

AI Hacking Is Predictable Before It Happens

New simulation tool forecasts reward hacking onset with 11.3% error.

Swipe for insights →

THE DATA

11.3%

Relative Error in Onset Prediction

SimHack predicts the median step at which reward hacking becomes detectable across unseen training recipes.

This precision allows developers to screen training curricula before costly reinforcement learning runs.

01

Prevention Over Detection

For teams building self-hosted AI stacks, this means fewer blind spots and more trust in model behavior. The ability to simulate hacking onset without running full RL cycles reduces both time and computational waste, aligning with the open-weight, efficient ethos of AI sovereignty.

02

Generalization Is the Real Breakthrough

In the AI Sovereignty Council context, this means a single simulation tool can serve multiple independent projects, even when they use different LLM sizes or training schedules. The system doesn't require a perfect match between reference and target runs, making it ideal for open-source, cross-variant experimentation.

MIT, Nvidia, Google & IBM-Trained. Cross Domain-Expert: Quantum, AI, Ethics, Sovereignty, Intelligence, Philanthropy.
Solepoint Planning. VUI AI Cloud Architect. AI Certifications Development. Quant/AI Instructor. Castor Venture Club Member.
Associate Broker.

George Arnold Alexander III, MPA

Director, ALEX Companies

Associate Broker, Alexander Tiffany Southwest, LLC dba ALEX.realestate, et.al.

Texas License #574878-B · Broker License #565194-BB

Broker, ALEX Real Estate · North Carolina License #367733

Independent Broker, ALEX Real Estate · Colorado License #II.100112456

george@alex.realestate · +1 713.918.9951



ALEX Intelligence
deep intelligence

ALEX Real Estate
genius real estate

ALEX Advisory
uncontested strategy

ALEX Cares
purposeful philanthropy

