wjert, 2025, Vol. 11 Issue 1, 60-88 *Review Article* ISSN 2454-695X

World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 7.029

EXPLAINABILITY IN AGENTIC AI: A CONCEPTUAL FRAMEWORK FOR BALANCING INTERPRETABILITY, SECURITY, AND PERFORMANCE

Siddharth Nandagopal[1*]

[1]Cambridge, Massachusetts 02139. USA. Unaffiliated/Independent.

Article Received on 09/11/2024 **Article Revised on 30/11/2024** **Article Accepted on 20/12/2024**

***Corresponding Author**
Siddharth Nandagopal
Cambridge, Massachusetts
02139. USA.
Unaffiliated/Independent.

ABSTRACT

This paper introduces the Secure-Explainable Agentic AI (SEAAI) framework to address critical trade-offs between explainability, security, and performance in agentic AI systems. The proposed conceptual model combines theoretical constructs, metrics, and game-theoretic approaches to achieve balanced outcomes across diverse application domains. By integrating explainability compliance measures, security risk assessments, interpretability scoring methods, performance trade-off calculations, and adaptive explainability mechanisms, the framework guides decision-makers in navigating complex design choices. Through conceptual analysis and qualitative case studies, the work establishes a robust foundation that accounts for adversarial influences, regulatory demands, and user expectations. The framework's game-theoretic perspective incorporates defenders and attackers into a multi-stakeholder environment, ensuring dynamic resilience against evolving threats. The result is a flexible and generalizable reference point for aligning ethical standards and legal requirements with operational goals. Ultimately, this research fosters better understanding of how to configure explainability without compromising security or performance. Such insights empower developers, regulators, and users to embrace agentic AI responsibly. Future efforts can refine and extend the SEAAI framework, encouraging interdisciplinary collaboration and continuous improvement. In doing so, the journey toward trustworthy, interpretable, and safe autonomous systems advances, paving the way for widespread real-world adoption.

KEYWORDS: Agentic AI, Explainability, Security, Performance, Multi-Stakeholder, Trade-offs.

INTRODUCTION

Background and Motivation

Agentic artificial intelligence (AI) refers to systems that exhibit autonomy in decision-making and interaction, impacting both physical and digital environments (Chan et al., 2023; Kenton et al., 2023; Lieberman, 1997; Liu et al., 2023; Ruan et al., 2023; Shavit et al., 2023; Sumers et al., 2023). Applications of such AI include planning tasks, automating research, and executing administrative processes, as seen in tools that manage vacation planning, email communications, and even financial portfolios (Bran et al., 2023; Chan et al., 2023; Ruan et al., 2023; Schick et al., 2023). The rapid adoption of Agentic AI is evident across domains like healthcare, where it assists in diagnostics, education through adaptive learning systems, and retail for personalized recommendations (Chan et al., 2023; Nigon et al., 2024).

As these applications grow, the demand for explainability becomes increasingly critical. Explainability ensures users and regulators can comprehend decisions made by AI, aligning with ethical, legal, and operational standards (Díaz-Rodríguez et al., 2023). In high-stakes sectors, like autonomous healthcare diagnostics, legal decision-making, education, the capacity to understand and validate an AI’s reasoning process is paramount for trust and accountability (Chan et al., 2023; Duan et al., 2024; Nigon et al., 2024; Saarela et al., 2021). This rising necessity underscores a crucial trade-off between transparency and the complexity of the AI systems being deployed (Pillai, 2024).

Challenges in Balancing Explainability and Security

Explainability introduces significant security challenges. Revealing sensitive mechanisms of AI models can expose them to adversarial vulnerabilities, as attackers may exploit these insights to manipulate outputs (Akhtar, Kumar & Nayyar, 2024; Kuppa & Le-Khac, 2020). For instance, adversaries might deceive AI models in fraud detection or manipulate decision-making in automated trading systems or deception in healthcare diagnostics (Baniecki & Biecek, 2024; Park et al., 2024; Zbrzezny & Grzybowski, 2023).

Overreliance on AI agents introduces significant challenges in ensuring a balance between explainability and security, particularly in critical areas such as finance and legal systems (Akhtar, Kumar & Nayyar, 2024; Kuppa & Le-Khac, 2020). AI agents are increasingly entrusted with tasks traditionally performed by humans, such as hiring decisions and medical diagnoses, despite the potential for design flaws or malfunctions. Such malfunctions can trigger cascading failures, resulting in widespread disruptions to public services and essential infrastructure. The complexity of these systems often obscures immediate detection of errors, further exacerbating their impact (Akhtar, Kumar & Nayyar, 2024; Kuppa & Le-Khac, 2020; Li et al., 2021).

Another pressing concern is the potential misuse of AI agents for harmful purposes. Autonomous systems designed for scientific experimentation or resource planning could be exploited to develop dangerous tools, such as bioweapons, by entities lacking conventional expertise (Rose & Nelson, 2023). Additionally, highly persuasive AI agents may facilitate the spread of misinformation or support targeted influence campaigns, complicating efforts to regulate and monitor their activities (Hajli et al., 2022).

The delayed and diffuse nature of certain risks further complicates governance. AI agents tasked with long-term objectives, like optimizing hiring processes or managing supply chains, may embed systemic biases or inefficiencies that become entrenched over time. For example, biases in algorithmic hiring processes can perpetuate inequalities that are difficult to reverse after widespread implementation. Similarly, agents managing human communication or social media platforms could contribute to significant societal and psychological impacts, mirroring the challenges associated with large-scale digital platforms (Fiske, Henningsen & Buyx, 2019; Hajli et al., 2022; Jabarian, 2024; Li et al., 2021).

The interconnected nature of multi-agent systems also heightens risks. Systems relying on multiple agents, such as automated trading platforms, can create destabilizing feedback loops, as exemplified by the 2010 flash crash. Shared foundational components across agents may amplify vulnerabilities, potentially leading to systemic failures across multiple domains. Moreover, the capability of AI agents to create specialized sub-agents compounds the issue, as monitoring and controlling these sub-agents becomes increasingly challenging, raising the likelihood of unintended consequences (CFTC & SEC, 2010; Dorri, Kanhere & Jurdak, 2018).

Addressing these challenges requires carefully designed frameworks that ensure explainability does not expose vulnerabilities while simultaneously preserving the security and robustness of AI agents (Akhtar, Kumar & Nayyar, 2024; Kuppa & Le-Khac, 2020).

