unified manner to achieve a holistic understanding and ensure that inputs and outputs are not fragmented. This requires them to combine language and visual interpretation along with reasoning abilities. Multimodal expertise helps agents' piece together the complete picture of a situation while enabling them to adapt to improve collaboration.

iii.Context Persistence

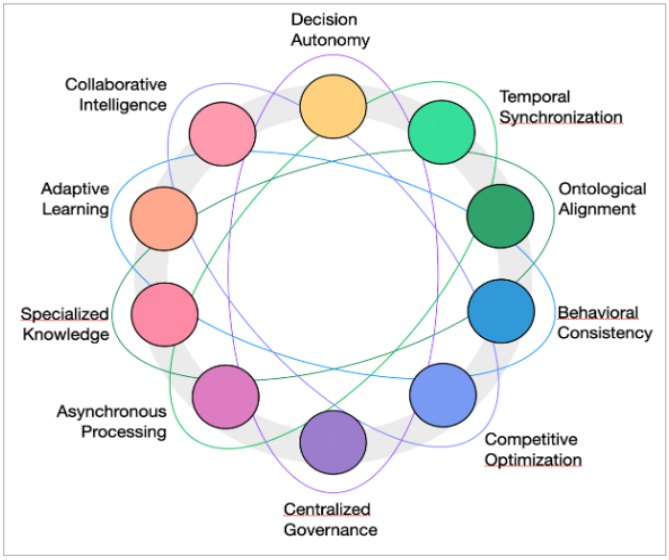
Context persistence allows AI agents to remember and retrieve what has transpired in previous interactions in a given situation, without the need to reestablish objectives or the workflow. This shared memory and context preservation help teams of agents pick up where they left off and continue their tasks. It is especially crucial when agents undertake complex workflows and decisions evolve gradually over time. This helps agents achieve coordinated interaction, maintain coherence across all activities, and even rely on what worked best in similar situations.

IX. CROSS-CUTTING PRINCIPLES OF THE MAI² FRAMEWORK

The six fundamental layers of the MAI² framework operate in tandem rather than in isolation. They are interconnected through a set of guiding principles that make the framework interdependent. These cross-cutting principles help autonomous agents with shared values, expectations, and constraints, and act as a glue for collective intelligence and execution. They ensure that the six layers do not become black boxes. Value alignment with human goals is the most crucial principle that cuts across the layers. This prevents misaligned incentives and keeps the agents grounded without losing sight of the human needs they are meant to serve. Interpretability and traceability are another principle that enables flexible, cross-platform agent collaboration, making each decision or interaction auditable and understandable. Trust, reciprocity, and fair contribution are the third principles that make agent teams' work reliable and non-exploitative. This ensures that agents clearly reveal their assumptions to decision paths. Robust access controls and understandable explanations keep the AI-to-AI interactions accountable. The fourth principle, conflict resolution and ethical boundaries, helps identify conflicting goals to resolve them constructively. The fifth principle, adaptive learning and ambiguity handling, helps agents learn through a multistep process while continuously improving. It also helps manage uncertainty through reasoning, negotiation, and collective coordination. Together, these five cross-cutting principles of the framework serve as conduits, making the diverse work of AI agents collaborative and aligned with human values and goals.

X. DIVERGENT STRUCTURAL TENETS

While the cross-cutting principles help agentic AI systems remain coherent, some parts of the MAI² framework tend to overlap or pull in different directions. To avoid conflicts and reinforce agent integrity, we have identified five sets of divergent structural tenets that need to be optimised amid competing priorities. They would have to be weighed against each other and, in some cases, negotiated for a workable middle path to ensure balanced outcomes.



[Fig.4: Divergent Structural Tenets]

i. Decision Autonomy vs. Centralized Governance

AI-to-AI interaction design must deal with the familiar tension between autonomy and governance. On the one hand, too much independence makes the agents unpredictable and too much control makes them inflexible. A practical middle path is to let agents act on their own in everyday situations and to introduce clear guidelines and boundary conditions for unfamiliar use cases. Policy-driven autonomy helps agents operate with sufficient flexibility without compromising oversight. In a smart manufacturing plant, robotic arms work independently to ensure the assembly line operation runs smoothly and efficiently. At the same time, quality monitoring agents who continuously assess the robotic arms' output quality need to plunge in and take over the operation when quality drops below required levels to fine-tune the parameters and fix quality issues. While autonomy helps

each agent do its job well, oversight prevents small mistakes from turning into significant failures.

