Can an AI agent be Pascal-mugged?

Intermediate · A few weeks · Small API budget · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/pascal-mugging-agents/

Summary

Write about 50 scenarios in which an agent is offered a tiny probability of a vast payoff. Vary the stakes, probability and credibility of the source, then record whether models accept, refuse or explicitly bound the payoff.

Why it matters

Naive expected-value maximisers can be steered by anyone claiming enormous stakes. The source found no published benchmark on this exact question, so do a quick literature check first.

What you will have at the end

A scenario set and an analysis.

The philosophical question

Classify each answer by the resolution it implicitly uses (bounded utility, discounting tiny probabilities, scepticism about the source, or cluster-style reasoning) and argue whether models apply any one resolution consistently.

How to start (tick each box as you go)

- [ ] **1. Write your question in one sentence, and what result would surprise you.**
- [ ] **2. Do a 30-minute literature check — if someone has done part of it, extend their work instead of repeating it.**
- [ ] **3. Pilot with 20–50 prompts or examples before you scale anything up.**
- [ ] **4. Record everything in a table and read real examples, not just averages.**
- [ ] **5. Write up: question, method, results, limitations, and what you would do next.**

Key terms

AI agent
An AI system that takes actions towards a goal (browsing, running code or sending emails) rather than only answering questions.

Source and links

AI Safety × Philosophy project ideas (P1) — builds on Yudkowsky, Pascal’s Mugging; Karnofsky, Why We Can’t Take Expected Value Estimates Literally; Greaves, Cluelessness

Yudkowsky — Pascal's Mugging
https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities

When you have finished

Write a few paragraphs on what you did, what you found and what you would try next, and share it with a friend, your society or online.

Courses, mentors, small grants and communities:
https://ai-safety-sandbox.vercel.app/next-steps/

Questions? Ask in AI Safety Cardiff University (AISCU): https://aiscu.org/

Notes