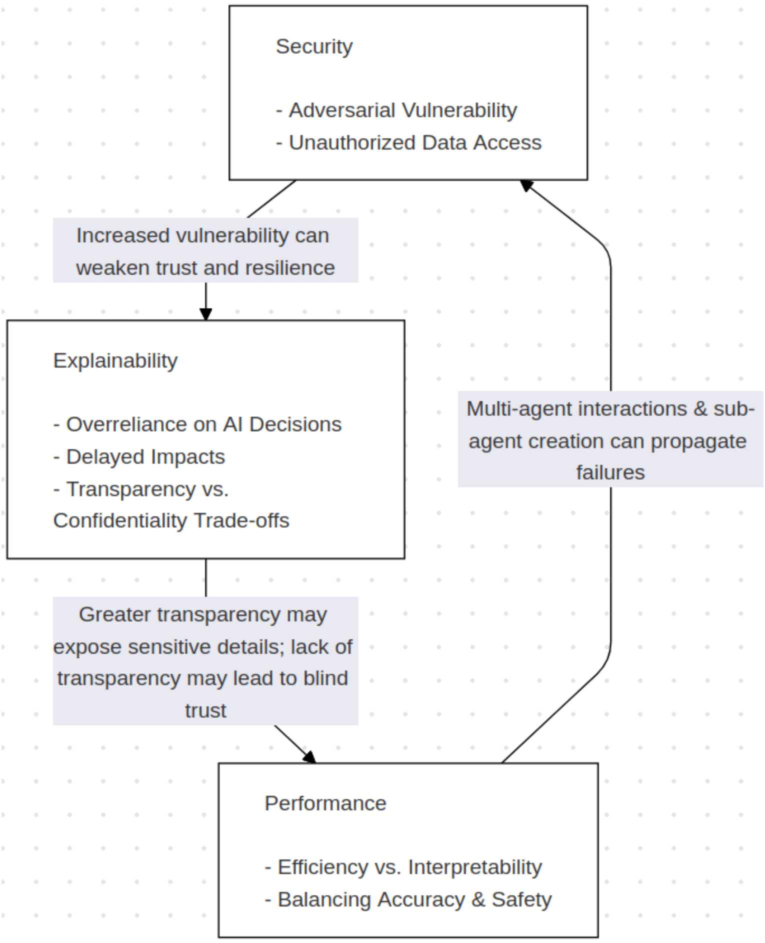
Security
- Adversarial Vulnerability
- Unauthorized Data Access
Increased vulnerability can weaken trust and resilience
Explainability
- Overreliance on AI Decisions
- Delayed Impacts
- Transparency vs. Confidentiality Trade-offs
Multi-agent interactions & sub-agent creation can propagate failures
Greater transparency may expose sensitive details; lack of transparency may lead to blind trust
Performance
- Efficiency vs. Interpretability
- Balancing Accuracy & Safety

Figure 1: Illustrates the interdependencies between identified challenges, emphasizing the trade-offs inherent in balancing explainability and security in Agentic AI systems.

Research Problem and Objectives

The effective governance of Agentic AI systems requires explainability, which is defined as the ability to provide evidence or reasoning for system outputs. Explainability must adhere to four principles: delivering explanations, ensuring meaningfulness for intended users, maintaining accuracy in reflecting system reasoning, and operating within defined knowledge limits to ensure confidence (Phillips et al., 2021). These principles are critical for aligning interpretability with security and performance in applications across finance, healthcare, and education.

Balancing security and explainability presents a dual challenge. While transparency builds trust and accountability, it can also expose vulnerabilities to adversarial attacks or compromise operational efficiency (Akhtar, Kumar & Nayyar, 2024; Kuppa & Le-Khac, 2020). Addressing this challenge requires a robust theoretical framework that integrates conceptual analysis with qualitative case studies. This framework will propose practical and generalizable solutions for ensuring accountability, mitigating risks, and optimizing the trade-offs between interpretability, security, and performance.

LITERATURE REVIEW

Explainability in AI: Definition and Methods

Explainability in AI refers to the ability of a system to provide clear and understandable reasons for its outputs or decisions. It plays a critical role in ensuring transparency and trust, particularly in sensitive domains like healthcare, finance, and education, where decisions can have significant consequences (Phillips et al., 2021). As depicted in Figure 2, Explainability techniques are broadly categorized into **global** and **local** methods. Global techniques focus on explaining the overall model behavior, identifying generic operating rules, while local techniques explain individual predictions, detailing how the model arrived at a specific output (Huber et al., 2021).

Some prominent explainability techniques include

1. **SHAP (SHapley Additive exPlanations):** Explains individual predictions by assigning importance values to input features.
2. **LIME (Local Interpretable Model-agnostic Explanations):** Provides approximate explanations for local predictions by perturbing the input and observing output changes.
3. **Permutation Importance:** Measures feature importance by analyzing the impact of shuffling feature values on model performance.
4. **Partial Dependence Plot (PDP):** Visualizes the relationship between a feature and the predicted outcome while holding other variables constant.
5. **Integrated Gradients:** Computes feature attributions for deep learning models by integrating gradients along the input path.
6. **Tree Surrogates:** Uses decision trees to approximate and interpret complex models.

Explainability techniques address different levels of detail. For instance, **global methods** like PDP or Tree Surrogates provide insights into model-wide behavior, while **local methods** like SHAP or LIME explain individual predictions (Dwivedi et al., 2023). These methods are critical for applications such as fraud detection, where transparency helps in auditing decisions, and healthcare, where clinicians need to trust diagnostic recommendations (Phillips et al., 2021). Additionally, explainability is essential for meeting governance and regulatory needs, ensuring accountability and compliance in AI-driven systems (Phillips et al., 2021).

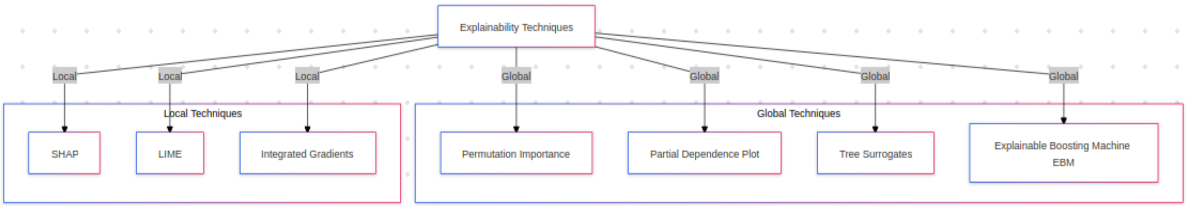

Figure 2 categorizes prominent explainability techniques into global and local levels.

Security Concerns in XAI Systems

While explainability enhances transparency, it also introduces security vulnerabilities (Akhtar, Kumar & Nayyar, 2024; Das & Rad, 2020; Kuppa & Le-Khac, 2020). Adversarial attacks exploit the transparency provided by explainability techniques, manipulating inputs to deceive the model and produce incorrect outputs. For instance, attackers can use SHAP values to identify critical features and craft inputs that bypass fraud detection systems (Wagle, 2021).

Explainability mechanisms may also leak sensitive information. For example, revealing feature importance in medical datasets could compromise the privacy of individuals (Ezzeddine, 2024; Hulsen, 2023), violating privacy regulations like GDPR. Techniques such as **differential privacy computations** mitigate these risks by ensuring that aggregated insights are shared without revealing individual data points (Abadi et al., 2016). These methods strike a balance between transparency and privacy, enabling the use of explainable AI in sensitive applications like patient diagnostics or financial audits.

The dual challenge lies in maintaining explainability while ensuring security (Akhtar, Kumar & Nayyar, 2024; Das & Rad, 2020; Kuppa & Le-Khac, 2020). Autonomous systems used in scientific research or resource planning, if exploited, could facilitate harmful activities such as bioweapon development or market manipulation (Rose & Nelson, 2023). These risks highlight the need for robust mechanisms that integrate explainability with security measures to prevent unintended misuse.

Existing Approaches and Gaps

Several frameworks have been developed to address the balance between explainability and security in AI systems. Privacy-preserving machine learning models and explainable boosting machines (EBMs) are examples of systems that aim to provide transparency while maintaining robust security. However, these frameworks often lack scalability and adaptability, limiting their effectiveness in diverse and dynamic environments (Dwivedi et al., 2023; Xu, Baracaldo, & Joshi, 2021).

One major gap is the absence of mechanisms to modulate access to data. Current systems struggle to provide varying levels of granularity in data logs, which are essential for regulatory oversight. For example, regulators may require detailed logs for investigations, while only aggregated data may be necessary for general audits. The inability to control data granularity creates challenges in balancing transparency with privacy (de Santana et al., 2023; Elkhawaga, Abu-Elkheir & Reichert, 2022; European Commission's Artificial Intelligence Act, 2024).