ii. Collaborative Intelligence vs. Competitive Optimization

All AI agents in an autonomous system may not have the same goals. While they are expected to collaborate primarily, they may also be required to compete. They also get to decide when to work together and when to act on their own. Cooperation typically helps a team of agents achieve their objectives; they are expected to compete with other agents and improve their own performance. The agent interaction strategy should accommodate both. Usually, agents are rewarded for collaboration, which benefits both the group and individual agents. They can also be incentivised to compete and improve. If only cooperation is rewarded, agents become too complacent to improve; if competition dominates, they can become selfish, which can have a detrimental impact on outcomes. The goal is to create guidelines and rules that restrain manipulation while still helping each agent excel in its tasks and behaviour.

iii. Adaptive Learning vs. Behavioural Consistency

As the AI agents continue to adapt and evolve, they also need to be predictable. If they are too slow to change, they become outdated. If they are changing relentlessly, they become unreliable. Agent evolution needs to be within clearly defined parameters. One way to do this is to assign core dimensions that remain consistent, while other parts can continue to evolve. And for the parts that are changing, versioning logs help record and communicate those changes to stakeholders. For example, in healthcare scenarios, when an agent encounters a new disease pattern or symptoms it has never experienced before, it should make the update visible to all agents. And explicitly state that the recommendation is based on newly acquired data. This ensures that doctors and patients receive improved diagnostics without abrupt, unexplained shifts in agent behaviour or communication.

iv. Temporal Synchronization vs Asynchronous Processing

AI agents need to combine synchronous and asynchronous processing to handle interactions more effectively. Some tasks need to be performed by different agents at the same time, while others can run in the background without a dependency on any agent's task or goal. This makes it essential to include timing-related metadata, not just when an action was taken, but also how long it lasted, when a response was sent, and whether the action aligned with the timing of other agents' tasks. This helps every agent understand what requires immediate attention and what can wait. In a fleet of autonomous cars, when a vehicle detects a potential hazard, it shares collision-avoidance data with all other vehicles likely to be affected in real time. Simultaneously, each car starts exploring alternate routes while calculating the time to the destination, without compromising safety or critical response times.

v. Ontological Alignment vs Specialized Knowledge Domains

Ontological alignment describes how different AI agents agree on the meaning of things. They need a shared understanding of concepts, definitions, and interpretations for both specific and generic activities. If this alignment is missing, agents may interact with others without really understanding what each agent means. To achieve this alignment, agents need broad-based ontologies or knowledge structures with detailed reference points that explain the concepts across their workflows. Apart from these shared knowledge structures, which provide a common understanding, the agents also need to rely on detailed, domain-specific knowledge constructs to ensure better outcomes. For example, financial and legal AI systems sometimes need to work together on the same contract. To do that, they rely on a shared vocabulary for basic ideas such as the parties involved, their obligations, and the timelines they agree to. Within this broad-based understanding, even when each agent operates within its expertise cluster, it needs a shared sense of other agents' domains. For example, financial agents focus on risk ratios and payment projections, while legal agents view interactions through the lens of precedent. When these two agents interact, they may not exchange information purely through their technical language. They might even rely on a translation agent to interpret concepts from each domain and make the insights available in a shared vocabulary.

XI. CONCLUSION

The Meta Framework for AI-to-AI Interaction (MAI²) provides a structured approach that enables AI agents to collaborate more reliably and efficiently. The framework lays out six interconnected layers that define how AI agents interact, communicate, share goals, make decisions and complete tasks. Across these layers, eighteen different components address various nuances of AI-to-AI interaction in professional settings. The framework takes a holistic approach to AI-to-AI interactions by addressing intrinsic motivations, reasoning abilities, and collaborative strategies at one end, and planning, execution, real-time coordination, and interactions at the other. It also highlights several cross-cutting principles that act as the connective tissue for all the layers. Aligning with human values, resolving conflicts, negotiating outcomes, learning from other agents, reposing trust, and handling ambiguity are among the key principles of the Framework. It also defines the divergent structural tenets, such as finding the right balance between freedom and accountability, deciding when to compete and when to collaborate with other agents, and the need to evolve constantly while maintaining predictable outcomes. Thus, MAI² helps AI agents work smoothly with people and other agents to handle diverse and demanding situations in real-world professional contexts.

