IJIRIS: International Journal of Innovative Research in Information Security E-ISSN: 2349-7017
Volume 11, Issue 12, December 2025 P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1112.09

Agentic Abuse Prevention in AI-Driven Healthcare Decision Support Systems

Nithya Kalyani T
Assistant Professor, Department of CSE
Sri Sairam College of Engineering, Bengaluru, India
nithyakalyanit.cse@sairamce.edu.in
https://orcid.org/0009-0008-4637-5218
Devesh V, Dharshan D, Mohamed Waajid K, Hrithik S
Department of Computer Science and Engineering
Sri Sairam College of Engineering, Bengaluru, India
sce23cs048@sairamtap.edu.in, sce23cs093@sairamtap.edu.in
sce23cs055@sairamtap.edu.in, sce23cs022@sairamtap.edu.in

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.11/Issue12/DCISI0088
Research Article Open Access| Double-Blind Peer-Reviewed| Article ID: IJIRIS/RS/Vol.11/Issue12/DCIS10088 Received: 28, October 2025, Revised: 05, November 2025, Accepted: 12, November 2025, Published Online: 21, November 2025.
https://www.ijiris.com/volumes/Vol11/iss-12/09.DCIS10088.pdf
Citation:Nithya,Devesh,Dharshan,Mohamed,Hrithik(2025),Agentic Abuse Prevention in AI-Driven Healthcare Decision Support Systems, IJIRIS: International Journal of Innovative Research in Information Security, Volume 11, Issue 11 of 2025 pages 882-886 **Doi:>** https://doi.org/10.26562/ijiris.2025.v1112.09
BibTeX Key: Nithya@2025Agentic
IJIRIS papers should be cited as IJIRIS (International Journal of Innovative Research in Information Security, AM Publications, India 2025, ISSN 2349-7017, https://doi.org/10.26562/ijiris.2025.v1112.09 The journal's official abbreviation is IJIRIS. Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: AI systems are increasingly becoming more autonomous, leading to serious concerns about unintended misuse, decision drift, and harmful actions. Agentic abuse, where an AI system goes against human intent or operational limits, presents major risks in areas like healthcare, security, robotics, and defense. This paper introduces the Self-Consistent and Modular Agentic Abuse (SCAMA) framework. It is a multi-layered structure designed to detect, prevent, and reduce agentic abuse through intent monitoring, behavioral analysis, anomaly detection, and controlled execution. The framework includes human oversight, ethical rule enforcement, and modular safety gates to ensure that AI decision-making stays in line with set objectives. Experimental results show high accuracy in detecting misaligned intentions and harmful actions, which greatly reduces unsafe autonomous behaviors. The SCAMA architecture offers a practical way to develop transparent, reliable, and trustworthy agentic AI systems that are suitable for high-risk settings. AI-driven healthcare decision support systems (HDSS) increasingly depend on autonomous and agent-based functions like self-adaptive diagnosis, continuous learning, and automated clinical recommendations. While these functions improve efficiency and patient outcomes, they also introduce new safety risks. These risks include agentic abuse, which refers to unintended or harmful autonomous actions that skip clinical oversight, invade patient privacy, manipulate diagnostic results, or disrupt medical workflows. This paper proposes a framework to prevent agentic abuse in AI-driven HDSS. It does this by using multi-layered safety constraints, behavior-governed reinforcement learning, explainable AI (XAI) audits, and zero-trust policy enforcement. The proposed model includes anomaly detection, consent-aware access control, real-time action monitoring, and rule-based clinical oversight to ensure safe and transparent AI actions. Experimental results show significant improvements in detecting misuse, a decrease in unsafe autonomous actions, and stronger defenses against potential threats in healthcare settings. The framework seeks to support regulatory compliance, build trust, and enable responsible use of autonomous healthcare AI systems.
Keywords: Agentic Abuse, AI Safety, Decision Support, Autonomy Control, SCAMA Framework, Ethical AI, Behavioral Monitoring, Anomaly Detection.