Another critical limitation is the lack of tools to track sub-agent creation. In multi-agent systems, agents often delegate tasks to sub-agents, which can operate independently and create cascading risks. Without proper monitoring, these sub-agents can introduce vulnerabilities, leading to failures that are difficult to trace or mitigate. Developing mechanisms to track and control sub-agent activities is essential for enhancing the robustness and reliability of explainable AI systems (CFTC & SEC, 2010; Dorri, Kanhere & Jurdak, 2018).

METHODOLOGY

The development of the proposed framework relies on a methodological approach, as depicted in Figure 3, that combines conceptual analysis with qualitative case study examination. This dual approach ensures a comprehensive understanding of how critical constructs interact in real-world applications, providing a robust basis for balancing explainability, security, and performance in Agentic AI systems.

Conceptual Analysis

The foundation of the methodology lies in identifying and analyzing key constructs such as security, interpretability, and performance. Security refers to the resilience of AI systems against adversarial attacks, data leakage, and malicious misuse (Moskalenko et al., 2023). Interpretability involves the ability to explain model decisions to stakeholders (Erasmus, Brunet, & Fisher, 2021), while performance encompasses the accuracy, efficiency, and scalability of AI models (Aggarwal & Liu, 2023). These constructs are interrelated and often involve trade-offs. For instance, increasing interpretability might expose vulnerabilities (Akhtar, Kumar & Nayyar, 2024; Das & Rad, 2020; Kuppa & Le-Khac, 2020), while prioritizing performance could reduce transparency (Ioku, Song & Watamura, 2024). A clear conceptual map of these interactions is essential for building a theoretical framework.

Qualitative Case Study Examination

To validate the relevance of these constructs, qualitative case studies from real-world applications are examined. In healthcare, for example, AI systems like IBM Watson for Oncology have demonstrated the need for explainability to justify diagnostic decisions, while maintaining security to protect patient data (Ezzeddine, 2024; Hulsen, 2023; Martens, De Wolf & De Marez, 2024). In retail, fraud detection systems often face the challenge of balancing interpretability for audit purposes with the need for robust security measures to prevent adversarial attacks (Baniecki & Biecek, 2024; Park et al., 2024). These case studies provide empirical insights into how these constructs manifest in practice and highlight the importance of tailoring solutions to specific contexts.

Integration Approach

The methodology integrates theoretical constructs with qualitative evidence to create a practical and generalizable framework. This involves synthesizing findings from the conceptual analysis and case studies, ensuring that the proposed framework addresses the unique challenges of different domains. For example, the framework incorporates adaptive mechanisms to balance explainability and security dynamically, based on the specific requirements of the application (Akhtar, Kumar & Nayyar, 2024; Kuppa & Le-Khac, 2020). This integration approach ensures that the framework is grounded in both theoretical rigor and real-world applicability.

Outcome

The outcome of this methodology is a robust conceptual framework that provides actionable insights into the trade-offs between security, interpretability, and performance. By grounding the framework in both theory and practice, the methodology ensures that it is not only relevant to current challenges but also adaptable to future developments in Agentic AI systems. This comprehensive approach supports stakeholders in making informed decisions, fostering trust, and ensuring accountability in AI applications (Molnar, 2022).

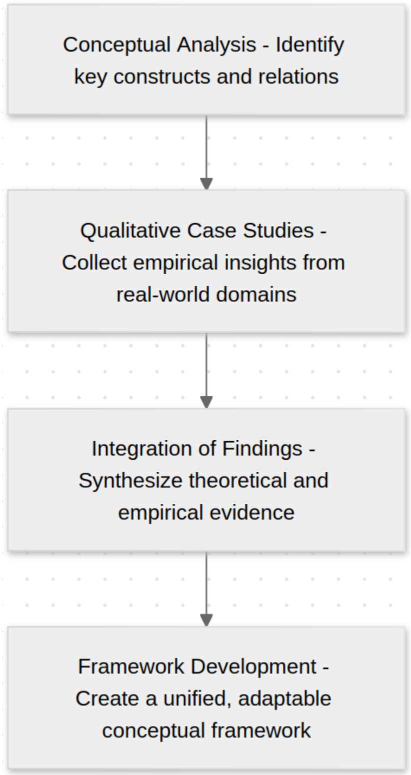
Conceptual Analysis - Identify key constructs and relations
Qualitative Case Studies - Collect empirical insights from real-world domains
Integration of Findings - Synthesize theoretical and empirical evidence
Framework Development - Create a unified, adaptable conceptual framework

Figure 3: Illustrates the methodological approach, detailing the interplay between theoretical constructs, case study insights, and framework development.

Conceptual Framework: Secure-Explainable Agentic AI (SEAAI)

The Secure-Explainable Agentic AI (SEAAI) framework addresses the complex interplay of explainability, security, and performance in AI systems. By integrating modular principles, targeted mechanisms, and theoretical foundations, this framework ensures that AI systems remain transparent, secure, and efficient in high-stakes environments.

Core Principles

1) Modular Explainability

Modular explainability ensures that the components of an AI system responsible for generating explanations are isolated, preventing unintended exposure of sensitive mechanisms. This modularity reduces the risk of information leakage that could be exploited for adversarial purposes (Zheng et al., 2022). For example, isolating interpretability components in fraud detection systems safeguards the system from attackers who might manipulate feature importance to bypass detection.

2) Differential Privacy in Explanations

The integration of differential privacy techniques into explanations ensures that outputs provide valuable insights without exposing sensitive data. Differential privacy methods, such as those used in healthcare diagnostics, allow AI models to share aggregate trends while preserving individual privacy, aligning with regulatory frameworks like GDPR (Abadi et al., 2016).

3) Performance-Explainability Equilibrium (PEE)

PEE is a strategic approach to balancing interpretability with performance. This principle ensures that explanations do not overly compromise the system's efficiency or security. For instance, employing selective feature importance visualizations tailored for stakeholders in financial applications prevents overloading users with unnecessary technical details, thereby optimizing interpretability without reducing performance (Casalicchio, Molnar & Bischl, 2019).

4) Explainability-Focused Observability Mechanisms

The SEAAI framework incorporates mechanisms to monitor the activities of deployed AI agents, addressing the risks associated with their operational autonomy. Observability mechanisms include agent logs and activity dashboards that ensure explainability without compromising operational privacy. This principle is particularly relevant for multi-agent systems, where dependencies and behavioral changes must be tracked to prevent cascading failures (de Santana et al., 2023; Dong, Lu & Zhu, 2024; Elkhawaga, Abu-Elkheir & Reichert, 2022; European Commission's Artificial Intelligence Act, 2024).

5) Multi-Agent Interactions and Sub-Agent Creation

The framework emphasizes mechanisms for monitoring interactions and dependencies within multi-agent systems. Sub-agent creation, where agents generate smaller specialized agents for specific tasks, increases the risk of unintended behaviors. Techniques such as dependency graphs and behavioral analysis tools are proposed to detect and mitigate risks from sub-agents operating beyond the intended scope (Arora et al., 2024; Dehimi et al., 2023).

Components of the SEAAI Framework

1) Security Layer

The Security Layer of the SEAAI framework is designed to protect AI systems from adversarial attacks, ensure data privacy, and maintain robustness in high-risk applications. This layer integrates multiple techniques, including encryption, adversarial robustness strategies, and privacy-preserving mechanisms, to safeguard both the system and its users.

Encryption Techniques

Encryption serves as a foundational tool for securing sensitive data in AI systems. Techniques such as homomorphic encryption allow computations to be performed on encrypted data without needing decryption, thereby preserving confidentiality throughout the process. This method is particularly useful in healthcare, where patient data must remain secure while enabling AI systems to analyze medical records (Acar, 2018).

