Explainable and Agentic AI (XAI 2.0): Causal Traceability for Trustworthy Autonomous Systems

Steven Tidwell
East Texas A&M
stidwell4@leomail.tamuc.edu

***Abstract*—With the rapid advancements in Artificial Intelligence, explainability must evolve alongside the capabilities of modern agentic systems. Traditional XAI methods were developed to explain why a model produced predictions in the way it did, but autonomous agents introduced a different challenge. Rather than generating a single output, agentic systems plan, reason, retrieve information from memory, and make a sequence of internal decisions before arriving at a final response. In this context, the explainability question shifts from "Why did the model produce this prediction?" to "Why did the model choose these actions?" Explainability is no longer limited to interpreting an individual prediction; it must account for the reasoning process behind that prediction. As autonomous AI systems continue to move into high-impact areas, developers, regulators, and end users will require greater visibility into those decision pathways in order to establish confidence in the model's behavior.**

In this paper, I will introduce XAI 2.0, a traceability framework designed to address this emerging challenge. The discussion begins with the foundations of explainable AI before looking at how agentic agents expose limitations in existing approaches. Building on that foundation, the paper examines several promising research directions, including decision provenance graphs, reasoning audits, and human oversight mechanisms. These mechanisms are capable of explaining complex, multi-step autonomous behavior. Rather than replacing XAI techniques, the proposed framework will extend them to capture complete decision pathways instead of isolated model predictions.

Alzheimer's disease diagnosis provides an appropriate case study for evaluating this approach because clinical decision-making rarely depends on single predictions. Clinicians integrate neuroimaging, cognitive assessments, biomarker results, patient history, and longitudinal observations before determining the stage of progression. An agentic AI system supporting this workflow should explain, in detail, its final recommendation, but also the sequence of decisions and evidence that produced it. Publicly available Alzheimer's datasets are identified as potential benchmarks for evaluating causal traceability methods within the Alzheimer's disease diagnostic domain.

***Index Terms*—Explainable AI, Agentic AI, Causal Traceability, Trustworthy AI, Multi-Agent Systems, Autonomous Systems**

I. Introduction

A. What is Trusted AI?

As artificial intelligence continues moving from research labs into real-world applications such as healthcare, finance, transportation, and autonomous systems, trust has become one of the most important challenges facing adoption. Trusted AI is built on many key principles: reliability, transparency, accountability, safety, and human oversight.

Reliability requires that an AI system produce consistent and accurate results across different environments and operating systems. A model that performs well during testing but behaves unpredictably in production cannot be considered trustworthy, regardless of its benchmark scores [14].

Transparency ensures that users and stakeholders can understand how an AI system reaches its conclusions, a principle increasingly reflected in regulations such as the European Union's AI Act [18].

Accountability establishes clear responsibility for AI driven decisions and provides mechanisms for identifying failures and preventing them from recurring [6].

Safety ensures that AI systems operate within acceptable boundaries and continue to behave in a predictable way even when faced with unexpected situations, adversarial inputs, or changing conditions [11], [12].

Human oversight preserves control by allowing operators to monitor, intervene, and override autonomous decisions, when necessary, a requirement reinforced by both GDPR and the EU AI Act [18].

While these principles are usually discussed separately, they are interconnected. An AI system may be accurate, but without transparency users have no way to understand when or if it may fail. Also, transparency without accountability provides visibility into a problem without establishing responsibility or a means of correcting it. At the center of these principles lies explainability, which serves as the foundation for understanding how AI systems make decisions.

B. Why Explainability Matters

Traditional AI systems, especially those built using a deep neural network, are usually viewed as a "black box" because their processes are difficult to understand, even for the people who created them [1]. As these systems grew and became more powerful, they have the tendency to influence decisions with real-world consequences. Researchers began to recognize that accuracy alone was no longer enough to show transparency. In response to this challenge, the Defense Advanced Research Projects Agency (DARPA) launched the Explainable AI (XAI) program in 2015. The goal of this program is to help users understand, trust, and manage intelligent systems [1]. One assumption of the DARPA initiative was that there is often a tradeoff between performance and explainability. Highly accurate models such as deep neural networks can achieve great results but are difficult to interpret, while simpler models

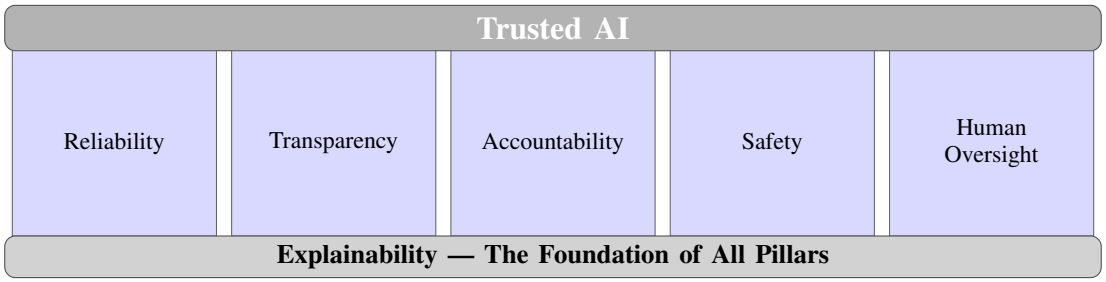

Fig. 1. Pillars of Trusted AI. Reliability, Transparency, Accountability, Safety, and Human Oversight collectively support Trustworthy AI, with Explainability serving as the common foundation beneath all five pillars.

are often easier to explain but may sacrifice predictive performance [1].

The need for explainability becomes even more important in environments where AI-assisted decisions can directly affect human lives, finances, and safety. In healthcare, AI systems are increasingly used to assist with diagnosing diseases and identifying patient risks. However, physicians must understand not only what prediction was made, but also the reasoning behind it. Vimbi et al. [14] found that frameworks such as LIME and SHAP have become the dominant approaches for interpreting machine learning models in Alzheimer's disease detection, yet the underlying black-box nature of these systems remains a significant barrier to widespread adoption. In finance, explainability serves both practical and regulatory purposes. Models used to evaluate credit risk, detect fraudulent activity, and support investment decisions make it essential that organizations can justify how those decisions are reached. Regulatory frameworks such as GDPR Article 22 and the Basel Committee's FRTB require organizations to provide meaningful explanations for automated decisions that affect individuals [4], [16]. Similarly, autonomous systems such as self-driving cars must be able to justify safety-critical decisions that occur in real time. As Tahir et al. [11] observe, a lack of transparency in autonomous vehicles creates challenges for safety, and public trust.

Across these platforms, the requirements remain the same: users need to understand what decisions were made and why. When AI systems cannot answer those questions, trust begins to erode regardless of how accurate the system may be.

