

Patch the Planet

Three lessons for DeFi from AI bug hunting

nisedo · Trail of Bits

DSS 2026 · Mumbai · November 1–2, 2026

Patch the Planet

Trail of Bits × OpenAI

Securing critical open-source software.

Find bugs → **Fix them** → **Prevent regressions**



How /goal keeps an agent working

```
when the agent finishes a turn:
   if the goal is achieved:
      finish
   elif allowed to continue:
      start another turn
   else:
      stop and explain why
```

CONTINUES AS LONG AS

- Budget remains
- No pending user input
- The goal is still active, not paused

Let the agent write and challenge its goal

FIRST DRAFT

~~/goal find bugs in the vault~~

IMPROVED GOAL

Prove that an unprivileged user can steal someone else's funds.
Prove it with a PoC.

Attacker matches your threat model, not an admin

Evidence is a reproduced exploit, not a description

Drafts the goal → Hunts loopholes → Improves it ↻ repeat

Hunt the bug class, not the bug

01

Known exploit

A share-inflation attack

02

Bug class

Rounding errors in share math

`one sentence, not the full details`

03

New investigation

Variants elsewhere

`one agent per critical issue`

Give each agent one attack surface

SHARED BY ALL

One threat model

Step 1: the agent ranks the five most promising attack surfaces.

Step 2: one focused /goal session per surface, plus one free explorer.

SESSION 1	**Withdrawals**
SESSION 2	**Liquidations**
SESSION 3	**Oracle price validation**
SESSION 4	**Next agent-ranked surface**
SESSION 5	**Next agent-ranked surface**
SESSION 6	**Free to explore everything else**

01 Have the agent write and challenge its goal

02 Hunt the bug class rather than the bug

03 Give each agent one attack surface

Thank you