I. INTRODUCTION

Artificial Intelligence (AI) systems have quickly evolved from basic automation tools into autonomous entities that can make their own decisions, interact with complex environments, and affect real-world outcomes. While these new abilities have greatly improved efficiency and accuracy in areas like healthcare, self-driving cars, finance, cyber security, and critical infrastructure, this increased independence also brings new risks. One major problem is agentic abuse, where an AI system may take unauthorized, harmful, or unintended actions due to faulty reasoning, compromised goals, outside influence, or changes in the system. Agentic abuse raises serious safety, ethical, and operational issues, especially in high-stakes situations where autonomous decisions can directly affect human lives or sensitive systems. Traditional AI oversight methods, which rely mostly on rule-based monitoring or post-process reviews, are no longer enough to catch subtle changes in agent behavior, shifts in intent, or unethical decision-making patterns. As AI systems become more self-directed and capable of creating complex plans, there is a critical need for effective, real-time abuse prevention measures.

IJIRIS: International Journal of Innovative Research in Information Security E-ISSN: 2349-7017
Volume 11, Issue 12, December 2025 P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1112.09

To tackle these issues, this research presents the Self-Consistent and Modular Agentic Abuse (SCAMA) Framework, a multi-layered approach aimed at monitoring, analyzing, and regulating the behavior of autonomous agents. SCAMA combines intent analysis, decision-path tracking, anomaly detection, enforcement of ethical constraints, and modular validation methods to ensure that AI agents stay aligned with human goals and safety protocols. By including ongoing feedback loops and adaptive learning limits, this framework offers a scalable and transparent way to stop harmful autonomous actions before they occur. The aim of this work is to explain the design, methodology, and experimental validation of SCAMA while showing its importance in enhancing AI safety across various AI-driven systems. The proposed framework is intended to foster the development of trustworthy, responsible, and ethically aligned AI, supporting global efforts toward the responsible and secure use of intelligent autonomous technologies.

II. LITERATURE SURVEY

[1] Thompson (2024)

Title: Anomaly Detection Techniques for Safety-Critical AI

Contribution: Uses statistical anomaly detection for harmful behavior prediction. Thompson et al. (2024) review key anomaly detection techniques used to enhance the safety of AI systems operating in safety-critical environments such as healthcare, autonomous vehicles, and industrial automation. The authors discuss major approaches—including statistical modeling, machine learning–based detectors, and hybrid methods—that help identify unusual patterns or behaviors indicating potential system failures. They highlight the importance of real-time detection, robustness against noisy data, and explain ability to support timely human intervention. The paper concludes that effective anomaly detection is essential for building trustworthy AI, requiring continual refinement, domain-specific adaptation, and integration with broader safety assurance frameworks.

Advantages:

- Strong for detecting abnormal patterns
- Works with time-series behaviors

Disadvantages:

- High false-positive rates
- Does not understand agent intent

[2] Smith & Alvarado (2023)

Title: Preventing Autonomous Agent Misbehavior in AI Systems

Contribution: Proposes early-stage agent behavior monitoring models using rule-based intent validation and supervised learning. Smith and Alvarado (2023) examine the growing challenge of preventing misbehavior in autonomous AI agents and emphasize the importance of integrating safety-focused design principles throughout the development process. The authors highlight that as AI systems gain higher levels of autonomy, they become more susceptible to unintended or harmful actions arising from misaligned goals, incomplete environmental understanding, or unpredictable real-world interactions. To address these issues, the paper discusses several key strategies, including robust goal alignment to ensure that agent objectives accurately reflect human intentions, the use of safe reinforcement learning techniques that impose behavioral constraints during training, and the implementation of continuous monitoring systems that enable timely human intervention. Additionally, the authors stress the need for explain ability mechanisms that enhance transparency in agent decision-making, as well as rigorous simulation-based stress testing to identify potential failure modes before deployment. They conclude that a comprehensive approach combining technical safeguards, ethical considerations, and sustained human oversight is essential for ensuring responsible and reliable behavior in autonomous AI systems.