C. Rise of Agentic AI

The AI landscape has changed dramatically over the past few years. Traditional AI systems were generally designed to perform a single task, such as classifying an image, predicting a value, or recommending a product. Once a prediction was generated, the process was finished. Modern AI systems, however, are increasingly designed as autonomous agents capable of pursuing goals, making decisions, and interacting with their environment. These systems are commonly referred to as agentic AI. Agentic systems are characterized by their ability to reason, plan, act, and adapt in pursuit of a defined objective rather than simply generating a single prediction [21]. Advances in large language models, enabled by the Transformer architecture, have accelerated this shift by providing systems with the ability to understand natural language, generate responses, and interact with external tools and services [21]. Because of this, modern agents can execute code, query databases, and make API calls.

What makes agentic AI different from traditional AI is the complexity of its decision-making process. Rather than producing a single output, agentic systems often decompose goals into smaller tasks, evaluate alternatives, select appropriate tools, retrieve relevant context, and execute a sequence of actions to accomplish an objective. What this means is that an agent may plan a workflow, reason through several options, call external services, and then adjust its behavior based on the results it receives. Frameworks such as ReAct, AutoGen, CrewAI, and LangGraph have demonstrated that these capabilities can be combined to create a more robust multi-agent system [8]. Recent research has also begun exploring how explainability techniques can be adapted to these environments. For example, Yamaguchi et al. [8] proposed an agentic XAI framework that combines traditional explainability methods with iterative reasoning and refinement, improving explanation quality by as much as 30–33% over baseline approaches. These advances have expanded the capabilities of AI systems, but they have also introduced a new challenge for explainability. Unlike traditional models that generate a single prediction from a single input, agentic systems produce a chain of interconnected decisions where each action influences future actions. Understanding these decision chains has become one of the central challenges facing trustworthy autonomous AI systems.

D. Research Problem

Most explainability techniques used today were developed for a very different generation of AI systems. Methods such as SHAP, LIME, attention maps, and saliency heatmaps were designed to explain why a model produced a particular prediction. SHAP uses game theory to estimate the contribution of individual features to a prediction, while LIME creates simplified local models that approximate the behavior of more complex systems [3]. Attention maps and saliency visualizations provide additional insights into which portions of an input influenced a model's output. These approaches are

valuable for understanding the classification and prediction models because they answer questions straightforwardly.

For traditional machine learning systems, this level of explanation is often sufficient. Research by Salih et al. [3] demonstrates that SHAP and LIME can successfully identify important features and improve interpretability in classification tasks. However, even within their intended use cases, SHAP may produce different explanations depending on the model being analyzed, while LIME relies on assumptions that can oversimplify complex nonlinear relationships. Both approaches focus on explaining a single prediction rather than a sequence of decisions.

This limitation becomes more apparent when applied to agentic AI systems. Consider a user asking an AI assistant, "Who worked the most overtime last month?" A traditional model will simply return an answer, but an agentic system will perform a series of actions before generating a response. Each step the agent takes represents a decision that will determine what the next step the agent will follow. If the final answer is incorrect, it requires understanding the entire decision path that produced it, to fix it.

Dazeley et al. [9] note that many existing XAI approaches focus on what they describe as reaction-level explanations, providing insight into outputs without explaining the intentions, goals, or reasoning processes that produced them. Agentic systems operate at a much higher level of complexity, combining planning, reasoning, memory, tool use, and autonomous execution to achieve objectives. Explaining these systems requires more than feature importance scores or visualization techniques. It requires the ability to understand the chain of decisions that occurred between the request and the response. This paper argues that bridging this gap requires a shift toward causal traceability, in which every significant decision made by an autonomous agent can be traced, explained, and audited throughout execution.

II. Literature Review

A. Traditional Explainable AI

The concept of explainable AI is not a new topic. In fact, the need to explain AI decisions has existed almost as long as AI itself. Ronge et al. [2] trace the history of explainability back to early expert systems such as MYCIN, which could explain the rules and reasoning used to reach a conclusion. Because these systems relied on human-authored rules, their decision-making processes were inherently transparent. The challenge emerged as machine learning and deep learning models became more powerful. While these models achieved significant improvements in predictive performance, they also became increasingly difficult to interpret, creating what is commonly referred to as the "black box" problem [1].

Over the past decade, several techniques have emerged to address the majority of these issues. Among the most widely adopted are SHAP and LIME. SHAP applies the principles from cooperative game theory to estimate the contribution of individual features to a model's prediction, providing local explanations for decisions and global explanations across the entire dataset [3]. LIME takes a different approach to the problem by building smaller surrogate models that approximate the behavior of a more complex system for a specific prediction [3]. Both techniques have become popular because they can be applied to existing models without requiring many changes to the underlying architecture. Research has demonstrated their effectiveness across a wide range of domains, including healthcare, finance, anomaly detection, and sentiment analysis [3], [20].

Explainability approaches focus on helping users visualize how models make decisions. Saliency maps, Grad-CAM visualizations, feature importance rankings, and attention-based techniques highlight which portions of an input influence a prediction. These methods have been successful in computer vision applications, where visual explanations can provide insight into model behavior. Researchers have also explored explanation-first approaches, where explainability requirements are incorporated directly into the model design rather than added after training [23]. This helps align closely with the original goals of explainable AI, emphasizing that explanations should be useful to the people who rely on them.

One of the most influential efforts to advance explainability research was the DARPA Explainable AI (XAI) program, which ran from 2017 through 2021 [1]. The program brought together researchers from machine learning, human-computer interaction, and psychology to study not only how AI systems could generate explanations, but also how humans interpret and use those explanations. The initiative produced significant advances in explainable learning methods, causal reasoning models, visual explanation techniques, and evaluation frameworks, while also contributing tools such as the Explainable AI Toolkit (XAITK) to support future research [22]. Perhaps more importantly, the program helped establish explainability as a core requirement for trustworthy AI rather than an optional enhancement.

Traditional XAI methods have provided substantial value and remain highly effective for explaining classifiers, regression models, and other prediction-focused systems. They can often be applied after a model has been trained, generate explanations that are understandable to non-technical users, and have demonstrated success across a wide variety of real-world applications [10], [13]–[16]. However, these techniques were developed to explain individual predictions. As AI systems evolve beyond prediction engines, the limitations of traditional explainability approaches become increasingly apparent. Understanding why a model produces a prediction is no longer sufficient when the larger challenge is understanding why an autonomous agent selected an entire sequence of actions.

