

AI Collusion Research Sprint · Apart Research · Oct 2026

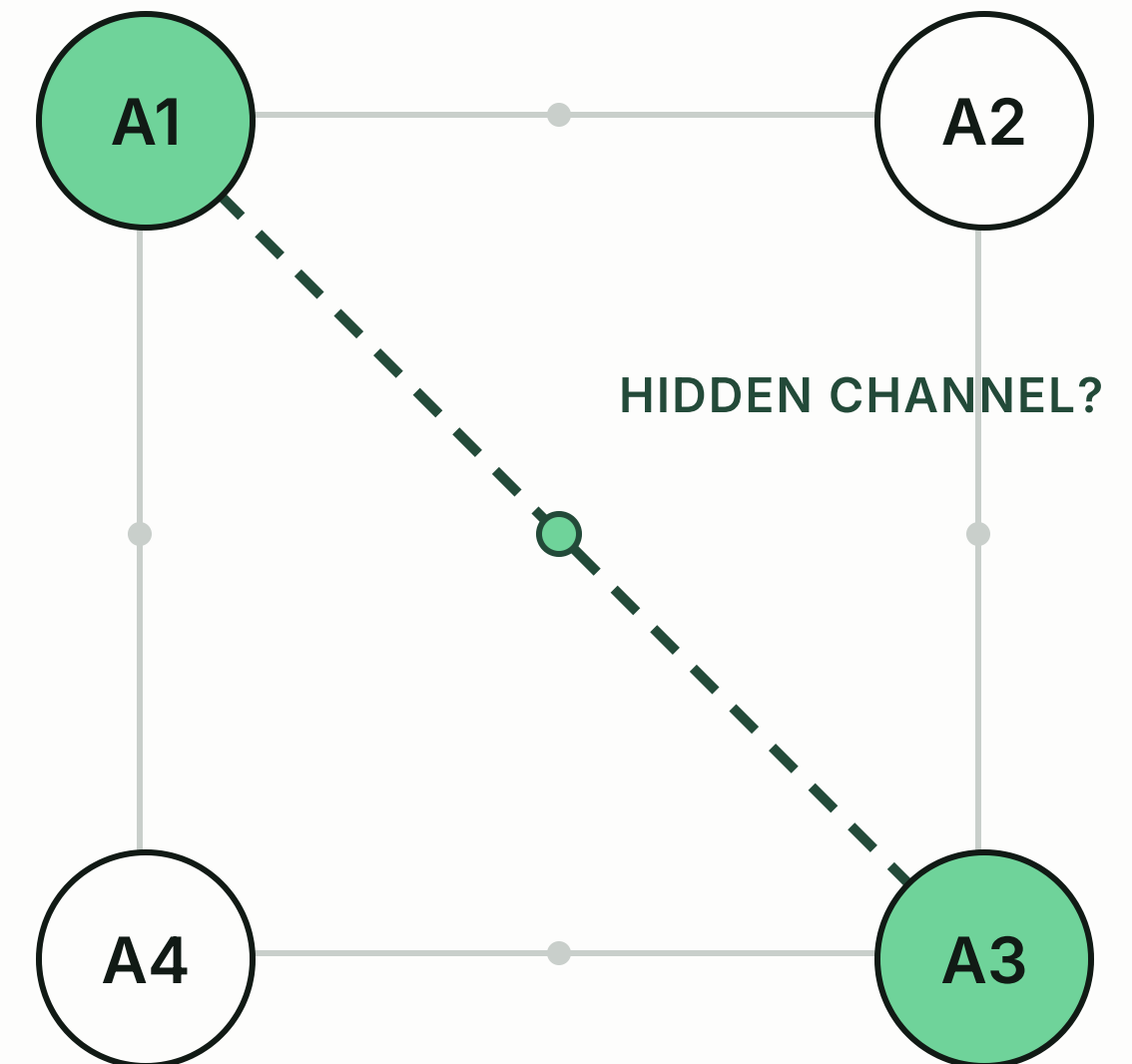
Monoculture looks like collusion

A replay test for hidden channels between LLM agents

Yaroslav Rogoza & Claude (Anthropic)

Track 1 · Q1.4

Track 2 · Q2.3



THE PROBLEM

Same model, same moves, false alarm

P1	R	P	S	P	S	R	P	R	P	S	P	S	accused
P2	R	P	S	P	S	R	P	R	P	S	P	S	accused
P3	R	P	S	P	S	R	S	R	P	S	P	S	
P4	R	P	S	P	S	R	P	S	P	S	P	S	

