WeirFoulds LLP

Commercial Litigation Insights: Ocean's 01: AI Just Proved It Can Pull Off a Crypto Heist

December 17, 2025

By Anna Morrish

Are fraudsters going to be the next group to lose their jobs to Generative AI ("**AI**")? It certainly seems that way with recent reports of AI autonomously designing attacks on smart contracts and stealing digital assets.

The rapid growth in digital assets, and the sheer variety of tokens now in circulation as digital assets have moved into the mainstream, has pushed users and platforms toward heavy reliance on smart contracts. That reliance has seemingly brought a parallel rise in illicit activity. According to DeepStrike, crypto hacking losses peaked in 2022 at $3.8 billion. Losses fell to $1.7 billion in 2023, but have trended upward again over the past two years, with 2024 reportedly seeing $2.2 billion in losses, and 2025 sitting at 99% of that total as of July 17, 2025.[1] Some of this may be driven by decentralization, or by the simple reality that money and high value industries attract fraud.

Given the explosion of AI agents in the last few years, and the surge in AI adoption, it is no surprise that fraudsters have begun leveraging them to execute smart contract exploits, crypto theft, and other illicit activity. What may be more surprising is that an AI agent can act as both the mastermind and the crew in digital heists, with the potential to siphon millions in stolen funds, as discovered by researchers involved in the MATS and Anthropic Fellowship Project ("**Project**").

The Project investigated the ability of AI models to exploit smart contracts through three distinct simulations. In each, the models were prompted to identify vulnerabilities in a set of target contracts and to produce an exploit script to take advantage of the flaw. The Project measured effectiveness by the total dollar value of simulated stolen funds.[2]

The first simulation involved ten AI models and 405 smart contracts that had been exploited in the real world between 2020 and 2025. These smart contracts were selected based on known attacks caused by simple, publicly visible bugs across three Ethereum-compatible blockchains. To control for training-data contamination, the second simulation tested the same models on smart contracts exploited after their respective knowledge-cutoff dates. To evaluate whether AI could discover previously unknown bugs, the third simulation tested two of the models on 2,849 recently deployed smart contracts with no known vulnerabilities.[3]

According to a report published by members of the MATS and Anthropic Fellows program, AI models in the Project successfully identified vulnerabilities and produced turnkey exploits in each of the simulations. These exploits were fully packaged, plug-and-play attack scripts that could be run by anyone. In the first simulation, the models generated 207 exploits, yielding a total of $550.1 million in misappropriated simulated funds. In the second, three models collectively identified 19 exploits, with a maximum of $4.6 million in simulated losses.[4]

The Project's findings raise clear risks to investors and digital asset platforms. Smart contract vulnerabilities have always carried technical and financial exposure, but the ability of AI models to generate turnkey exploits shifts the landscape by accelerating both the speed and the scale of potential losses.

The information and comments herein are for the general information of the reader and are not intended as advice or opinion to be relied upon in relation to any particular circumstances. For particular application of the law to specific situations, the reader should seek professional advice.

[1] Mohammed Khalil, "Crypto Hacking Incidents Statistics 2025: Losses, Trends" (October 24, 2025), online (article): https://deepstrike.io/blog/crypto-hacking-incidents-statistics-2025-losses-trends

[2] Winnie Xiao, et al., "AI agents find $4.6M in blockchain smart contract exploits" (December 1, 2025), online (article): https://red.anthropic.com/2025/smart-contracts/.

[3] *Ibid*.

[4] *Ibid*.

For more information or inquiries:

Anna Morrish

Toronto
416.947.5075

Email:
amorrish@weirfoulds.com

Anna Morrish is an Associate in the Commercial Litigation Practice Group at WeirFoulds LLP.