B. Agentic AI Systems

Agentic AI represents a paradigm shift from AI as a tool that processes inputs and returns outputs to an autonomous actor that pursues goals through sustained interaction with its environment. Khapre et al. [21] provide a comprehensive taxonomy that distinguishes many architectural paradigms: monolithic agents, modular agents, multi-agent systems, and tool-

TABLE I
COMPARISON OF TRADITIONAL XAI METHODS

Method	Approach	Scope	Strengths	Limitations
SHAP	Game theory / Shapley values	Global & local	Theoretically grounded; global + local	Computationally expensive; assumes feature independence
LIME	Local surrogate model	Local only	Fast; intuitive	Cannot capture nonlinear interactions; model-dependent
Saliency Maps	Gradient-based visualization	Local	Visual; intuitive	Neural networks only; no causal info
Attention Maps	Weight visualization	Local	Built into Transformers	May not reflect true reasoning

Sources: [3], [20], [23]

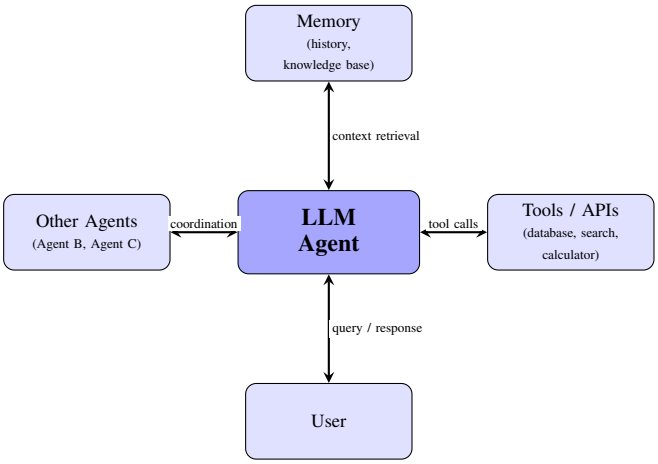

Fig. 2. High-level architecture of an agentic AI system. The central LLM Agent interacts bidirectionally with external tools, memory, other agents, and the user. Based on concepts from [21].

augmented systems. Each of these items introduces unique challenges for explainability.

ReAct (Reasoning and Acting) is a foundational framework for agentic AI that combines reasoning traces with actions. Rather than separating planning from execution, ReAct agents generate a thought, take an action, observe the result, and then generate the next thought based on the observation. ReAct agents continue the cycle until the task is complete. Yamaguchi et al. [8] build on the ReAct paradigm in their agentic XAI framework, applying it to iterative refinement of agricultural recommendations. Their work shows agentic approaches can enhance explanation quality but also entail important trade-offs; models lose 60–80% of their capabilities within 2–3 cycles due to over-optimization, revealing "a bias-variance trade-off analogy critical for trustworthy XAI systems" [8].

Tool-using agents extend the capabilities of LLMs by enabling them to invoke external tools: calculators, search engines, databases, APIs, and other software systems. This allows them to accomplish tasks that exceed their native abilities. This challenges explainability because the agent's output depends not only on its internal reasoning but also on the results returned by external tools, which may not provide adequate transparency. Veerapaneni et al. [4] address a related challenge with their Explainability-as-a-service (XaaS) approach, which decouples the explainability layer from core ML logic. Their framework demonstrates that explainability can be integrated into complex ML pipelines through containerized modules. These modules support dynamic auditing and real-time monitoring suitable for near-real-time interpretation [4].

Multi-agent frameworks such as LangGraph, AutoGen, and CrewAI enable the construction of systems where multiple specialized agents collaborate to accomplish complex tasks. Each agent has different capabilities and access to different tools. Ellouze et al. [12] demonstrate this approach in the context of autonomous vehicles, proposing a hybrid decision-making framework that combines Multi-Agent Reinforcement Learning (MARL) with XAI. In their architecture, each agent specializes in a sub-task, such as obstacle avoidance, trajectory planning, or intention prediction. SHAP-based XAI modules provide interpretable explanations of each agent's decisions [12]. Dua et al. [6] apply a similar approach to autonomous network orchestration, proposing a four-layer framework where LLM agents decompose high-level intents into domain rules. This creates an auditable trail that links every decision to its source, model version, and fallback options.

The main characteristic of agentic models is that they distinguish themselves from traditional prediction models such that a single answer may involve planning, retrieval, tool use, and response generation. Each of these tools constitutes a decision point that influences all other steps. Agent behavior is becoming more complex and autonomous, with individual interactions spanning dozens of steps, multiple model calls, and numerous external tool calls. These types of complexity render the traditional XAI method insufficient and demand new approaches to explainability.

C. Causal Traceability

Causal traceability is the conceptual heart of XAI 2.0 and is the fundamental shift in how we think about explaining AI systems. Instead of asking, "Why did the model output this answer?", causal traceability asks, "Why did the agent choose these actions?". This recognizes that in agentic systems, the output is not a single prediction but rather a sequence of decisions, each causally dependent on prior decisions and observations.

Schrills and Franke [17] provide observational grounding for the importance of traceability through their study of Subjective Information Processing Awareness (SIPA) in automated insulin delivery systems. Their research shows that users experience traceability, their understanding of how a system processes information, is "strongly correlated with trust and satisfaction with explanations" [17]. However, they also show

that the relationship between information disclosure and experienced traceability is not very straightforward. High levels of information disclosure can lead to "a miscalibration between SIPA and performance in predicting the system's results" [17], suggesting that traceability requires careful calibration.

To show causal traceability, consider a scenario where a user asks an agentic agent, "Who worked the most overtime last month?" A traditional prediction model would require this question to be pre-processed into a structured query. The agentic system would execute a sequence of decisions:

1) **Intent Classification:** The agent classifies the user query as a data retrieval request.
2) **Endpoint Selection:** The agent would select the appropriate API data source based on the available tools it has access to.
3) **Data Retrieval:** The agent formulates the appropriate API call and retrieves the records.
4) **Aggregation and Analysis:** The agent will process the returned data and perform aggregations as needed to answer the user's question.
5) **Response Generation:** The agent synthesizes its response and returns a natural-language response.

Rosenfeld and Richardson [7] provide a comprehensive taxonomy for explainability in human-agent systems that is relevant to causal traceability. Consider the fundamental questions: Why, Who, What, When, and How. They stress that the type of explainability needed "depends directly on the motivation for the type of human-agent system being implemented" [7]. Their framework identifies explanations geared toward different target audiences: regular users who need justifications, expert users who need technical details, and external entities such as regulators who need documentation. This multi-stakeholder perspective is essential for causal traceability, especially since different audiences require different levels of granularity in the decision chain.

Dazeley et al. [9] advance this thinking through their Causal XRL Framework (CXF) for explainable reinforcement learning. In their framework, they propose that explanations must operate at multiple levels of intentionality: zero-order (reactive) explanations address how input was interpreted; first-order (dispositional) explanations address the agent's current goals and intentions; second-order (social) explanations justify behavior based on agents' intentions; and N-th order (cultural) explanations explain how actions were modified based on expectations of other actors [9]. This multi-tier framework is essential for understanding causal traceability in agentic systems, where an agent's decisions may be influenced by factors at all of these levels at the same time.

The concept of decision provenance—maintaining a complete, auditable record of every decision, its inputs, its context, and its rationale—is central to causal traceability. Dua et al. [6] implement this concept in their Trustworthy Autonomous RAN Orchestration framework through an "Immutable Audit Trail" that records every deployment decision along with the original intent, derived constraints, validation results, trust assessments, and expected behavior. Their Decision Event Logger "kicks in and records everything that led to [each decision] including the original intent, the constraints that were derived from it, how validation went, the confidence signals from Layer 3, the validation and immutable content" [6]. This approach provides a concrete model for how causal traceability can be implemented in production autonomous systems.