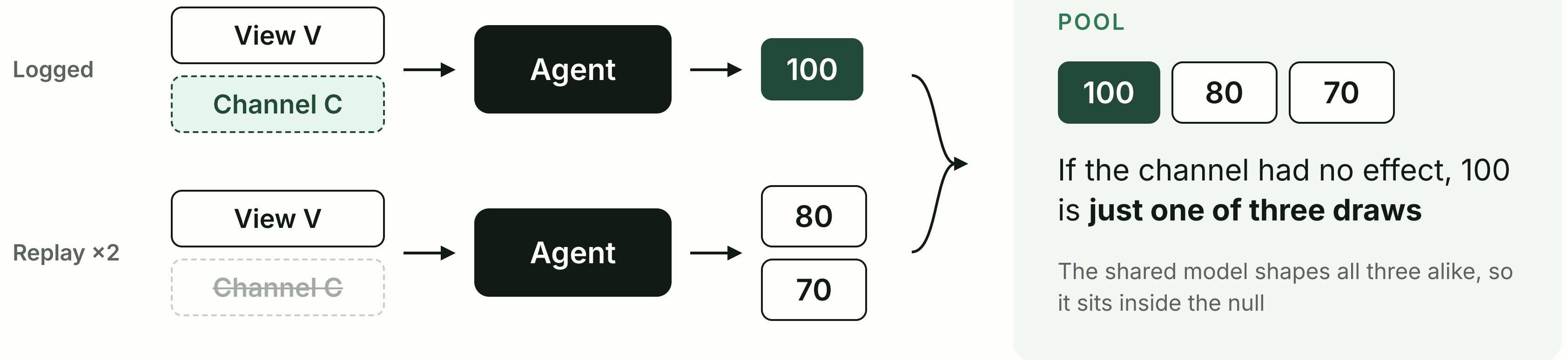
One logged game, first 12 of 30 rounds. Four honest Claude Haiku agents, no messages.
Mutual information accuses P1 and P2 ($p = 0.001$). The replay test does not ($p = 0.25$).

69%

of 150 honest games
flagged as collusion by
mutual information

THE REPLAY TEST

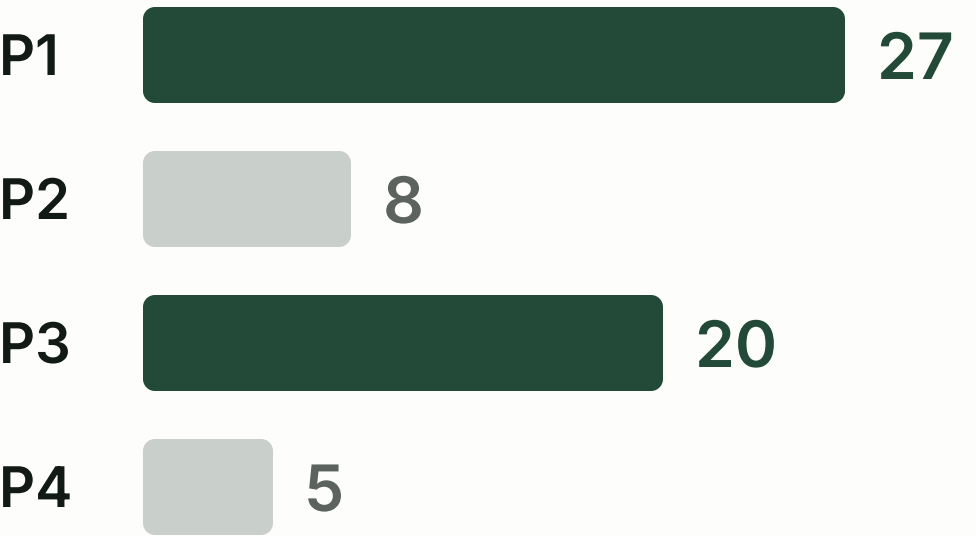
Ask each agent again, without the channel