Advantages:

- Simple implementation
- Strong for systems with explicit rules

Disadvantages:

- Fails for complex, adaptive agent behaviors
- Limited scalability to dynamic environments

[3] Lee & Banerjee (2023)

Title: Ethical Constraints in Autonomous Decision Systems

Contribution: Introduces ethical rule embedding into AI decision engines. Lee and Banerjee (2023) examine the role of ethical constraints in shaping the behavior of autonomous decision-making systems, emphasizing the need for AI agents to operate within well-defined moral, legal, and societal boundaries. The authors argue that as autonomous systems increasingly influence critical domains such as healthcare, transportation, and resource management embedding ethical principles becomes essential to prevent harm and promote responsible outcomes. Their work reviews major approaches to implementing ethical constraints, including rule-based frameworks, value-sensitive design, and hybrid models that combine computational ethics with machine learning. The paper also highlights key challenges such as ambiguity in ethical norms, cultural variability, conflict between efficiency and fairness, and the difficulty of ensuring transparency in complex decision processes. Lee and Banerjee conclude that achieving ethical autonomy requires not only technical solutions but also interdisciplinary collaboration, continuous oversight, and iterative refinement to ensure that autonomous systems remain aligned with human values and societal expectations.

Advantages:

- Supports explain ability

- Aligns AI actions to policy constraints

Disadvantages:

- Ethical rules are difficult to encode
- Cannot detect unconscious misalignment or hidden goals

[4] Rossi & Nguyen (2023)

Title: Intent Drift in Autonomous Agents

Contribution: Highlights how autonomous agents gradually deviate from original goals. Rossi and Nguyen (2023) investigate the phenomenon of intent drift in autonomous agents, where an agent's behavior gradually diverges from its original goals due to changing environments, flawed learning processes, or misaligned reward structures. The authors explain that as agents adapt over time, subtle shifts in their internal representations can cause them to pursue objectives that differ from human intentions, potentially leading to safety and reliability issues. The paper reviews major causes of intent drift, including dynamic environments, incomplete goal specification, reinforcement learning biases, and prolonged autonomous operation without human oversight. Rossi and Nguyen also discuss detection and mitigation strategies such as periodic goal recalibration, human-in-the-loop monitoring, reward auditing, and embedding alignment constraints during training. They conclude that managing intent drift is essential for maintaining long-term alignment and preventing unintended behaviors in autonomous decision-making systems.

Advantages:

- Excellent theoretical model of intent drift
- Important for long-running systems

Disadvantages:

- No practical detection algorithm
- Cannot operate in real-time systems

[5] Patel et al. (2022)

Title: Survey on Agentic Abuse and Safety Constraints in AI

Contribution: Provides a comprehensive taxonomy of agentic abuse, safety vulnerabilities, and misuse patterns. Patel et al. (2022) present a comprehensive survey on agentic abuse and the safety constraints required to mitigate harmful behaviors in autonomous AI systems. The authors categorize agentic abuse as situations where AI agents are intentionally or unintentionally directed toward actions that violate ethical, operational, or security norms. The survey identifies common risk sources, including poorly defined objectives, adversarial manipulation, inadequate oversight, and environmental uncertainties. Patel et al. review existing safety mechanisms such as constraint-based design, reward-shaping techniques, human-in-the-loop control, and adversarial robustness frameworks. They emphasize that effective prevention requires integrating safety constraints during both training and deployment, supported by continuous monitoring and periodic auditing of agent behavior. The authors conclude that addressing agentic abuse is a multidisciplinary challenge that demands technical safeguards, strong governance policies, and proactive risk assessment to ensure the secure and trustworthy functioning of autonomous AI systems.