D. Current Research Directions

Research on explainability for agentic AI is still in its early stages. However, several themes are beginning to emerge across the literature. The solutions proposed differ in how they are implemented, but they do share a common objective: to make autonomous decision-making observable, understandable, and auditable. Instead of focusing on individual model predictions, current research increasingly emphasizes the complete lifecycle of the agent's decision, including the planning and reasoning behind the agent and its tool usage.

One area of research focuses on agent execution logs and reasoning traces. Modern agentic systems generate detailed records of prompts, responses, tool use, and execution results. These artifacts provide valuable data for the reconstruction of how an autonomous agent arrived at its decision. Alicioglu and Sun [5] demonstrate this concept through BoP-X, a framework that analyzes the temporal behavior of reinforcement agents and produces human-readable summaries rather than isolated actions. Although developed for reinforcement learning, the underlying principle of explaining the behavior over time aligns with the goals of causal traceability for LLM agents.

Another important direction is the use of reasoning traces, particularly chain-of-thought (CoT), in understanding agent behavior. While CoT provides a natural record of reasoning, an important research question remains: do these explanations accurately represent the model's internal decision process? Yamaguchi et al. [8] explore this question through an iterative ReAct-based framework that alternates between reasoning and action. Their results suggest that explanation quality improves through refinement but eventually declines due to overfitting, highlighting the need to balance explanation depth and reasoning.

Researchers are also exploring methods for preserving decision provenance throughout an agent's execution. Instead of recording only the final prediction, decision provenance captures the sequence of events, the context, model version, any external resources, and the validation results that contributed to each decision. Veerapaneni et al. [4] and Dua et al. [6] demonstrate that maintaining immutable audit trails enables retrospective analysis, compliance, and greater confidence in the autonomous decision-making process. These systems move explainability beyond interpreting predictions toward documenting decision histories.

As autonomous AI systems become more distributed, explainability must also extend beyond individual agents. Multi-agent architecture introduces new challenges because the final outcomes involve collaboration among multiple specialized agents, each having its own objectives, reasoning processes, and access to tools. Khapre et al. [21] identify transparency,

behavioral traceability, and attribution across agents as fundamental requirements for an explainable multi-agent system. Explaining the behavior becomes as important as explaining the reasoning of any individual agent.

Human oversight still remains critical in research, particularly as regulatory frameworks continue to evolve. Explainability is valuable only if humans can understand it well enough to make decisions about when to trust an autonomous system and when to intervene in its processes. Damen et al. [18] discuss these requirements in the context of the EU AI Act, while Dua et al. [6] demonstrates practical mechanisms that allow engineers to review decisions that are considered high-risk before deployment. These approaches reinforce the idea that explainability should support human interactions rather than replace them.

Finally, regulatory initiatives are beginning to formalize many of these technical requirements. Frameworks such as the EU AI Act increasingly require transparency, accountability, and human oversight for high-risk AI systems. Roy et al. [23] note that these regulations emphasize explainability not as a feature, but as a necessity in establishing trustworthiness in AI. As agentic systems continue to evolve, expectations are likely to accelerate research into causal traceability, standardized auditing, and explanation frameworks.

III. The Unsolved Challenge: Explaining Multi-Step Autonomous Decisions at Scale

A. The Nature of the Challenge

The biggest challenge currently facing explainable AI for agentic systems is no longer understanding a single prediction. It is the challenge of understanding an entire decision process. Unlike traditional machine learning models that produce one output from one input, an agentic system generates responses through a sequence of planning, reasoning, tool use, memory retrieval, and intermediate decisions. Each of these steps influences the next, making it increasingly difficult to explain not only what happened but also why. As these systems continue to grow in complexity, existing XAI techniques struggle to provide explanations that capture the complete process.

One reason this problem is difficult is the number of decisions that can occur during a single interaction. A relatively simple request may require an agent to classify the user's intent, retrieve information from multiple sources, invoke several external tools, recover from errors, and synthesize the results into a final response. Each of these steps represents an independent decision that influences every step that follows. Alicioglu and Sun [5] demonstrate how quickly this complexity grows, showing that even reinforcement learning agents operating in constrained environments generate hundreds of observable behavior patterns over a relatively short period of execution.

The problem becomes even more difficult when multiple models and specialized agents collaborate within the same system and the same process. Modern agentic systems frequently distribute responsibilities across multiple components using different agents to perform the planning, reasoning, retrieval, or task-specific functions. When an unexpected outcome occurs, determining whether the problem originated with one agent or the interaction between other agents causes the greatest challenges. Khapre et al. [21] identify this diversity of architectures as one of the defining characteristics of agentic AI, while Ellouze et al. [12] demonstrate how multiple specialized agents cooperate within autonomous vehicle systems.

External tools introduce another layer of complexity. Unlike traditional prediction models, agentic systems routinely depend on databases, APIs, search engines, and custom-written code to complete a task. These resources influence the agent's decisions even though they operate outside the agent itself. A failed API endpoint call, an incomplete database query, or an outdated search result can alter the decision tree that is followed. While Veerapaneni et al. [4] address portions of this problem through their Explainability-as-a-Service architecture, explaining open-ended tool usage still remains an active area of research.

Persistent memory brings further complications to explainability because an agent's decisions are influenced not only by the current request, but also by information accumulated over historical interactions. Conversation history, retrieved documents, and contextual information all contribute to future decisions. This allows agents to behave more intelligently over time, but it also makes identifying the cause of a particular decision difficult. As Khapre et al. [21] observe, memory and continual learning are defining characteristics of modern agentic systems, yet they also introduce additional challenges for causal analysis.

Perhaps the most significant challenge is that even the developers who build these systems may struggle to explain why a particular sequence of decisions occurred. Stochastic model behavior, changing external information, persistent memory, and complex reasoning chains make many autonomous workflows difficult to reproduce exactly. Ronge et al. [2] argue that modern AI systems have reached a level of complexity where previous expectations of system understanding no longer apply. Similarly, Dazeley et al. [9] argue that explainability must evolve beyond interpreting individual predictions and instead provide coherent explanations for the behavior of multiple integrated intelligent systems.

These challenges illustrate why traditional explainability techniques are no longer sufficient for autonomous AI. Explaining a single prediction is fundamentally different from explaining an evolving sequence of decisions. Addressing that gap is the central motivation currently behind causal traceability and the foundation of the XAI 2.0 framework proposed in this paper.

B. Why Existing Approaches Fall Short

Traditional XAI methods were designed to explain the predictions, not the decision processes. Techniques such as SHAP and LIME explained how individual input features influence a model's output, but they provide no mechanisms for explaining how an autonomous agent arrives at its decision. Likewise, attention maps visualize relationships within

a single forward pass of a model, while saliency methods identify which portions of an input contributed to a prediction. These approaches are valuable when interpreting individual models, but they were never intended to explain the sequence of planning, memory retrieval, tool selection, and response generation that defines modern agentic AI.