XII. LIMITATIONS

AI-to-AI interaction design is still a young field. New ideas emerge quickly, but most of them have not yet been tested outside controlled environments. With few scalable implementations, we have only a limited sense of how well these concepts might hold up in real-world situations. We have studied and analysed publicly available open-source multi-agent architectures and experimental platforms which are in pilot stages. Since several proprietary agentic systems currently under development in large technology companies

are not accessible for evaluation, there is a noticeable gap in achieving a holistic understanding of the emergent phenomena. These proprietary systems might offer valuable insights that could further inform this framework's development. While using recurring patterns and ideas from Human-AI interactions is helpful, they may not be fully adaptable to AI-to-AI exchanges. Another limitation is that the framework focuses on technical and design considerations. This leads to overlooking critical socio-cultural implications of AI-to-AI interaction.

XIII. FUTURE WORK

The Meta Framework for AI-to-AI Interaction opens up several possible new directions. Developing technical guidelines and reusable tools for AI-to-AI interaction design can help organisations build scalable, reliable agentic systems. The second priority is to test the framework in varied professional contexts. These cross-industry trials would help identify the strengths and highlight areas for improvement in the framework. Third, metrics and benchmarks need to be developed to compare different agentic systems that can measure success. Finally, there is also a need for a deeper study into how autonomous AI agents make decisions with conflicting goals and operate simultaneously across different domains. These directions would broaden this research and support the vision of building agentic AI systems that are responsible, reliable and trustworthy.

DECLARATION STATEMENT

Some of the references cited are Outdated, noted explicitly as [1]. However, these works remain significant for the current study, as they are pioneering in their fields.

As the article's author, I must verify the accuracy of the following information after aggregating input from all authors.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organisations or agencies. This independence ensures that the research is conducted objectively and without external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Stone, P., & Veloso, M. (2000). Multiagent systems: A survey from a machine learning perspective. Autonomous Robots, 8(3), 345–383. https://doi.org/10.1023/A:1008942012299, works remain significant, see the declaration
2. Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson. http://lib.ysu.am/disciplines_bk/efdd4d1d4c2087fe1cbe03d9ced67f34.pdf
3. McMahan, H. B., Moore, E., Ramage, D., & Hampson, S. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) (Vol. 54). PMLR. https://proceedings.mlr.press/v54/mcmahan17a.html
4. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S. J., Belfield, H., Farquhar, S., Lyle, C., … Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1802.07228
5. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608
6. Marri, S. (2024). 12 Conversational Archetypes for Human-AI Interaction. International Journal for Multidisciplinary Research, 6*(3), Article 23226. https://doi.org/10.36948/ijfmr.2024.v06i03.23226
7. Marri, S. (2024). Conceptual Frameworks for Conversational Human-AI Interaction (CHAI) in Professional Contexts. International Journal of Current Science Research and Review, 7(10), 7842–7853. https://doi.org/10.47191/ijcsrr/V7-i10-42

AUTHOR'S PROFILE

Dr. Shridhar Marri, CEO and Co-Founder, Senseforth AI Research Private Limited. Dr Shridhar Marri is the CEO and Co-Founder of Senseforth AI Research Private Limited, a multi-experience AI Platform serving global enterprises across various domains. He is a serial entrepreneur working at the intersection of technology and design. At Senseforth, he has built Flyfish.ai, a generative AI product to enhance the customer experience, and PitchDark, an agent-based sales transformation platform. Under Shridhar's leadership, Senseforth was granted patents in the US and India for its proprietary technology. He has also co-founded Moonraft Innovation Labs, one of India's largest experience design groups. Before this, he served as a Vice President at Infosys Technologies and worked with Fortune 500 companies in the US, Europe and APAC. As an alumnus of the Indian Institute of Management, Calcutta, and the National Institute of Design, Ahmedabad, he brings a unique blend of B-School strategy and D-School thinking.

Dr. Gayatri Menon, Activity Vice-Chairperson, Education. Dr Gayatri Menon is a principal faculty member at the National Institute of Design with nearly 25 years of experience in design education, research, and practice. Currently, she heads the Design Foundation program at NID. She has also worked as a project head and design consultant for several industries, including public-sector and socially relevant projects such as the UNIDO project for the development of SSI, the GI project for the craft sector, the setting up of the Jharkhand Institute of Craft and Design (JICD), Think Design Space for TCS, etc. She has been a board member of ITRA (International Toy Research Association), a member of the expert committee for NCF and NEP2020 at NCERT (National Council of Educational Research and Training), and is currently a member of the Senate at IITJammu and IIT-Gandhinagar. She has been invited to serve as a tutor for International Creativity Workshops held in Italy, the UK, and Germany, and as a visiting faculty member at design schools in India, South Africa, and Canada.