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

Steven Tidwell
East Texas A&M
Email: 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, retrieving 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 of 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 agent 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 their area of expertise

Keywords—Explainable AI, Agentic AI, Casual Traceability, Trustworthy AI, Multi-Agent Systems, Autonomous Systems

I. INTRODUCTION

A. What is Trusted AI?

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.

FIG 1. PILLARS OF TRUSTED AI

TRUSTED AI				
Reliability	Transparency	Accountability	Safety	Human Oversight
EXPLAINABILITY				

B. Why Explainability Matters

Traditional AI systems, specially 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 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 that are used to evaluate credit risk, detect fraudulent activity, and support investment decisions, making it essential that organizations can justify how those decisions are reached. Regulatory frameworks such as GDRP 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], [14], [15], [16]. However, these techniques were developed to explain individual predictions. As AI systems evolve from 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.

TABLE I

Comparison of Traditional XAI Methods				
Method	**Approach**	**Scope**	**Strengths**	**Limitations**
SHAP	Game theory-based	Global + Local	Strong theoretical foundation	Computationally expensive
LIME	Local surrogate methods	Local	Fast; intuitive explanations	Cannot capture nonlinear interactions
Saliency Maps	Gradient-based visualization	Local	Visual and intuitive	Limited to neural networks
Attention Maps	Attention weight visualization	Local	Built into transformers	Attention does not necessarily reflect reasoning

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

B. Agentic AI Systems

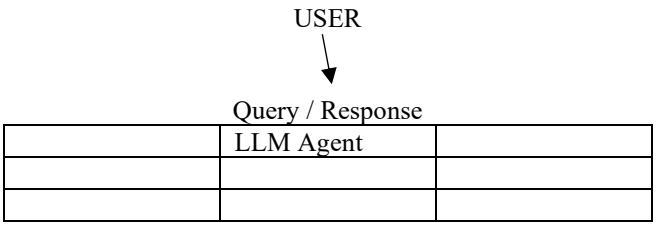
Agentic AI represents a paradigm shift from AI as a tool that processes inputs and returns outputs to a tool that becomes an autonomous actor that pursues goals through sustained interaction with its environment. Khapre et al. [21] provide a comprehensive taxonomy that distinguishes many architectural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

ReAct (Reasoning and Acting) is a foundational item 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 the 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 explainability. 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 would have 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.

FIG 2



C. Causal Traceability

If you want, you can have subsections for the main topics in the background information.

D. Current Research Directions

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

After the text edit has been completed, the report is ready for the template. Duplicate the template file by using the Save As command. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your report; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your report. There are two types: component heads and text heads.

Component heads identify the different components of your report and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the report title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named "Heading 1", "Heading 2", "Heading 3", and "Heading 4" are prescribed.

B. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I. TABLE TYPE STYLES

Table Head	Table Column Head		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a.] Sample of a Table footnote. (*Table footnote*)

Fig. 1. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT *(Heading 5)*

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

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 Transactions on Artificial Intelligence, 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," Advanced Intelligent Systems, 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 in Technology (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. Machine Learning and Applications (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 and Optical Communications Conf. (WOCC), 2026, pp. 1–6. doi: 10.1109/WOCC69802.2026.1155617

[7] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," Autonomous Agents and Multi-Agent Systems, 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 Computing and Applications, 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 Informatics, 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," Journal of Food Composition and Analysis, 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 Research Letters, 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, Article 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," Maandblad voor Accountancy en Bedrijfseconomie, 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