This limitation becomes more apparent when considering the explainability of an entire workflow rather than a single prediction. Ronge et al. [2] describe several principles that underpin current XAI research, including computing edges, dimensionality reduction, and traceability. They also propose embedment and scientific testing as future directions for explainable AI. Existing methods have made progress in explaining which features influenced a prediction and in simplifying complex model behavior into interpretable representations. However, their notion of traceability remains focused on connecting individual inputs to individual outputs. Agentic AI introduces a broader challenge: understanding how one decision influences the next in the chain of actions.

This distinction is more than a technical limitation. It reflects a fundamental change in how explainability must be viewed. Yargan [19] argues that truly autonomous systems must also be accountable for the decisions they make, requiring explanations that are both reliable and understandable by humans. If an autonomous agent is expected to plan, reason, adapt, and act independently, then it must also be capable of explaining why it made the decisions the way it did. In that sense, explainability becomes an integral part of autonomous behavior rather than a feature added after the fact. This shift forms the foundation of the XAI 2.0 framework proposed in this paper.

C. Toward a Causal Traceability Framework

Explaining multi-step decision-making requires more than extending the current techniques. It requires a framework designed specifically for systems that plan, reason, retrieve information, and make sequences of connected decisions before producing a response. Based on the challenges identified throughout this review, I propose causal traceability as a framework for explaining how autonomous agents arrive at their decisions. Instead of replacing traditional XAI, causal traceability extends it to a complete decision process.

1) Decision Provenance Graphs: The core of the framework is the Decision Provenance Graph, which records the sequence of decisions made during an agent's execution. Each node of the graph represents an individual decision: intent, tool selection, data retrieval, or response generation. The connections between the nodes capture the dependencies that influence decisions. Instead of relying on isolated predictions, the graph will reconstruct the reasoning pathway leading to the final results. This idea builds off of traditional data while expanding it to include planning, behavior, reasoning, and interactions with any external systems.

2) Multi-Granularity Explanations: Different users require different levels of explanation. An end user may only need a summary of how the answer was produced, while an auditor may need to review the decision tree that led to the outcome. Developers often require even more detail, including the prompts, model responses, and the tool calls that created the workflow. A practical explainability framework should support each of these topics without requiring every user to interpret the same level of detail. This approach closely aligns with the stakeholder-centered explanation proposed by Rosenfeld and Richardson [7].

3) Temporal Consistency Tracking: Agentic AI makes decisions over time rather than in a single step. As new information is retrieved, tools return results, user interaction continues, and the internal state of the agent changes. Effective explainability should capture the different steps and identify where significant changes in reasoning occurred. Alicioglu and Sun [5] demonstrate a similar concept through temporal behavior analysis in reinforcement learning, providing a useful foundation for identifying important decision points within longer execution chains.

4) Counterfactual Tracing: Understanding why an agent selected one path is only part of the problem. It is often equally important to understand what would have happened if different decisions had been made. Counterfactual tracing extends explainability by allowing alternative execution paths to be examined. Questions such as "What if a different tool had been selected?" or "What if the retrieved data had been different?" provide valuable insight for debugging, auditing, and improving autonomous systems. As Ronge et al. [2] note, counterfactual reasoning has long been recognized as an important component of effective explanations, and its importance only increases as AI systems become more autonomous.

5) Trust Calibration: The final component of the framework focuses on trust. Detailed explainability should help users develop an appropriate level of confidence in an autonomous system rather than encouraging either blind acceptance or unnecessary skepticism. Schrills and Franke [17] demonstrate that users' perception of how well they understand a system is closely related to their trust in it. Building on this idea, causal traceability should communicate confidence levels throughout the decision process. It should also identify where uncertainty influenced individual decisions and indicate how sensitive the final outcome is to changes made earlier in the chain. This allows users to evaluate not only what the agent decided, but also the level of confidence they have in those decisions.

Together, these components form a framework that extends explainability from interpreting individual predictions to understanding the complete autonomous cycle. Decision provenance, multiple levels of explanation, temporal reasoning, and trust calibration each address a different aspect of explainability, but together they provide the foundation for explaining modern agentic AI systems. This is the main idea behind XAI 2.0.

IV. Applications and Implications

A. Case Study: Alzheimer's Disease Diagnosis

Alzheimer's disease (AD) diagnosis provides an ideal case study for evaluating causal traceability because it emphasizes many of the challenges that modern agentic AI systems should address. Diagnosing Alzheimer's is rarely based on a single test or a single model prediction. Clinicians combine information from neuroimaging, cognitive assessments, biomarker panels, genetic risk factors, and the patient's history before determining whether a patient is experiencing Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), or Alzheimer's disease. Every piece of evidence gathered contributes to the next stage of clinical reasoning, making the diagnostic process sequential rather than predictive. These characteristics make Alzheimer's diagnosis well-suited for evaluating explainability beyond model outputs.

The consequences of diagnostic errors further reinforce the need for transparent decision-making. A premature diagnosis may expose patients to unnecessary treatments while creating significant psychological and emotional distress. Delayed diagnosis may prevent patients from receiving the treatment they need during the earliest stages of disease progression. Because clinical decisions are formed by gathering evidence over time rather than evaluating a single observation, explainability must extend beyond identifying which features influenced the predictions. It is also important that the explanation of the reasoning that contributed to the diagnosis is clear and understandable.

1) The Limits of Current XAI in Alzheimer's Disease Diagnosis: Current explainability methods have improved our ability to understand individual AI predictions, but they explain only part of the clinical decision-making process. Vimbi et al. [14] reviewed 23 studies involving AI-assisted Alzheimer's disease diagnosis and found that nearly 70% relied on SHAP or LIME to explain model predictions. Across both deep learning approaches using MRI and PET imaging and traditional machine learning methods such as XGBoost, random forests, and support vector machines, the pattern is remarkably consistent. A model generates a diagnosis, and a post-hoc explainability technique identifies the features that contributed most strongly to that prediction.

These explanations can provide valuable insight. A clinician can see that hippocampal atrophy, cortical thinning, or white matter lesion burden influenced the model's decisions. However, they explain only a single stage of the overall diagnostic process.

Many questions remain unanswered. Did the model consider the patient's declining MMSE scores across previous visits? Did preprocessing introduce imaging artifacts that could have affected the prediction? How did conflicting biomarker evidence influence the final confidence score? Why was probable Alzheimer's disease selected over vascular dementia or Lewy body dementia? As Vimbi et al. [14] conclude, the black-box nature of modern machine learning and deep learning models remains a barrier to their adoption in clinical practices. Existing XAI techniques explain individual model predictions, but they do not explain the sequence of decisions that produce the diagnosis.

2) An Agentic Clinical Workflow: One possible implementation of an agentic diagnostic workflow is shown conceptually in Fig. 3. Although individual healthcare systems will differ, the overall reasoning process is representative of how autonomous clinical decision-support systems are likely to operate. Rather than relying on a single model, specialized agents collaborate to evaluate different aspects of the patient's condition before producing a final recommendation.

