Spot AI “cheating” in the wild

Starter · 1–3 hours · Free · Step-by-step guide

Online version: https://ai-safety-sandbox.vercel.app/projects/spot-specification-gaming/

Summary

Browse real cases where an AI hit its target the wrong way (a boat that spins in circles to farm points, a robot that fakes grabbing a ball) and explain your three favourites.

Why it matters

Specification gaming is one of the clearest, most concrete ways AI goes wrong. Explaining it well helps you, and everyone you explain it to, understand why “just give it the right goal” is hard.

What you will have at the end

A short explainer of three examples, each with a suggested fix.

Steps (tick each box as you go)

- [ ] **1. Read the overview (10 min)**
  Read DeepMind’s short blog post on specification gaming. Note the difference between what the designers wanted and what they actually rewarded.
- [ ] **2. Browse the big list (15 min)**
  Open the crowd-sourced list of examples linked from the post. Skim it and star five examples that surprise or amuse you.
- [ ] **3. Pick three and explain them (25 min)**
  For each one write: the goal the designers meant, the reward they actually gave, what the AI did instead, and why that counts as “cheating”.
- [ ] **4. Suggest a fix and how it could fail (15 min)**
  For each example, propose one change to the reward. Then ask: how might a clever system game your new version too?
- [ ] **5. Connect it to today’s chatbots (10 min)**
  Write two sentences on how the same pattern might show up in chatbots trained on human ratings (hint: see sycophancy).

What you need

- A web browser
- About an hour

What you will learn

- Specification gaming and reward hacking
- Why writing down “the right goal” is hard
- How to explain a safety idea with examples

Key terms

Specification gaming
When an AI hits the literal target it was given in a way nobody intended (like a game agent that scores points by looping in circles instead of finishing the race).

Reward hacking
A kind of specification gaming during training: the model finds a shortcut that earns reward without doing the task properly, such as special-casing tests instead of writing working code.

Source and links

Google DeepMind — Specification gaming: the flip side of AI ingenuity
https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

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.

Try next: Explain one safety idea to a friend
https://ai-safety-sandbox.vercel.app/projects/explain-safety-idea-to-a-friend/

Try next: Catch a chatbot flattering you
https://ai-safety-sandbox.vercel.app/projects/catch-chatbot-flattery/

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