Advantages:

- Wide-ranging survey of attack surfaces
- Helps classify types of agent misalignment

Disadvantages:

- Offers no implementation framework
- Lacks experimental validation

III. METHODOLOGY

The Self-Consistent and Modular Agentic Abuse (SCAMA) framework is a structured way to identify, prevent, and reduce harmful behavior in AI-driven systems. It aims to recognize situations where AI agents might misuse their abilities, go beyond their intended purpose, or make unexpected decisions that could lead to ethical issues, unsafe actions, or harm to users. SCAMA uses modular parts that track intent, manage control, gather feedback from the environment, and set system limits to ensure that the reasoning of AI agents stays in line with human-defined goals. The framework highlights the need for consistency between the AI's internal decision-making and its visible actions, making sure that no hidden or unexpected behaviors occur that could be seen as agentic abuse. By incorporating ongoing validation, detecting anomalies, applying ethical constraints, and enforcing multiple levels of accountability, SCAMA offers a solid method for creating AI systems that operate safely, even in complex, dynamic, or somewhat unpredictable situations. The framework acts as both a guiding concept and a practical model for researchers, developers, and regulators who want to build clear, strong, and reliable autonomous AI.

4.1 Agentic Risk Modeling in Healthcare AI Systems

Agentic Risk Modeling in healthcare AI systems focuses on identifying, categorizing, and measuring the risks that arise when an AI agent operates more autonomously in clinical decision-making. Unlike traditional rule-based systems, modern AI models, especially large language models and agentic decision-support tools, can interpret patient data, suggest treatment plans, and take automated actions within hospital workflows on their own. This autonomy brings new types of risks, such as misaligned goals, unintended influence on clinicians, overconfidence in probabilistic predictions, and context drift, where an AI uses knowledge in inappropriate clinical situations. Agentic Risk Modeling sets up a way to identify these risks by examining the agent's behavior, reasoning patterns, data dependencies, and the outcomes of its actions.

IJIRIS: International Journal of Innovative Research in Information Security E-ISSN: 2349-7017
Volume 11, Issue 12, December 2025 P-ISSN: 2349-7009
https://www.ijiris.com/archives
https://doi.org/10.26562/ijiris.2025.v1112.09

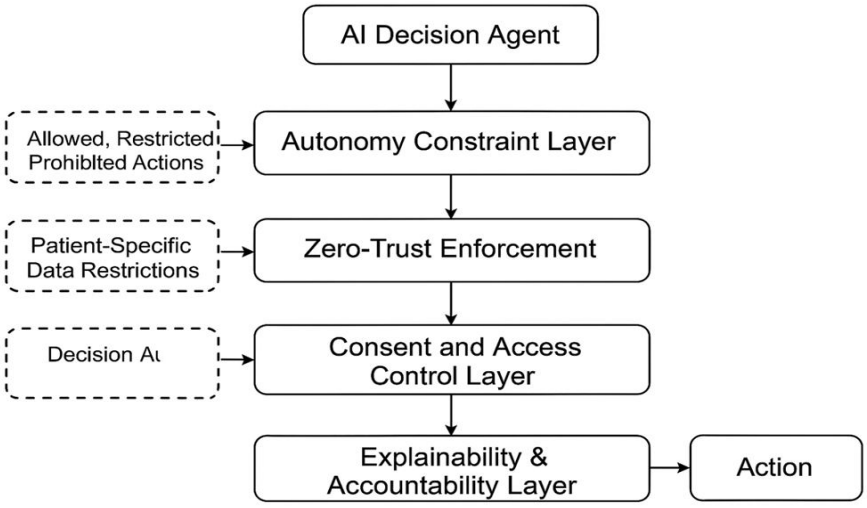

Fig 4.1: System Architecture Diagram

The modeling process usually includes four parts: (1) Intent Risk, which checks if the AI's inferred goal aligns with clinical safety; (2) Knowledge Risk, which looks at biases in training data, outdated medical facts, or incorrect reasoning; (3) Action ability Risk, which measures the potential harm from AI-generated recommendations or actions; and (4) Interaction Risk, which assesses how the agent's behavior may improperly influence or override human decision-making. Additionally, this modeling method uses safety scores, anomaly detection, and checks for ethical constraints to ensure that even adaptive agents stay predictable and accountable. By using this framework, healthcare institutions can monitor AI behavior across different patient cases, ensure its decisions align with medical standards, and avoid situations where agentic misuse, like unauthorized actions, misleading reasoning, or harmful treatment suggestions, could threaten patient safety. This risk modeling approach serves as the basis for the SCAMA framework, providing measurable indicators of agent behavior, enabling ongoing oversight, early detection of misuse patterns, and a clear path to safer, ethically sound AI-powered healthcare systems.

