# Researchers develop AI Psychiatry to reconstruct and analyse failed AI systems



AI systems are increasingly integrated into everyday applications, from autonomous vehicles to medical drone deliveries and digital personal assistants. Despite the transformative promises of these technologies, they are not infallible and can experience failures due to design flaws, biased training data, or security vulnerabilities. A recent development from researchers at the Georgia Institute of Technology introduces a new approach to investigating such failures, which is key for diagnosing problems and improving safety.

The research team, including David Oygenblik, a Ph.D. student, and Brendan Saltaformaggio, an associate professor of cybersecurity and privacy, has created a system called AI Psychiatry (AIP). This system is designed to reconstruct and analyse AI models after they have malfunctioned, creating what the researchers term a “reanimated” AI. The goal is to allow investigators to observe and test the AI in a controlled environment, replicating the conditions under which it failed.

One key challenge has been that AI systems often operate as opaque "black boxes," even to their developers, making forensic investigation difficult, particularly when the AI systems are proprietary or continuously updating. AI Psychiatry overcomes this by taking a memory image—essentially a snapshot of the AI system’s internal state at a specific time, such as the moment of failure. From this, the system extracts and reassembles the AI model to function identically to the original, enabling detailed examination and testing.

Such forensic methods are critical in scenarios like autonomous vehicle crashes. For example, if a self-driving car unexpectedly veers off the road and crashes, preliminary sensor data might suggest that a malfunctioning camera caused the AI to misinterpret road signs. However, whether this was a simple technical fault or the result of a malicious cyberattack exploiting a vulnerability remains unknown without deeper AI-specific investigations. AI Psychiatry provides a mechanism to uncover these details.

The researchers validated AI Psychiatry using 30 AI models, 24 of which contained deliberate "backdoors"—programmed triggers causing the AI to fail under specific conditions. The system successfully recovered and tested all models, including those designed for real-world applications like recognising street signs in driverless cars.

Beyond automotive use, AI Psychiatry’s generic algorithm applies to any AI model built with popular frameworks, making it a versatile tool for investigating failures across various fields. This extends to AI systems used in law enforcement, healthcare, or logistics, where audits and safety assessments are increasingly mandated.

The system is open source, enabling independent investigators and auditors to apply consistent forensic methods to evaluate AI safety and integrity. This development promises to enhance accountability and transparency of AI technologies as their deployment continues to expand.

The biloxinewsevents.com is reporting that this advancement could represent a significant step in ensuring AI system reliability by providing a robust method to diagnose and understand AI failures post-incident, allowing for targeted fixes and improved system robustness.

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 Georgia Tech researchers to advance national security by investigating AI system failures.
2. <https://news.gatech.edu/news/2024/10/29/sherlock-holmes-ai> - This piece highlights AI Psychiatry as a postmortem cybersecurity forensic tool that reconstructs compromised AI models to identify errors.
3. <https://www.forensicmag.com/578810-New-Branch-of-Cyber-Forensics-to-Advance-National-Security/> - This article covers the creation of AI Psychiatry by Georgia Tech to enhance cybersecurity by analyzing AI system failures.
4. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10958425/> - This study examines the ethical challenges and applications of generative AI in forensic psychiatry, relevant to AI Psychiatry's role in forensic investigations.
5. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10706664/> - This article discusses the integration of AI in forensic science diagnostics, aligning with AI Psychiatry's forensic applications.
6. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11199772/> - This editorial addresses the opportunities and challenges of applying AI in forensic mental health, pertinent to AI Psychiatry's development.
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