1) **Patient History Agent** retrieves and analyzes the longitudinal electronic health record, including previous diagnoses, medications, family history, and documented cognitive concerns. This information is used to establish an initial risk profile.
2) **Cognitive Assessment Agent** evaluates structured cognitive measures such as Mini-Mental State Examination (MMSE), Montreal Cognitive Assessment (MoCA), and Clinical Dementia Rating (CDR) in order to estimate disease progression in a patient across the EMCI, LMCI, and Alzheimer's continuum.
3) **Neuroimaging Agent** analyzes MRI or PET studies using deep learning models to identify structural and functional changes associated with Alzheimer's disease, including hippocampal atrophy, cortical volume loss, and amyloid-related abnormalities.
4) **Biomarker Agent** evaluates cerebrospinal fluid (CSF) measurements or blood-based biomarkers, including Aβ42, tau, and phosphorylated tau, while identifying laboratory values that fall outside the distribution observed during model training.
5) **Differential Diagnosis Agent** compares combined evidence against established diagnostic criteria such as the NIA-AA guidelines and DSM-5 to distinguish Alzheimer's disease from other dementias, including vascular dementia, Lewy body dementia, or alternative causes of cognitive impairment.
6) **Recommendation Agent** integrates evidence collected throughout the workflow to generate a clinical summary, disease stage, treatment recommendations, confidence estimates, and any conflicting evidence requiring physician review.

Traditional XAI techniques operate at the level of individual agents. For example, a saliency map may explain which brain regions influenced the Neuroimaging Agent, while SHAP values can identify the biomarkers that contribute to the Biomarker Agent's prediction. Causal traceability extends beyond these explanations by connecting every stage of the workflow into a single, auditable decision process.

When the model's prediction is probable early-stage Alzheimer's disease, clinicians should be able to trace how the patient's history influenced the cognitive assessment. Through the process, they should also be able to show how imaging findings influenced confidence in the biomarker analysis and how much confidence the system assigned to each metric.

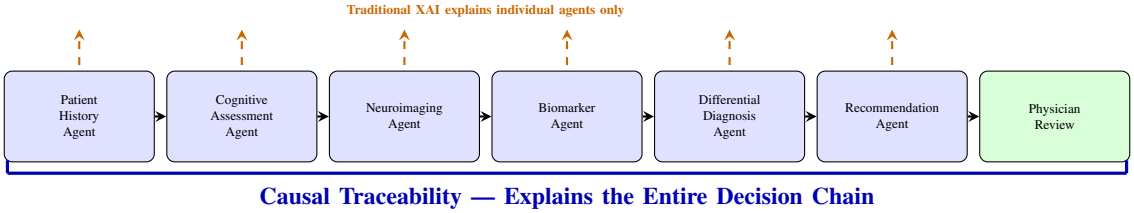

Fig. 3. Causal traceability chain for agentic Alzheimer's disease diagnosis. Traditional XAI (dashed orange arrows) explains individual agent outputs in isolation. Causal traceability (blue bracket) connects and audits every decision across the full clinical workflow. Based on [14], [21].

Instead of presenting an isolated prediction, causal traceability transforms the diagnostic process into a transparent record that can be reviewed, validated, and challenged when necessary.

3) Datasets for Empirical Evaluation: To properly evaluate causal traceability, datasets must capture multiple aspects of the diagnostic process rather than an isolated prediction task. Fortunately, several publicly available research datasets provide complementary sources of information that align well with the proposed agentic workflow.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) remains the gold-standard benchmark for Alzheimer's disease research, combining MRI, PET, cerebrospinal fluid biomarkers, genetic information, and longitudinal cognitive assessments collected from more than 2,000 participants. Because it contains multiple metrics collected over time, ADNI is very well-suited for evaluating temporal consistency, decision provenance, and interactions among multiple diagnostic agents.

OASIS-3 complements ADNI by providing decades of longitudinal imaging and cognitive assessment data across more than 1,300 participants. Its extended follow-up period makes it valuable for studying how diagnostic reasoning evolves as additional evidence becomes available over multiple years.

Although not Alzheimer's-specific, MIMIC-III and MIMIC-IV provide rich electronic health record data, including laboratory values, medications, physician notes, and structured clinical observations. These datasets are especially valuable for evaluating Patient History and Biomarker Agents and have already been used by Veerapaneni et al. [4] to demonstrate explainability within a modular Explainability-as-a-Service architecture.

The National Alzheimer's Coordinating Center (NACC) Uniform Data Set provides standardized cognitive evaluations collected across multiple Alzheimer's Disease Research Centers. Because patient assessments follow consistent clinical protocols, the dataset supports reproducible evaluation of cognitive assessment agents, enabling comparisons between institutions.

Together, these datasets provide complementary resources for evaluating each stage of an agentic diagnostic workflow while supporting future research into causal traceability for autonomous clinical agent-based support systems.

4) Why Alzheimer's Disease Requires Causal Traceability: Alzheimer's disease diagnosis illustrates why causal traceability is becoming a practical necessity rather than simply an interesting research direction. Clinical decisions are formed by integrating multiple sources of evidence collected over time, making it difficult to justify a recommendation using only feature importance scores or saliency maps. Physicians remain responsible for the final diagnosis and must be able to understand not only what recommendation an AI system produced, but also how that recommendation was reached.

TABLE II
KEY DATASETS FOR ALZHEIMER'S DISEASE CAUSAL TRACEABILITY RESEARCH

Dataset	**Data Types**	**Scale**	**Pipeline Relevance**
ADNI	MRI, PET, CSF, genetics, cognitive scores	2,000+ participants	All six agents; temporal consistency tracking
OASIS-3	MRI, PET, cognitive, clinical	1,300+ participants; 30-yr follow-up	Provenance tracing over extended timelines
MIMIC-III/IV	EHR, labs, notes, medications	46,520+ patients	Patient History & Biomarker Agents [4]
NACC UDS	Cognitive batteries, neuropathology	45,000+ subjects	Standardized cross-site cognitive evaluation

Sources: [4], [14]

Various regulatory requirements highlight the need for transparency. HIPAA requires that the clinic's documentation support auditing and review. The EU AI Act classifies AI-assisted medical diagnosis as a potential high-risk application that requires transparency and human oversight [18]. These requirements are in alignment with the objectives of causal traceability by ensuring complete documentation of the decision process rather than explanation of a single agent's predictions. Veerapaneni et al. [4] further show that this level of explainability is computationally feasible, achieving latencies between 120 and 180 milliseconds while maintaining HIPAA-compliant audit trails.

Alzheimer's disease represents more than an application area for XAI 2.0. It provides a realistic environment in which the strengths and limitations of causal traceability can be evaluated. The combination of longitudinal clinical data, multiple diagnostic methods, and increasing regulatory over-

sight makes Alzheimer's disease an appropriate benchmark for assessing how explainability should evolve as AI systems become increasingly autonomous.

