# Georgia Tech unveils AI Psychiatry to 'reanimate' and audit failing AI models



Artificial intelligence (AI) is becoming increasingly integrated into daily life, with applications ranging from autonomous vehicles to drones delivering medical supplies, as well as digital assistants. These systems promise significant advances, but they are not infallible and can fail due to design flaws, biases in training data, or software vulnerabilities that may be exploited by hackers.

One challenge facing investigators and engineers alike is that AI systems are often opaque, making it difficult to determine the root cause of a failure or attack. This difficulty is compounded when investigating proprietary AI, where access to internal data is restricted, hampering forensic efforts.

Researchers at the Georgia Institute of Technology have developed a novel forensic tool named AI Psychiatry (AIP) designed to address these challenges by effectively “reanimating” failed AI models. This system reconstructs the AI model from a memory image—a snapshot of its internal state during operation—which can then be tested rigorously in a controlled environment for vulnerabilities or errant behaviour.

In a typical scenario, such as an autonomous vehicle crash precipitated by the car unexpectedly swerving off the road, investigators must determine whether the failure was caused by a sensor malfunction, a software bug, or an external malicious attack. Traditional investigative methods are often insufficient for fully examining the AI components, especially since advanced AI models can update themselves continuously, complicating efforts to capture the relevant state at the time of failure.

AI Psychiatry's forensic process involves extracting data representing the AI’s decision-making model and reconstructing it to behave identically to the original. This enables security analysts or investigators to input potentially malicious data and observe the AI’s responses, thereby uncovering hidden behaviours or backdoors.

The research team evaluated AI Psychiatry on 30 distinct AI models, including 24 that had deliberately embedded backdoors designed to produce incorrect outcomes upon specific triggers. The tool successfully recovered, rehosted, and tested every model, including those for tasks relevant to real-world applications like street sign recognition in autonomous vehicles.

Beyond autonomous transport, AI Psychiatry's underlying algorithm is framework-agnostic, making it adaptable to any AI system built using popular development platforms. This expands its utility to various AI-powered applications, such as recommendation systems or autonomous drone operations. Crucially, the tool is open source, facilitating broader access for investigators and auditors.

The researchers also highlight the potential of AI Psychiatry as a preventive audit tool to ensure AI systems operate correctly before deployment or prior to problem detection. With governmental agencies integrating AI into critical workflows—ranging from law enforcement to child protective services—standardised auditing methods become imperative. AI Psychiatry offers a consistent methodology that can be applied across different AI implementations to verify safety and reliability.

By providing a means to systematically diagnose AI failures and potentially expose security breaches, AI Psychiatry represents a significant advancement in the emerging field of AI forensics, promising benefits to both AI developers and users affected by AI-driven processes.

This report is based on research and insights shared by David Oygenblik, Ph.D. Student in Electrical and Computer Engineering, and Brendan Saltaformaggio, Associate Professor of Cybersecurity and Privacy at the Georgia Institute of Technology. The biloxinewsevents.com is reporting.

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

## Bibliography

1. <https://www.gtri.gatech.edu/newsroom/gtri-georgia-tech-develop-ai-psychiatry-advance-national-security> - This article discusses the development of AI Psychiatry by the Georgia Tech Research Institute (GTRI) to enhance national security by addressing AI system vulnerabilities.
2. <https://news.gatech.edu/news/2024/10/29/sherlock-holmes-ai> - This piece highlights AI Psychiatry's role in recovering compromised deep-learning models to understand AI system failures.
3. <https://www.forensicmag.com/578810-New-Branch-of-Cyber-Forensics-to-Advance-National-Security/> - This article covers the collaboration between GTRI and Georgia Tech to develop AI Psychiatry, a new branch of cyber forensics aimed at enhancing national security.
4. <https://cyfi.ece.gatech.edu/publications/> - This page lists publications from the Cyber Forensics Innovation Laboratory (CyFI Lab) at Georgia Tech, including research on AI Psychiatry and its applications in cyber forensics.
5. <https://www.azolifesciences.com/news/20250305/AI-powered-Forensic-Tool-Predicts-Traumatic-Brain-Injuries-With-9425-Accuracy.aspx> - This article discusses the development of an AI-powered forensic tool that predicts traumatic brain injuries with high accuracy, showcasing AI's growing role in forensic assessments.
6. <https://arxiv.org/abs/2503.21003> - This research presents a novel approach to forensic detection and source attribution of AI-generated images, highlighting advancements in AI forensics.
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