

Overview

Watched a game theory video.

Decided I wanted a version where the opponent learns in real time.

So this is a fast tabular Q-learning implementation of the Iterated Prisoner's Dilemma with a human in the loop and a UI that shows exactly how your bad decisions affect the policy, observable in 30 rounds.

Under the hood

Agent:

- Tabular $Q(s,a)$
- ϵ -greedy action selection
- exponential ϵ decay
- discounted future reward ($\gamma = 0.95$)

State space:

START

CC CD

DC DD

Which is enough for:

- reciprocity
- punishment
- trust recovery

Pretraining:

The agent plays against itself for multiple short episodes to avoid starting from a zero-knowledge policy.

Yes, this pushes it toward Nash equilibrium behaviour.

Yes, you can drag it back to cooperation if you're patient.

Visual analytics

The UI tracks:

- cooperation as a binary time series
- moving averages (behavioural trend)
- cumulative score differential
- per-round action tree

So you're not just playing; you're watching the policy move.

Behavioural observations

Even with minimal memory (one previous round), distinct strategies emerge:

- sustained cooperation if the human is consistent
- immediate retaliation after exploitation
- cautious re-cooperation after mutual defection cycles

Because the agent is pretrained in self-play, its default policy trends toward safe defection, but it will shift toward cooperative equilibria when the human provides a stable signal.

In other words, cooperation becomes a learned response, not a preset rule.