# Georgia Tech develops AI Psychiatry to reconstruct AI failures from memory snapshots



Artificial Intelligence (AI) systems are increasingly integrated into everyday technologies, ranging from drones delivering medical supplies to sophisticated digital assistants. These systems, however, are not infallible and can experience significant failures due to design flaws, biased training data, or security vulnerabilities. Investigating the causes of such failures, especially after critical incidents, poses a considerable challenge due to the opacity and complexity of AI models.

Researchers at the Georgia Institute of Technology have developed a novel system called AI Psychiatry (AIP), designed to improve forensic investigations of AI system failures. The system enables the recreation or "reanimation" of AI models from memory snapshots taken at the time of failure, allowing for systematic testing to determine what went wrong.

The necessity for such a tool is exemplified by incidents like those involving autonomous vehicles, where an AI might suddenly misinterpret sensor data and cause a crash. Investigators must discern whether a failure was caused by a hardware fault, such as a malfunctioning camera, or by a deeper issue within the AI itself, potentially including malicious attacks exploiting software vulnerabilities.

Traditional forensic methods for cyber-physical systems like drones or autonomous cars fall short when it comes to analyzing AI-specific failures, partly due to the dynamic and adaptable nature of advanced AI models. AI Psychiatry addresses these limitations by isolating the data behind the AI’s decision-making processes from a memory image captured during operation and reconstructing a fully functional model identical to the one in use at the time of failure.

In a series of tests involving 30 AI models, 24 of which contained intentional backdoors triggering incorrect behaviours under specific conditions, AI Psychiatry successfully recovered, rehosted, and tested every model. These tests included models utilised for real-world applications such as street sign recognition in autonomous vehicles. The system's capacity to identify vulnerabilities allows investigators to either confirm AI-related causes or rule out AI faults in accident investigations.

AI Psychiatry’s methodology is broadly applicable as it focuses on fundamental components common to all AI models developed using popular AI frameworks. This versatility means the system can be employed to examine various AI applications, from product recommendation bots to autonomous drone fleets.

Additionally, AI Psychiatry has potential uses beyond post-failure investigations. It offers a standardized forensic approach that can be valuable for auditing AI systems proactively, a practice that is gaining traction as government agencies and other institutions increasingly adopt AI technologies. The open-source nature of AI Psychiatry ensures that investigators and auditors have unrestricted access to this investigative tool.

Brendan Saltaformaggio, Associate Professor of Cybersecurity and Privacy at Georgia Institute of Technology, and David Oygenblik, a Ph.D. student in Electrical and Computer Engineering at the same institution, highlight that AI Psychiatry could bring significant benefits to AI creators and users alike by enabling clearer insight into AI failures and vulnerabilities.

The biloxinewsevents.com is reporting on this advancement, emphasising how AI Psychiatry represents a key innovation in the evolving field of digital forensics for artificial intelligence systems.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://ece.gatech.edu/news/2023/12/gtri-georgia-tech-develop-ai-psychiatry-advance-national-security> - This article discusses the development of AI Psychiatry by Georgia Tech researchers to enhance forensic investigations of AI system failures, including applications in autonomous vehicles and drones.
2. <https://news.gatech.edu/news/2024/10/29/sherlock-holmes-ai> - This piece highlights AI Psychiatry's role in recovering exact models from compromised machines to identify errors, emphasizing its application in self-driving cars and AI systems.
3. <https://www.researchgate.net/publication/378804903_Beyond_Discrimination_Generative_AI_Applications_and_Ethical_Challenges_in_Forensic_Psychiatry> - This study examines the ethical challenges of generative AI in forensic psychiatry, including issues like racial and gender bias in AI forensic risk assessment algorithms.
4. <https://www.frontiersin.org/articles/10.3389/fpsyt.2023.1209862/full> - This article discusses external challenges in using AI in forensic psychiatry, highlighting concerns about fairness, bias, and the potential for AI to reinforce existing social biases.
5. <https://www.frontiersin.org/articles/10.3389/fpsyt.2024.1435219/full> - This editorial addresses the applications of AI in forensic mental health, focusing on opportunities and challenges, including ethical concerns related to fairness and transparency.
6. <https://abpp.org/newsletter-post/implications-for-artificial-intelligence-in-forensic-psychological-practice-and-board-certification/> - This article discusses the implications of AI in forensic psychological practice, highlighting concerns about AI's limitations and potential biases in forensic settings.
7. <https://news.google.com/rss/articles/CBMixAFBVV95cUxONS1vcXJxV2FOcFFaU0ZYVThxQWdsWnJIYlBHMGs3UW9xMVg1Y2FyaTBfSzVMTUVyREtSWklBZ0FjV1pFQVNMYjVFTk1QeVlkVmR6ZV9tWTg1dmdmcEZvVlhyNk53X18xMjNtNXg3Z1B0NDczd2RBbDE2RzBwVnk2Rl9vcGlPSFhwOGRlZUJiOW1XdkR1VDNNSXJKNmF2Z1JyTXJfRkRQdGZOd2UyVlFiaTU5NTctUExwUHlYUklHSFhIRXYx?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data