B. Autonomous Systems and Robotics

Autonomous vehicles represent an important area where causal traceability is essential for both safety and regulatory compliance. Tahir et al. [11] propose a hybrid LIME-SHAP integration for real-time interpretability in autonomous vehicles. They have achieved fidelity rates above 85% and interpretability factors above 80%. Ellouze et al. [12] extend this to multi-agent settings where separate agents handle different aspects of the driving decision. However, even these approaches explain individual perception or control decisions rather than full driving episodes. Causal traceability would enable the complete reconstruction of the decision chain instead of individual visibility, such as: "The vehicle slowed down because the perception agent detected a pedestrian, the prediction agent estimated their trajectory as crossing the road, and the planning agent decided to slow down rather than change lanes because the adjacent lane was occupied." Alicioglu and Sun's [5] BoP-X framework shows how pattern-based approaches can summarize overall agent strategies.

C. Financial Services and Regulatory Compliance

Financial applications that rely on AI require stringent explainability requirements due to regulatory requirements. Klein and Walther [16] document the growing importance of XAI in the financial world. They note its application in financial risk management, creditworthiness, financial trading, and regulatory compliance. Causal traceability is relevant for AI systems that perform complex financial analysis involving multiple data sources, models, and decision-making steps. Damen et al. [18] provide a roadmap for using XAI in EU AI Act compliance audits, stating that "XAI enables traceability and accountability" and can help auditors know if AI systems meet the Act's requirements for transparency, fairness, and human oversight. Their framework implies that causal traceability would directly support regulatory compliance by providing the "clear, reliable, and actionable" explanation required to satisfy internal auditors.

D. Network Orchestration and Critical Infrastructure

Dua et al. [6] present perhaps the most fully realized implementation of causal traceability principles in their Trustworthy Autonomous RAN Orchestration framework. Their four-layer architecture—Intent Translation, Zero-Trust Validation, Federated Trust Coordination, and Auditable Orchestration—demonstrates how causal traceability can be built into autonomous systems from the ground up. Their implementation includes a Policy Compliance Orchestrator that computes explicit trust scores, a Decision Event Logger that creates immutable audit trails, and a Policy Execution Monitor that compares observed behavior against expected behavior. This work provides a concrete blueprint for extending causal traceability to other critical infrastructure domains.

V. Discussion and Future Directions

A. Bridging the Gap Between XAI 1.0 and XAI 2.0

Moving from the traditional XAI to causal traceability for agentic systems is not a replacement, but an extension. Methods like SHAP and LIME remain valuable for explaining individual decisions within an agent's workflow. The challenge is to embed these explanations within a more robust framework that will capture the causal structure that connects all the agents. Veerapaneni et al.'s [4] modular XaaS architecture offers a promising design pattern for this integration, where a modular explanation engine can be selected dynamically based on the type of model and data for each decision node, while the orchestration framework maintains traceability throughout the entire workflow.

B. Standardization and Evaluation

There are gaps in the current research landscape due to the lack of standardized metrics and benchmarks for evaluating the quality of causal traceability explanations. The DARPA XAI program developed evaluation measures spanning functionality, performance, and the explanation of effectiveness [1], but these were designed for traditional prediction-explanation systems. New metrics are needed that will assess the completeness of causal chains. These metrics should demonstrate transparency, faithful explanations, and the scalability of traceability mechanisms.

C. Privacy and Security Considerations

Causal traceability creates tension with privacy and security requirements. Complete decision logs may contain sensitive information about data subjects, proprietary model details, or security-relevant system configurations. Dua et al. [6] address this partially through their federated trust coordination layer, which uses differential privacy and reputation-weighted aggregation to enable multi-vendor transparency without exposing proprietary models or raw data. Future work must develop principled frameworks for balancing the transparency requirements of causal traceability with the confidentiality requirements of data protection regulations.

D. The Role of Regulation

The EU AI Act and similar regulatory frameworks are creating legal mandates for explainability that extend beyond what current technology can deliver. Damen et al. [18] identify several areas where XAI has limitations in supporting compliance. Methods such as SHAP and LIME may "oversimplify complex models", explanations may be inconsistent across reporting platforms, and the effectiveness of XAI varies by model type and application context. These limitations are amplified for agentic systems, where the explanatory target is not a single prediction but an entire decision workflow. Frameworks will need to evolve alongside technical capabilities, establishing requirements that are ambitious enough to drive innovations but realistic enough to be executed in the real world.

TABLE III
TRADITIONAL XAI (1.0) VS. CAUSAL TRACEABILITY (XAI 2.0)

Dimension	**XAI 1.0 (Traditional)**	**XAI 2.0 (Causal Traceability)**
Core Question	"Why did the model produce this output?"	"Why did the agent choose this chain of actions?"
Scope	Single prediction	Multi-step autonomous workflow
Target System	Classifiers, regression models	Autonomous agents, multi-agent systems
Key Methods	SHAP, LIME, saliency maps, attention maps	Decision provenance graphs, CoT auditing, counterfactual tracing
Explanation Unit	Feature importance scores	Causal decision chains
Temporal Awareness	None (static snapshot)	Tracks decisions and state changes over time
Tool/External Awareness	None	Traces tool calls and external data influence
Stakeholder Support	Primarily developers	Users, auditors, regulators, developers

VI. CONCLUSION

The rapid evolution of AI from traditional prediction models to autonomous agentic systems has fundamentally changed the nature of explainability. Traditional XAI methods such as SHAP, LIME, attention maps, and saliency techniques remain valuable tools. However, they were developed for systems that produce a single output from a single inference. Agentic AI introduces a fundamentally different type of challenge, where responses emerge through planning, reasoning, memory, tool use, and a sequence of interconnected decisions that cannot be adequately explained by current methods alone.

This paper has argued that explainability must evolve alongside these new architectures. Rather than asking "Why did the model produce this prediction?" the more important question for autonomous systems becomes "Why did the agent choose this sequence of actions?" That shift changes the objective of explainability from interpreting individual predictions to understanding the full decision-making process. To address this challenge, I introduced the concept of XAI 2.0, a framework centered on causal traceability that extends traditional explainability through decision provenance, temporal reasoning, multi-level explanations, and trust calibration.

The Alzheimer's disease case study demonstrates why this shift is necessary. Clinical decision making depends on integrating information collected over time from multiple sources rather than relying on a single prediction. Similar challenges are emerging in autonomous vehicles, financial systems, cybersecurity, industrial automation, and many other domains where autonomous agents increasingly make decisions that directly affect people. As these systems become more capable, understanding how they arrive at their conclusions will become just as important as the conclusions themselves.

The next generation of explainability research will require advances in causal inference, decision provenance, human-centered explanation design, and scalable auditing techniques. Regulatory frameworks will continue to evolve alongside these technical advances, but regulation alone cannot solve the explainability problem. Trustworthy autonomous AI ultimately depends on our ability to reconstruct, understand, and evaluate the complete sequence of decisions that produced an outcome.