Adversarial Robustness

Adversarial robustness involves designing AI models to resist manipulations by malicious actors. Techniques such as adversarial training, where models are exposed to adversarial examples during training, enhance the system's ability to identify and mitigate attacks. For instance, fraud detection systems that employ adversarial training are better equipped to handle deceptive inputs designed to bypass detection mechanisms (Baniecki & Biecek, 2024; Park et al., 2024; Zbrzezny & Grzybowski, 2023).

Privacy-Preserving Mechanisms

Privacy-preserving mechanisms, such as differential privacy and federated learning, are vital for protecting user data while enabling explainability. Differential privacy ensures that system outputs cannot be traced back to individual data points, making it a key tool in compliance with regulations like GDPR. Federated learning allows AI models to train across decentralized devices without sharing raw data, reducing the risk of data leakage in distributed environments (Abadi et al., 2016).

Application Programming Interface (API)-Based Monitoring and Access Control

APIs play a critical role in regulating access to AI systems and ensuring explainability. APIs can enforce rules such as rate limits, scope of services, and user verification to prevent misuse. For example, financial institutions may restrict access to AI models unless the requesting entity has verified credentials, reducing the risk of unauthorized usage (Bakar & Selamat, 2018; Jung et al., 2012). This layer also includes tools like CAPTCHA systems to differentiate between genuine human users and automated agents attempting to bypass controls.

Preventing AI Mimicry and Exploitation

The Security Layer addresses the challenge of AI agents mimicking human behavior to exploit systems. Techniques such as behavioral analysis (Arora et al., 2024; Dehimi et al.,

2023) and advanced CAPTCHA mechanisms (Yasur et al., 2023) are proposed to detect and block disguised AI activities. For high-risk scenarios, identity verification protocols, including biometric authentication or "know-your-customer" (KYC) regulations, further enhance security. However, these tools must evolve to counteract increasingly sophisticated AI capabilities, including the generation of fake document or fake identification or forged credentials. Understanding mechanisms that balance identity verification and privacy can provide practical solutions (Phillips et al., 2021).

Balancing Privacy and Access

To mitigate the trade-off between privacy and system utility, the Security Layer incorporates mechanisms to modulate access granularity. Aggregated, de-identified, or differentially private data are shared for general insights, while identifiable information is made accessible only under strict conditions. For instance, regulators investigating high-risk transactions may request specific logs upon approval from third-party adjudicators, ensuring privacy is not compromised unnecessarily (de Santana et al., 2023; Elkhawaga, Abu-Elkheir & Reichert, 2022; European Commission's Artificial Intelligence Act, 2024).

Adaptive Threat Detection

Finally, adaptive models that continuously monitor and learn from new threats are integrated into the Security Layer. These models use real-time analytics to detect anomalies and potential risks, enabling systems to adapt dynamically to evolving security challenges. This proactive approach is critical in domains such as autonomous vehicles and digital infrastructure, where real-time responses are essential for safety and functionality (Aminu, Akinsanya & Dako, 2024).

2) Explainability Layer

The explainability layer selectively exposes agent activities using agent cards containing identifiable or aggregated data. These cards help organizations provide explanations tailored to different stakeholders, ensuring that data privacy is maintained. Privacy assurances like no logging of inputs and outputs in language model APIs and the ability to delete logs are critical in this layer. For example, adhering to GDPR requirements ensures that customer data remains protected while allowing explanations to address regulatory needs (Abadi et al., 2016).

Granularity and access control play a central role in this layer. Aggregated data offers general insights, while identifiable data enables specific investigations. Access must be minimized to legitimate objectives, with identifiable logs made available only under compelling need, as approved by third-party adjudicators. These measures ensure that explainability does not lead to over-surveillance or data misuse (Papagni et al., 2023; Zhou, Boussard & Delaborde, 2021).

3) Performance Layer

The performance layer incorporates adaptive learning models designed to optimize domain-specific needs. These models balance correctness, accuracy, and interpretability dynamically. For example, EBMs provide interpretable insights without significantly sacrificing accuracy, making them suitable for sensitive applications like healthcare diagnostics or fraud detection (Akhtar, Kumar & Nayyar, 2024; Dwivedi et al., 2023; Kuppa & Le-Khac, 2020; Xu, Baracaldo, & Joshi, 2021).

Theoretical Foundations

1) Integrated Compliance and Performance Model (ICPM)

To effectively balance the competing objectives of explainability, security, and performance in Agentic AI systems, a mathematical framework is essential. This paper proposes the **Integrated Compliance and Performance Model (ICPM)** to quantify, represented in Equation (1), and evaluate these trade-offs across diverse application domains. The model incorporates key metrics such as the **Explainability Compliance Index (ECI)**, **Security Risk Index (SRI)**, **Interpretability Score (IS)**, **Performance Trade-Off Ratio (PTR)**, and **Adaptive Explainability Score (AES)**, providing a structured approach to assess system effectiveness.

The ICPM is represented as:

$$ICPM = w_1 \cdot ECI - w_2 \cdot SRI + w_3 \cdot IS - w_4 \cdot PTR + w_5 \cdot AES \qquad (1)$$

Key Variables and Relationships

1. **Trade-Off Metrics**

1. **Security Risk Index (SRI):** Evaluates potential vulnerabilities, influenced by susceptibility to adversarial attacks and data privacy risks. Lower SRI implies a more secure system. It is represented in Equation (2).

$$SRI = f(A_t, D_p) \qquad (2)$$

Where $A_t$is the susceptibility to adversarial attacks, and $D_p$is data privacy risk.

2. **Interpretability Score (IS):** Reflects the system's clarity for stakeholders, emphasizing simplicity and logical coherence. Higher interpretability enhances decision-making. It is represented in Equation (3).

$$IS = f(S_u, C_l) \qquad (3)$$

Where $S_u$is simplicity for the user, and $C_l$is clarity of logic.

3. **Performance Trade-Off Ratio (PTR):** Represents the balance between model performance (e.g., accuracy, latency) and explainability. Lower PTR indicates a better balance. It is represented in Equation (4).

$$PTR = PerformanceMetrics(P) \, / \, ExplainabilityMetrics(E) \qquad (4)$$

2. **Compliance and Adaptive Explainability Metrics**

1. **Explainability Compliance Index (ECI):** Measures adherence to explainability standards, calculated as a function of model transparency and completeness of explanations. Higher compliance improves user trust. It is represented in Equation (5).

   $$ECI = f(T_x, C_x) \qquad (5)$$

   where $T_x$is transparency of the model, and $C_x$is completeness of explanations.

2. **Adaptive Explainability Score (AES):** Assesses the system's ability to adjust explanations based on context and feedback. Higher AES indicates better adaptability. It is represented in Equation (6).

   $$AES = f(C_d, U_f) \qquad (6)$$

   where $C_d$is contextual demand, and $U_f$is user feedback effectiveness.

Dimensions and Interaction Effects

- **Weights (w_1, w_2, w_3, w_4, w_5):** Importance assigned to each metric, depending on the application domain (e.g., healthcare, education).
- For example, in healthcare: $w_1 > w_3 > w_5 > w_2 > w_4$, emphasizing compliance and interpretability
- **Dynamic Adjustments:** Feedback loops that recalibrate metrics based on evolving contexts, such as changing security risks or interpretability demands.

- **Interaction Effects:** Dependencies between variables (e.g., higher interpretability may slightly increase security risks due to transparency).