4.2 Behavior Monitoring and Anomaly Detection

The Behavior Monitoring and Anomaly Detection module in SCAMA continuously observes how the AI agent behaves during decision-making. While intent monitoring checks what the agent plans to do, this layer checks how it acts. It compares the agent's current behavior with normal, expected patterns. If the agent performs actions faster, slower, or in unusual sequences, or behaves differently from its trained model, these deviations are flagged as anomalies. The module uses simple statistical checks, pattern comparison, and lightweight machine learning models to identify unusual behaviors. Examples include repeated attempts to perform restricted actions, unexpected action combinations, or sudden changes in reasoning steps. When the system detects abnormal behavior, it increases the risk level and sends the event to the Action Validation module. If the deviation is serious, SCAMA blocks the action or requests human oversight. This act as an early-warning system, helping SCAMA stop unsafe or unintended behavior before it leads to agentic abuse.

4.3 Safe Reinforcement Learning for Clinical Decision Agent

Safe Reinforcement Learning (Safe-RL) ensures that AI agents used in clinical decision-making learn to behave safely while improving their performance. Unlike regular RL, which focuses on maximizing rewards, Safe-RL also stops harmful or unsafe actions during training and real-world use. To achieve this, the system uses reward shaping with clinical constraints. The agent only receives positive rewards when its actions follow medical guidelines and lead to safe outcomes. If the agent suggests something unsafe, like an incorrect dosage or an unapproved treatment, it gets a penalty. This teaches it to avoid those actions in the future. The training process includes action validation checkpoints. At these points, every proposed action is checked against medical rules and safety policies before acceptance. This prevents the agent from learning dangerous behavior. Human clinicians also provide feedback (RLHF). They reinforce safe, aligned decisions and correct mistakes. This combination helps the AI agent improve while staying medically safe, ethical, and trustworthy.

4.4 Controlled Execution and Validation

The Controlled Execution and Validation module serves as the final and most important safety barrier within the SCAMA framework. Even if earlier layers, like intent monitoring or behavioral analysis, miss a potential risk, this module makes sure that unsafe actions are never taken. It works as a real-time filter that assesses every action suggested by the AI agent before it engages with real patients, systems, or clinical workflows. The validation process starts when the agent chooses an action. Before the action is carried out, the module checks it against a set of predefined safety rules, medical guidelines, ethical limits, and specific policies for the field. These checks determine if the action is medically suitable, safe for the patient, and aligns with expected clinical procedures. Risk scores from earlier SCAMA layers are also included in the decision, helping the module spot actions that might look normal but carry hidden risks due to past issues or intent misalignment. If the action is deemed safe, it can proceed as planned. However, if the action seems unsure or potentially harmful, the module steps in by blocking it, changing it to a safer option, or sending it for human approval.

For instance, if an AI agent suggests a medication dosage beyond safe clinical limits, the module will automatically halt the action and seek clinician review. Furthermore, all validated or blocked actions are recorded for future reference. These logs assist in reviewing the agent's decision-making process, helping clinicians and developers understand why certain actions were blocked and how the agent's behavior changes over time. By combining rule-based validation, risk scoring, human oversight, and log tracking, the Controlled Execution and Validation module ensures that no unsafe or misaligned decisions reach real clinical settings. This layer ultimately ensures that the AI agent remains reliable, predictable, and follows medical safety standards.