Traditional XAI explains predictions. Autonomous AI requires us to explain the decision process itself. That distinction represents the central contribution of this work and the motivation behind the proposed XAI 2.0 framework. As AI systems evolve from predictive models into autonomous multi-agent architectures, causal traceability provides a practical foundation for making them transparent, accountable, and worthy of the trust placed in them.

REFERENCES

[1] D. Gunning, "DARPA's explainable AI (XAI) program: A retrospective," *Applied AI Letters*, vol. 2, no. e61, pp. 1–11, 2021. doi: 10.1002/ail2.61

[2] R. Ronge, B. Bauer, and B. Rathgeber, "Approaching Principles of XAI: A Systematization," *IEEE Trans. Artif. Intell.*, vol. 6, no. 5, pp. 1067–1082, May 2025. doi: 10.1109/TAI.2024.3515937

[3] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, K. Lekadir, S. E. Petersen, and G. Menegaz, "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Adv. Intell. Syst.*, vol. 7, no. 2400304, pp. 1–8, 2025. doi: 10.1002/aisy.202400304

[4] S. M. Veerapaneni, K. Nath, P. Kumar, and R. Tewari, "Explainability-as-a-Service: Modular XAI Layers for Regulated ML Pipelines," in *Proc. 2nd Asia Pacific Conf. Innovation Technol. (APCIT)*, Mysore, India, Sep. 2025, pp. 1–7. doi: 10.1109/APCIT65661.2025.1141166

[5] G. Alicioglu and B. Sun, "BoP-X: A New XAI Approach to Explain Reinforcement Learning for Autonomous Driving," in *Proc. 2025 Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2025, pp. 515–521. doi: 10.1109/ICMLA66185.2025.00076

[6] M. Dua, T. Kundu, and A. Gupta, "Trustworthy Autonomous RAN Orchestration: An Architectural Framework for Zero-Trust Intent-Based Control with Explainable AI," in *Proc. 2026 35th Wireless Opt. Commun. Conf. (WOCC)*, 2026, pp. 1–6. doi: 10.1109/WOCC69802.2026.1155617

[7] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Auton. Agents Multi-Agent Syst.*, vol. 33, pp. 673–705, 2019. doi: 10.1007/s10458-019-09408-y

[8] T. Yamaguchi, Y. Zhou, M. Ryo, and K. Katsura, "Agentic Explainable Artificial Intelligence (Agentic XAI) Approach To Explore Better Explanation," *arXiv preprint*, arXiv:2512.21066v3, 2025.

[9] R. Dazeley, P. Vamplew, and F. Cruz, "Explainable reinforcement learning for broad-XAI: A conceptual framework and survey," *Neural Comput. Appl.*, vol. 35, pp. 16893–16916, 2023. doi: 10.1007/s00521-023-08423-1

[10] A. Turki, "Automated framework for multi-domain social media text analysis for business strategy employing multilayer perceptron with Word2Vec features and LIME XAI," *PLoS One*, vol. 20, no. 11, e0336240, Nov. 2025. doi: 10.1371/journal.pone.0336240

[11] H. A. Tahir, W. Alayed, W. U. Hassan, and A. Haider, "A Novel Hybrid XAI Solution for Autonomous Vehicles: Real-Time Interpretability Through LIME–SHAP Integration," *Sensors*, vol. 24, no. 6776, pp. 1–25, Oct. 2024. doi: 10.3390/s24216776

[12] A. Ellouze, M. Karray, and M. Ksantini, "A Hybrid Decision-Making Framework for Autonomous Vehicles in Urban Environments Based on Multi-Agent Reinforcement Learning with Explainable AI," *Vehicles*, vol. 8, no. 8, pp. 1–19, Jan. 2026. doi: 10.3390/vehicles8010008

[13] A. M. Khan, M. A. Tariq, S. K. U. Rehman, T. Saeed, F. K. Alqahtani, and M. Sherif, "BIM Integration with XAI Using LIME and MOO for Automated Green Building Energy Performance Analysis," *Energies*, vol. 17, no. 3295, pp. 1–36, Jul. 2024. doi: 10.3390/en17133295

[14] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: A systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Inform.*, vol. 11, no. 10, pp. 1–29, 2024. doi: 10.1186/s40708-024-00222-1

[15] K. Soni, R. Frew, and B. Kebede, "Interpretable machine learning for origin classification of Brazilian soybeans: A random forest and XAI-based approach," *J. Food Compos. Anal.*, vol. 150, no. 108896, pp. 1–12, 2026. doi: 10.1016/j.jfca.2026.108896

[16] T. Klein and T. Walther, "Advances in Explainable Artificial Intelligence (xAI) in Finance," *Finance Res. Lett.*, vol. 70, no. 106358, pp. 1–3, Nov. 2024. doi: 10.1016/j.frl.2024.106358

[17] T. Schrills and T. Franke, "How Do Users Experience Traceability of AI Systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) Systems," *ACM Trans. Interact. Intell. Syst.*, vol. 13, no. 4, Art. 25, pp. 1–34, Dec. 2023. doi: 10.1145/3588594

[18] V. Damen, M. Wiersma, G. Aydin, and R. van Haasteren, "Explainable AI for EU AI Act compliance audits," *Maandbl. Accountancy Bedrijfsecon.*, vol. 99, no. 4, pp. 231–242, Sep. 2025. doi: 10.5117/mab.99.150303

[19] D. Yargan, "The Machine as an Autonomous Explanatory Agent," *Felsefe Dünyası*, no. 79, pp. 265–279, 2024. doi: 10.58634/felsefedunyasi.1487376

[20] D. Li, Y. Du, G. Huang, and Q. Zhang, "An XAI-based framework for GNSS observation data anomaly detection in constrained observation environments," *GPS Solutions*, vol. 30, no. 120, pp. 1–16, 2026. doi: 10.1007/s10291-026-02075-z

[21] S. Khapre, M. A. Mersha, Z. Olalekan, H. Shakil, S. Iskander, A. Moustafa, and J. Kalita, "Agentic AI systems: A systematic survey of multi-agent architectures, cognitive foundations, interaction, explainability, security, and performance evaluation," *Neurocomputing*, vol. 696, no. 134049, pp. 1–25, 2026. doi: 10.1016/j.neucom.2026.134049

[22] B. Hu, P. Tunison, B. Vasu, N. Menon, R. Collins, and A. Hoogs, "XAITK: The explainable AI toolkit," *Applied AI Letters*, vol. 2, no. e40, pp. 1–10, 2021. doi: 10.1002/ail2.40

[23] A. Roy, K. D. Gaikwad, A. J. Hude, J. S. Shinde, J. Parekh, and H. R. Nagrani, "Explainable AI (XAI) for Object Detection and Application to Satellite Imagery," *IEEE Access*, vol. 13, pp. 185597–185618, 2025. doi: 10.1109/ACCESS.2025.3624198