Illustrative Example

To illustrate, assume values for a hypothetical healthcare diagnostic system:

- ECI=0.85, SRI=0.25, IS=0.9, PTR=1.2, AES=0.8
- Weights: w_1= 0.4, w_2= 0.3, w_3= 0.2, w_4= 0.05, w_5= 0.05

Substitute into the equation (1):

$\text{ICPM} = (0.4 \cdot 0.85) - (0.3 \cdot 0.25) + (0.2 \cdot 0.9) - (0.05 \cdot 1.2) + (0.05 \cdot 0.8)$

$\text{ICPM} = 0.34 - 0.075 + 0.18 - 0.06 + 0.04 = 0.425$

The positive ICPM score indicates a well-balanced system prioritizing explainability and security while managing performance trade-offs.

Potential Applications

- **Healthcare:** Use ICPM to ensure safe, interpretable diagnostic AI systems.
- **Retail:** Apply to fraud detection systems balancing interpretability and security.
- **Education:** Implement to design transparent grading systems with privacy safeguards.

This equation (1) provides a structured approach to quantify and optimize the trade-offs inherent in AI system design, enabling balanced decision-making across diverse application domains.

2) Game-Theoretic Model: Multi-Stakeholder Trade-Off Game (MSTG)

Game-theoretic principles are applied to balance competing objectives of explainability, security, and performance. These models analyze risks and optimize trade-offs by simulating scenarios where attackers and defenders (developers, users, and regulators) interact. For instance, attackers exploiting interpretability insights to bypass security measures are modeled to identify optimal defense strategies (Abate et al., 2021; Abdallah et al., 2024; Seiler, 2023; Yang & Wang, 2019). The objective of the system is to achieve a **secure equilibrium** that balances the trade-offs while mitigating the impact of adversarial actions.

Mathematical Representation

- **Players**
  - ◦ P_1: Developers (optimize security, performance)
  - ◦ P_2: Users (prioritize interpretability, explainability)

- P_3: Regulators (enforce compliance, manage risks)
- P_4: Attackers (maximize security risks or reduce explainability)

- **Strategies (S_i)**: Each player chooses a strategy to either enhance (defenders) or disrupt (attackers) the system's metrics:
- Developers: $S_D = w_1^D, w_2^D, w_3^D, w_4^D, w_5^D$
- Users: $S_U = w_1^U, w_2^U, w_3^U, w_4^U, w_5^U$
- Regulators: $S_R = w_1^R, w_2^R, w_3^R, w_4^R, w_5^R$
- Attackers: $S_A = a_1, a_2, a_3, a_4, a_5$, where a_k represents an attack strategy targeting metric M_k.

- **Payoff Function**

- **Defender Payoff Function**: Each defender's utility function is now penalized by the impact of attackers' actions:

$$U_i = \sum_{k=1}^{5} w_k^i \cdot M_k - C_i - \alpha \cdot Impact(S_A)$$

Where:

- U_i: Payoff for player i
- M_k: Metric value for k (e.g., M_1=ECI, M_2=SRI, M_3=IS, etc.)
- C_i: Cost incurred by player i for achieving their strategy
- Impact(S_A): Represents the cumulative effect of attackers' strategies.
- α: A scaling factor representing the sensitivity of the system to attacks.

- **Attacker Payoff Function**: Attackers aim to maximize disruption by reducing key metrics:

$$U_A = \sum_{k=1}^{5} a_k \cdot (1 - M_k) - C_A$$

Where:

- C_A: Cost incurred by attackers to execute their strategies.

- **System-Wide Equation**
  - The system-wide equilibrium is now governed by the joint utility of defenders and the disruption caused by attackers:

$$ICPM = \sum_{k=1}^{4} \left(\sum_{k=1}^{5} U_i \right) - U_A$$

- **Variables and Interrelations**
  - **Attack Strategies (S_A)**
    - a_1: Exploiting compliance gaps (reducing ECI)
    - a_2: Introducing vulnerabilities (increasing SRI)
    - a_3: Obfuscating outputs (reducing IS)
    - a_4: Increasing system inefficiencies (increasing PTR)
    - a_5: Manipulating context adaptability (reducing AES)
  - **Impact Function (Impact(S_A))**: Measures the effectiveness of attack strategies:

$$Impact(S_A) = \sum_{k=1}^{5} \beta_k \cdot a_k$$

    - β_k: Weight representing the severity of the attack on metricM_k.
  - **Defender-Attacker Interaction**
    - Higher SRI increases Impact(S_A), amplifying attackers' utility.
    - Enhanced ECI, IS, and AES mitigate attackers' effectiveness by reducing β_k.

- **Example Scenario**
  - In a financial fraud detection system:
    - Developers and regulators prioritize SRI and ECI to safeguard compliance and security.
    - Users focus on IS and AES for transparency and adaptability.
    - Attackers exploit a_2(vulnerabilities) and a_3 (obfuscation) to bypass fraud detection.
  - By solving the game:

- Defenders identify optimal weights (w_k^i) to minimize the impact of S_A.
- Attackers adjust a_k to maximize disruption, revealing areas needing further resilience.

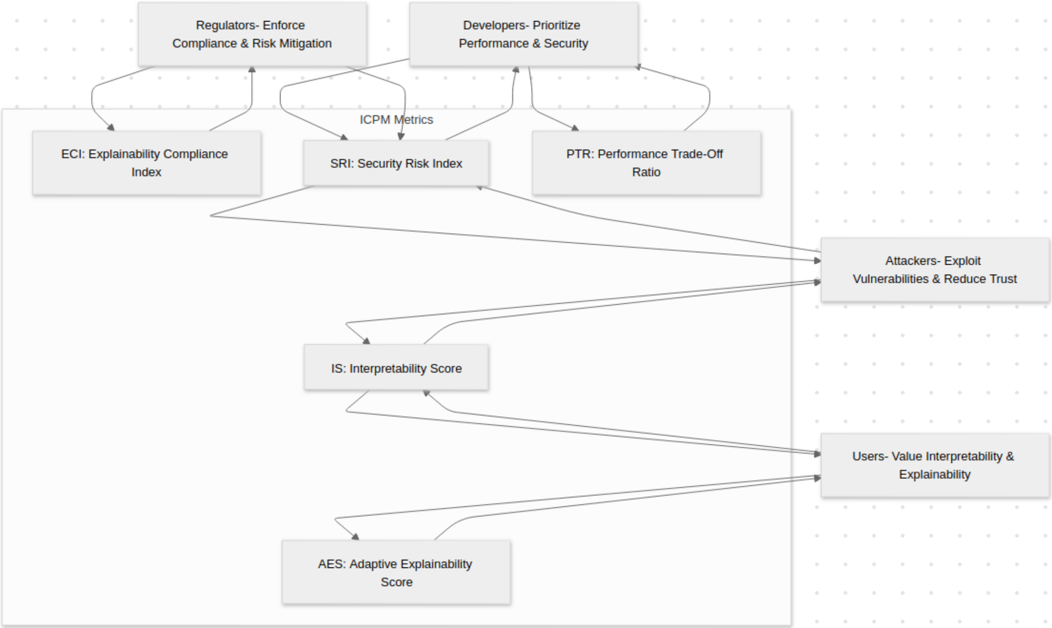
Regulators- Enforce Compliance & Risk Mitigation
Developers- Prioritize Performance & Security
ICPM Metrics
ECI: Explainability Compliance Index
SRI: Security Risk Index
PTR: Performance Trade-Off Ratio
Attackers- Exploit Vulnerabilities & Reduce Trust
IS: Interpretability Score
Users- Value Interpretability & Explainability
AES: Adaptive Explainability Score

Figure 4 illustrates the multi-stakeholder trade-offs in the MSTG model, highlighting the interdependence of explainability, security, and performance metrics in achieving equilibrium.