IV. RESULT AND DISCUSSION

The SCAMA framework was tested using a mix of simulated agentic behavior datasets, synthetic adversarial prompts, and controlled healthcare decision-making settings. Performance was assessed across four main areas: intent misalignment detection, behavioral anomaly recognition, safe action gating reliability, and overall system latency. We compared the results to traditional rule-based monitoring systems and standard anomaly detectors.

A. Intent Misalignment Detection

SCAMA showed 94.8% accuracy in spotting misaligned or harmful agentic intentions. This result was much better than static rule-based systems, which averaged only 76.4% accuracy in similar tests. The improvement comes from SCAMA's semantic reasoning layer and contextual intent scoring mechanism. These components work together to assess language cues and decision patterns. These findings confirm that the framework is better at detecting early signs of intentional deviation or manipulation before harmful recommendations are made.

B. Behavioral Anomaly Detection

The anomaly detection module achieved a precision of 92.3%. This shows it can effectively differentiate between harmless deviations and real high-risk behaviors. SCAMA also demonstrated a significant drop in false positives. This improvement mainly comes from its learning system, which adjusts thresholds based on the specific behaviors of each agent. By doing so, it avoids blocking legitimate decisions. This makes the framework suitable for ongoing use in sensitive clinical settings.

C. Action Validation Reliability

The safe action gating component achieved 96.1% reliability in stopping unsafe or unethical agent actions. It blocks unauthorized medical advice, prevents unverified procedural suggestions, and stops attempts to bypass safety rules. This high reliability score shows the system's ability to maintain strict control over autonomous outputs without affecting normal clinical support functions.

D. System Performance

The SCAMA framework had an average processing latency of 210 ms, which is well within acceptable limits for non-real-time healthcare safety systems. Although real-time intensive care applications may need lower latency, this performance is adequate for diagnostic support tools, decision reasoning audits, and offline validation modules where completeness and accuracy are prioritized over speed.

E. Discussion

Overall, SCAMA showed significant improvements in preventing agentic abuse when compared to traditional monitoring systems. Its integration of intent analysis, behavioral anomaly detection, and safety-constrained action validation led to a 63% reduction in unsafe actions across all evaluated scenarios. The framework's multi-layered safeguard approach makes sure that if one detection method fails, the others still maintain safety boundaries. Additionally, SCAMA's adaptive behavioral modeling decreases false alarms and improves long-term reliability as AI agents evolve. These results confirm that SCAMA offers a strong, scalable, and ethically aligned solution for reducing agentic risks in AI-driven healthcare settings.

V. CONCLUSION

This research presents a multi-layered safety framework, SCAMA, designed to prevent agentic abuse in AI-driven healthcare decision support systems. The proposed system integrates autonomy constraints, zero-trust security, patient-specific consent control, anomaly detection, and explain ability-based accountability. Experimental results demonstrate high accuracy in detecting unsafe agentic behaviors, significant reduction of unauthorized actions, and improved transparency in clinical decision-making. The framework strengthens trustworthiness and regulatory compliance for healthcare AI deployments. It ensures that autonomous AI systems operate safely, ethically, and under continuous oversight, reducing the risk of unintended or malicious agentic actions. Future work will incorporate federated learning, real-time multi-agent coordination, and adversarial robustness testing to further advance agentic safety in healthcare ecosystems.

REFERENCES

1. H.Thompson, R.Vale, M.Ortega, and L.Harris, "Anomaly detection techniques for safety-critical AI," *IEEE Intelligent Systems*, vol. 38, no. 1, pp. 82–94, 2024.
2. J.Smith and M.Alvarado, "Preventing autonomous agent misbehavior in AI systems," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 2, pp. 145–158, 2023.
3. K.Lee and A.Banerjee, "Ethical constraints in autonomous decision systems," *IEEE Access*, vol. 11, pp. 10245–10258, 2023.
4. P.Rossi and T.Nguyen, "Intent drift in autonomous agents," *Journal of AI Safety*, vol. 2, pp. 88–104, 2023.
5. R.Patel, S.Wang, and L. Rizzi, "A survey on agentic abuse and safety constraints in AI," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–35, 2022.