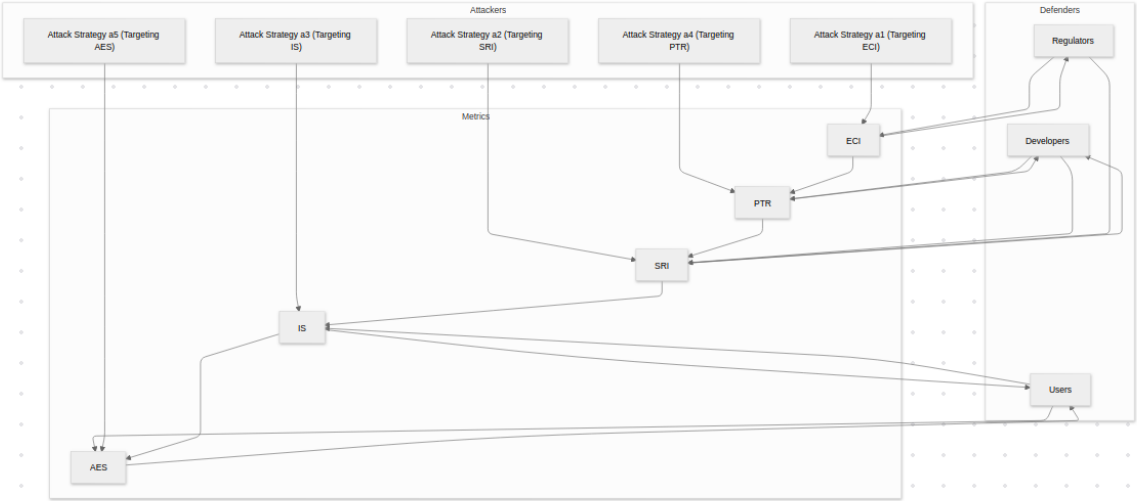
Attackers
Attack Strategy a5 (Targeting AES)
Attack Strategy a3 (Targeting IS)
Attack Strategy a2 (Targeting SRI)
Attack Strategy a4 (Targeting PTR)
Attack Strategy a1 (Targeting ECI)
Defenders
Regulators
Developers
Users
Metrics
ECI
PTR
SRI
IS
AES

Figure 5 illustrates the adversarial dynamics in the MSTG model, highlighting the interplay between defender strategies and attacker actions to maintain system equilibrium.

Application Scenarios

The SEAAI framework finds application in various critical sectors, each requiring a careful balance between explainability, security, and performance. By addressing specific challenges, the framework enhances transparency and trust without compromising data privacy or operational integrity.

Healthcare

In healthcare, explainability is crucial for diagnostic systems to build trust among practitioners and patients. AI-driven diagnostic agents must provide interpretable decisions to enable practitioners to validate recommendations. For example, systems like IBM Watson for Oncology provide insights into treatment plans based on patient data while maintaining privacy through encryption and differential privacy techniques (Acar, 2018; Abadi et al., 2016; Ezzeddine, 2024; Hulsen, 2023; Martens, De Wolf & De Marez, 2024).

Explainability ensures accountability in deployed diagnostic agents by offering explanations tailored to clinical requirements without compromising patient confidentiality. Overreliance on such systems, however, introduces risks if users blindly trust outputs without understanding limitations. Malfunctions or biased data can result in severe consequences, highlighting the need for human oversight in clinical decision-making (Bakken, 2023; Challen et al., 2019; Mosqueira-Rey et al., 2023).

Retail

Retail applications of AI benefit significantly from explainability in recommendation systems. Explainable algorithms ensure transparency in product suggestions, improving consumer trust while protecting proprietary data. For instance, techniques like SHAP or LIME provide feature attributions that clarify why a product was recommended (Chan et al., 2023; Dwivedi et al., 2023; Nigon et al., 2024; Phillips et al., 2021).

Fraud detection systems, another critical retail application, require explainability to mitigate adversarial risks. Adversaries often exploit feature importance or system vulnerabilities to bypass fraud detection mechanisms. API-based restrictions and identifier protocols can enhance security by limiting access to sensitive algorithms and ensuring only verified entities interact with the system. This dual approach of explainability and security strengthens fraud prevention (Baniecki & Biecek, 2024; Park et al., 2024).

Education

In education, explainable AI enhances transparency in grading systems and learning analytics. Transparent grading systems provide students and educators with clear insights into evaluation processes, ensuring fairness and accountability. For instance, AI systems using interpretable models like decision trees can highlight criteria that led to specific grades while protecting sensitive profiles through privacy-preserving methods (Chitti, Chitti & Jayabalan, 2020; Lünich & Keller, 2024; Memarian & Doleck, 2023).

AI tools must also safeguard data logs to prevent unauthorized access to student profiles and model predictions. Techniques like differential privacy and federated learning ensure that sensitive data remains secure, maintaining trust among stakeholders while enabling educational institutions to leverage analytics effectively (Abadi et al., 2016).

Generalizable Applications

The SEAAI framework can address cross-domain challenges in financial services, autonomous systems, and smart cities. Multi-agent systems in these domains often encounter feedback loops and systemic vulnerabilities. For example, in financial trading, autonomous agents interacting in volatile markets can lead to cascading failures, as observed in the 2010 flash crash (CFTC & SEC, 2010; Dorri, Kanhere & Jurdak, 2018).

Explainability mechanisms in such environments are essential for monitoring agent interactions and identifying potential risks. By providing insights into agent behavior and decision-making processes, the SEAAI framework facilitates proactive mitigation strategies. Similarly, in smart cities, AI systems managing infrastructure must balance transparency and security to address citizen concerns while ensuring efficient service delivery (Arora et al., 2024; Dehimi et al., 2023).

Prospective Applications and Future Evaluation

Although the framework remains conceptual, several opportunities exist to apply and assess it in practical settings. Integrating its metrics—such as the Explainability Compliance Index or Security Risk Index—into existing AI systems in healthcare, finance, or education could provide insights into how adjustments influence trust, security, and performance (Phillips et al., 2021). Rather than immediate large-scale deployment, initial pilot studies may help gauge the framework's effects. For instance, small clinics might trial explainability features to determine their impact on clinician comprehension and patient confidence, while a retail fraud detection system could evaluate privacy-preserving methods to enhance user understanding.

Data gathered from these limited trials would inform refinements and highlight areas needing improvement. Collaborations with industry partners, regulators, and research institutions would be valuable, as access to authentic scenarios and ongoing feedback would guide systematic enhancements. Retrospective analyses of historical datasets offer another avenue: simulating how outcomes might differ had the framework's principles guided earlier decisions.

A phased evaluation approach could start with controlled simulations, advance to restricted field tests, and eventually progress to broader adoption if initial results prove promising. By combining quantitative metrics—such as changes in accuracy or reductions in vulnerabilities—with qualitative input from stakeholders, a holistic understanding of the framework's value emerges. Over time, evidence-based validation can strengthen the framework's credibility, ensuring that Agentic AI systems evolve from theoretical constructs into trusted, widely applicable solutions.

DISCUSSION

Challenges and Limitations

Implementing the SEAAI framework may face resistance due to perceived complexity and resource demands. Advanced measures like differential privacy and federated learning can improve explainability and security but also strain computational and financial resources, especially in settings with limited infrastructure (Abadi et al., 2016). Similarly, balancing ease of access with stringent explainability controls is difficult. Overly restrictive conditions could reduce usability, while insufficient restrictions risk security breaches. Regulatory oversight may require granular data access, raising privacy concerns if stakeholders view such measures as intrusive. Moreover, achieving an optimal equilibrium among accuracy, interpretability, and resilience often involves resource-intensive adaptive models.

Opportunities for Future Research

Future work can focus on tailoring the SEAAI framework to resource-constrained contexts, ensuring its principles reach a wider range of sectors. Establishing standardized protocols to measure trade-offs between security, explainability, and performance would create reference points for comparing different systems. Solutions to detect disguised agent activities, such as enhanced CAPTCHA techniques or advanced behavioral analysis, remain essential. Success in these areas would enable seamless traceability and oversight in decentralized environments, supporting the framework's applicability to autonomous systems and smart cities. By exploring these directions, future research can continue refining SEAAI's scope and effectiveness.

CONCLUSION

This research presents the SEAAI framework, a novel conceptual model designed to balance explainability, security, and performance in Agentic AI systems. The framework incorporates modular explainability, privacy-preserving mechanisms, and adaptive performance strategies to address complex trade-offs across diverse applications. The proposed theoretical foundations, including metrics and game-theoretic models, provide actionable insights for real-world deployment, emphasizing resilience against adversarial attacks. The study highlights the necessity for interdisciplinary collaboration to refine and operationalize SEAAI, ensuring safe and transparent AI integration across industries.

Declaration of interests

The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of funding

The author declare that they did not receive any funding for this paper.

Statement Generative AI

During the development of this manuscript the author used AI tools in order to assist with:

- Structuring ideas and thoughts more coherently.
- Checking for grammar, spelling, and sentence clarity.
- Rephrasing to improve readability and flow.

While these tools helped enhance clarity and structure, all intellectual contributions, theoretical frameworks, arguments, and analyses are entirely by the author. All sources are properly cited. After using the AI tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

REFERENCES

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16),* 2016. https://doi.org/10.1145/2976749.2978318
2. Abate, A., Gutierrez, J., Hammond, L., Harrenstein, P., Kwiatkowska, M., Najib, M., Perelli, G., Steeples, T., & Wooldridge, M. Rational verification: Game-theoretic verification of multi-agent systems. *Applied Intelligence,* 2021; *51:* 6569–6584. https://doi.org/10.1007/s10489-021-02658-y
3. Abdallah, M., Bagchi, S., Bopardikar, S. D., Chan, K., Gao, X., Kantarcioglu, M., Li, C., Liu, P., & Zhu, Q. Game theory in distributed systems security: Foundations, challenges, and future directions. *IEEE Security & Privacy,* 2024; 2–13. https://doi.org/10.1109/MSEC.2024.3407593
4. Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (CSUR),* 2018; *51*(4): 79, 1–35. https://doi.org/10.1145/3214303
5. Aggarwal, N., & Liu, A. (2023). KPIs for gen AI: Why measuring your new AI is essential to its success. *Google Cloud*. Retrieved December 10, 2024, from https://cloud.google.com/transform/kpis-for-gen-ai-why-measuring-your-new-ai-is-essential-to-its-success
6. Akhtar, M.A.K., Kumar, M., Nayyar, A. Privacy and Security Considerations in Explainable AI. In: Towards Ethical and Socially Responsible Explainable AI. Studies in Systems, Decision and Control, 2024; 551. Springer, Cham. https://doi.org/10.1007/978-3-031-66489-2_7
7. Aminu, M., Akinsanya, A., & Dako, D. A. Enhancing cyber threat detection through real-time threat intelligence and adaptive defense mechanisms. *International Journal of Computer Applications Technology and Research,* 2024; *13*(8): 11–27. https://doi.org/10.7753/IJCATR1308.1002
8. Arora, D., Sonwane, A., Wadhwa, N., Mehrotra, A., Utpala, S., Bairi, R., Kanade, A., & Natarajan, N. (2024). MASAI: Modular architecture for software-engineering AI agents. *arXiv*. https://doi.org/10.48550/arXiv.2406.11638
9. Bakar, N. A., & Selamat, A. Agent systems verification: Systematic literature review and mapping. *Applied Intelligence,* 2018; *48*(5): 1251–1274. https://doi.org/10.1007/s10489-017-1112-z

10. Bakken, S. AI in health: Keeping the human in the loop. *Journal of the American Medical Informatics Association*, 2023; *30*(7): 1225–1226. https://doi.org/10.1093/jamia/ocad091
11. Baniecki, H., & Biecek, P. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 2024; *107:* 102303. https://doi.org/10.1016/j.inffus.2024.102303
12. Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. ChemCrow: Augmenting large-language models with chemistry tools. *ArXiv*, 2023. https://doi.org/10.48550/arXiv.2304.05376
13. Casalicchio, G., Molnar, C., Bischl, B. Visualizing the Feature Importance for Black Box Models. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science, 2019; 11051. Springer, Cham. https://doi.org/10.1007/978-3-030-10925-7_40
14. CFTC (U.S. Commodity Futures Trading Commission) & SEC (U.S. Securities and Exchange Commission). (2010). *Preliminary findings regarding the market events of May 6, 2010.* Retrieved December 11, 2024, from https://www.sec.gov/sec-cftc-prelimreport.pdf
15. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 2019; *28*(3): 231–237. https://doi.org/10.1136/bmjqs-2018-008370
16. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., & Maharaj, T. Harms from increasingly agentic algorithmic systems. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT'23)*, 2023; 651–666. Association for Computing Machinery. https://doi.org/10.1145/3593013.359403
17. Chitti, M., Chitti, P., & Jayabalan, M. Need for interpretable student performance prediction. *2020 13th International Conference on Developments in eSystems Engineering (DeSE)*, 2020; 269–272. https://doi.org/10.1109/DeSE51703.2020.9450735
18. Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. ArXiv:2006.11371. https://doi.org/10.48550/arXiv.2006.11371

19. Dehimi, N. E. H., Galland, S., Tolba, Z., Allaoua, N., & Ferkani, M. Distributed, Dynamic and Recursive Planning for Holonic Multi-Agent Systems: A Behavioural Model-Based Approach. *Electronics*, 2023; *12*(23): 4797. https://doi.org/10.3390/electronics12234797
20. De Santana, V. F., Fucs, A., Segura, V., de Moraes, D. B., & Cerqueira, R. Predicting the need for XAI from high-granularity interaction data. International Journal of Human-Computer Studies, 2023; 175: 103029. https://doi.org/10.1016/j.ijhcs.2023.103029
21. Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, 2023; *99:* 101896. https://doi.org/10.1016/j.inffus.2023.101896
22. Dong, L., Lu, Q., & Zhu, L. (2024). AgentOps: Enabling observability of LLM agents. *arXiv*. https://doi.org/10.48550/arXiv.2411.05285
23. Dorri, A., Kanhere, S. S., & Jurdak, R. Multi-agent systems: A survey. *IEEE Access*, 2018; *6:* 28573–28593. https://doi.org/10.1109/ACCESS.2018.2831228
24. Duan, W., McNeese, N., Freeman, G., & Li, L. Mitigating gender stereotypes toward AI agents through an explainable AI (XAI) approach. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 2024; 430: 1–35. https://doi.org/10.1145/3686969
25. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys, 2023; 55(9): 194. https://doi.org/10.1145/3561048
26. Elkhawaga, G., Abu-Elkheir, M., & Reichert, M. Explainability of Predictive Process Monitoring Results: Can You See My Data Issues? *Applied Sciences*, 2022; *12*(16): 8192. https://doi.org/10.3390/app12168192
27. Erasmus, A., Brunet, T. D. P., & Fisher, E. What is interpretability? Philosophy & Technology, 2021; 34(3): 833–862. https://doi.org/10.1007/s13347-020-00435-2
28. European Commission. (2024). Artificial Intelligence Act. Regulation 2024/1689. (https://eur-lex.e uropa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689) (accessed December 11, 2024)
29. Ezzeddine, F. (2024). Privacy implications of explainable AI in data-driven systems. arXiv:2406.15789. https://doi.org/10.48550/arXiv.2406.15789
30. Fiske, A., Henningsen, P., & Buyx, A. Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and

psychotherapy. *Journal of Medical Internet Research,* 2019; *21*(5): e13216. https://doi.org/10.2196/13216

31. Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. Social bots and the spread of disinformation in social media: The challenges of artificial intelligence. *British Journal of Management,* 2022; *33*(3): 1238–1253. https://doi.org/10.1111/1467-8551.12554
32. Huber, T., Weitz, K., André, E., & Amir, O. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. Artificial Intelligence, 2021; 301: 103571. https://doi.org/10.1016/j.artint.2021.103571
33. Hulsen T. Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare. *AI*., 2023; 4(3): 652-666. https://doi.org/10.3390/ai4030034
34. Ioku, T., Song, J., & Watamura, E. Trade-offs in AI assistant choice: Do consumers prioritize transparency and sustainability over AI assistant performance? *Big Data & Society*, 2024; *11*(4). https://doi.org/10.1177/20539517241290217
35. Jabarian, B. (2024). *Black boxes: Mental models and AI models.* Retrieved December 11, 2024, from https://static1.squarespace.com/static/62a9c8018c274b7728fc5d89/t/6709517bdd656c29d17e5314/1728663933218/Jabarian_MentalModels.pdf
36. Jung, Y., Kim, M., Masoumzadeh, A., & Joshi, J. B. D. A survey of security issues in multi-agent systems. *Artificial Intelligence Review,* 2012; *37:* 239–260. https://doi.org/10.1007/s10462-011-9228-8
37. Kenton, Z., Kumar, R., Farquhar, S., Richens, J., MacDermott, M., & Everitt, T. Discovering agents. *Artificial Intelligence,* 2023; *322:* 103963. https://doi.org/10.1016/j.artint.2023.103963
38. Kuppa, A., & Le-Khac, N.-A. (2020). Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. *2020 International Joint Conference on Neural Networks (IJCNN),* 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206780
39. Li, L., Lassiter, T., Oh, J., & Lee, M. K. Algorithmic hiring in practice: Recruiter and HR professionals' perspectives on AI use in hiring. *In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21),* 2021; 166–176. Association for Computing Machinery. https://doi.org/10.1145/3461702.3462531
40. Lieberman, H. Autonomous interface agents. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '97),* 1997; 67–74. Association for Computing Machinery. https://doi.org/10.1145/258549.258592

41. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., & Tang, J. (2023). AgentBench: Evaluating LLMs as agents. *arXiv*. https://doi.org/10.48550/arXiv.2308.03688
42. Lünich, M., & Keller, B. Explainable artificial intelligence for academic performance prediction: An experimental study on the impact of accuracy and simplicity of decision trees on causability and fairness perceptions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24),* 2024; 1031–1042. Association for Computing Machinery. https://doi.org/10.1145/3630106.3658953
43. Martens, M., De Wolf, R., & De Marez, L. Trust in algorithmic decision-making systems in health: A comparison between ADA health and IBM Watson. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 2024; *18*(1): 5. https://doi.org/10.5817/CP2024-1-5
44. Memarian, B., & Doleck, T. Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence,* 2023; *5:* 100152. https://doi.org/10.1016/j.caeai.2023.100152
45. Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable.* Retrieved December 10, 2024, from https://christophm.github.io/interpretable-ml-book/
46. Moskalenko, V., Kharchenko, V., Moskalenko, A., & Kuzikov, B. Resilience and Resilient Systems of Artificial Intelligence: Taxonomy, Models and Methods. *Algorithms*, 2023; *16*(3): 165. https://doi.org/10.3390/a16030165
47. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 2023; *56:* 3005–3054. https://doi.org/10.1007/s10462-022-10246-w
48. Nigon, N., Tucker, J. D., Ekstedt, T. W., Jeong, B. C., Simionescu, D. C., & Koretsky, M. D. (2024). Adaptivity or agency? Educational technology design for conceptual learning of materials science. *Computer Applications in Engineering Education.* https://doi.org/10.1002/cae.22790
49. Papagni, G., de Pagter, J., Zafari, S., Filzmoser, M., & Koeszegi, S. T. Artificial agents' explainability to support trust: Considerations on timing and context. *AI & SOCIETY*, 2023; *38:* 947–960. https://doi.org/10.1007/s00146-022-01462-7

50. Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 2024; *5*(5): 100988. https://doi.org/10.1016/j.patter.2024.100988
51. Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. *Four principles of explainable artificial intelligence* (NISTIR 8312). National Institute of Standards and Technology, 2021. https://doi.org/10.6028/NIST.IR.8312
52. Pillai, V. Enhancing Transparency and Understanding in AI Decision-Making Processes. *Iconic Research and Engineering Journals*, 2024; *8*(1): 168-172.
53. Rose, S., & Nelson, C. (2023). *Understanding AI-facilitated biological weapon development.* Centre for Long-Term Resilience.
54. Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C. J., & Hashimoto, T. (2023). Identifying the risks of LM agents with an LM-emulated sandbox. *arXiv*. https://doi.org/10.48550/arXiv.2309.15817
55. Saarela, M., Heilala, V., Jääskelä, P., Rantakaulio, A., & Kärkkäinen, T. Explainable student agency analytics. *IEEE Access,* 2021; *9:* 137444–137459. https://doi.org/10.1109/ACCESS.2021.3116664
56. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv*. https://doi.org/10.48550/arXiv.2302.04761
57. Seiler, B. B. *Applications of Cooperative Game Theory to Interpretable Machine Learning*. Stanford University, 2023.
58. Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O’Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., & Robinson, D. G. (2023). Practices for governing agentic AI systems.
59. Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2023). Cognitive architectures for language agents. *arXiv*. https://doi.org/10.48550/arXiv.2309.02427
60. Wagle, J. M. (2021). *Utilizing the shap framework to bypass intrusion detection systems* (Master's thesis, The University of Bergen).
61. Xu, R., Baracaldo, N., & Joshi, J. (2021). Privacy-preserving machine learning: Methods, challenges, and directions. arXiv:2108.04417.
62. Yang, Y., & Wang, J. (2019). An overview of multi-agent reinforcement learning from a game theoretical perspective. *arXiv*. https://doi.org/10.48550/arXiv.2011.00583

63. Yasur, L., Frankovits, G., Grabovski, F. M., & Mirsky, Y. (2023). Deepfake CAPTCHA: A method for preventing fake calls. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security (ASIA CCS '23)* (pp. 608–622). Association for Computing Machinery. https://doi.org/10.1145/3579856.3595801

64. Zbrzezny, A. M., & Grzybowski, A. E. (2023). Deceptive tricks in artificial intelligence: Adversarial attacks in ophthalmology. *Journal of Clinical Medicine, 12*(9): 3266. https://doi.org/10.3390/jcm12093266

65. Zheng, H., Zang, Z., Yang, S., & Mangharam, R. (2022). Towards explainability in modular autonomous vehicle software. *arXiv*. https://doi.org/10.48550/arXiv.2212.00544

66. Zhou, Y., Boussard, M., & Delaborde, A. (2021). Towards an XAI-assisted third-party evaluation of AI systems: Illustration on decision trees. In D. Calvaresi, A. Najjar, M. Winikoff, & K. Främling (Eds.), *Explainable and transparent AI and multi-agent systems. EXTRAAMAS 2021. Lecture notes in computer science* (Vol. 12688). Springer, Cham. https://doi.org/10.1007/978-3-030-82017-6_10