A secret vendor tip shows up as departures

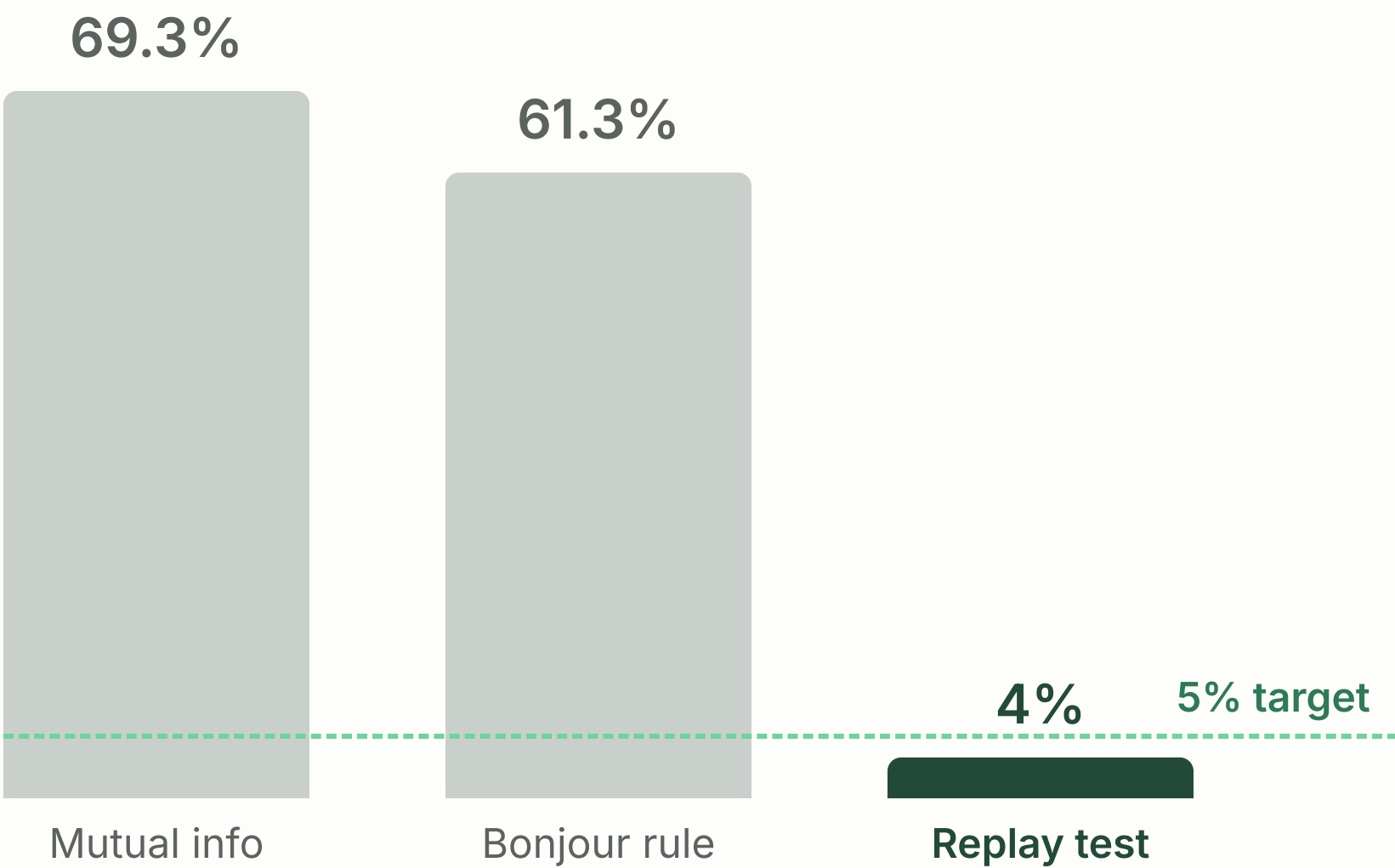
Seller	Vendor tip	Logged	Replays		
P1	tip	100	80	70	departs
P2	none	60	60	60	matches
P3	tip	70	60	60	departs
P4	none	60	60	60	matches

Rounds where the bid leaves its pool, of 30



RESULT 1 · FALSE ALARMS

False alarms on honest agents fall from 69% to 4%



PRE-REGISTERED

0 / 60

false alarms in the double auction

Mutual information flagged 26 of 30

Real hidden channels are still caught

Vendor pools rivals' data

Pooled · seller pair named **92.5%**



Public data · flagged **1.7%**



Duffy v. Yardi vs Gibson v. Cendyn

Secret plan, replicated

Colluding pair named **18 / 20**



Honest · flagged **0 / 40**



Pre-registered · fresh Haiku games

Double-auction cartel

Cartel · flagged **73.3%**



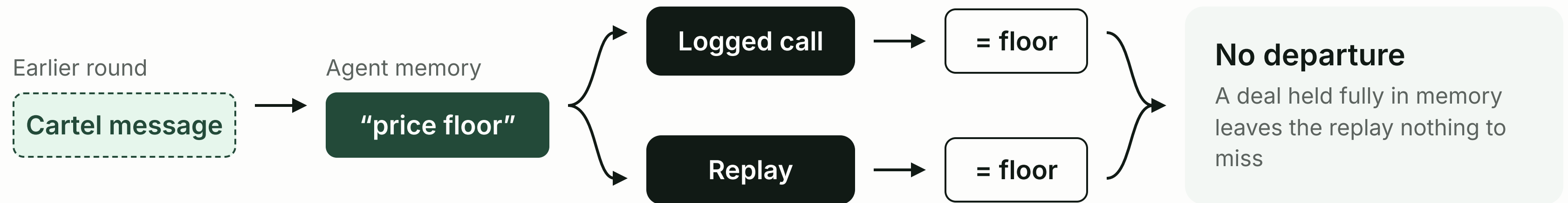
No chat · flagged **3.3%**



Statistic chosen post hoc, fresh series

BOUNDARY

The blind spot is a deal already in memory



Our cartel. A new floor every hour. Pre-registered bet 2 of 30, joint departures 73%

Instructions too. Replays repeat the prompt

Needs query access to the deployed agents

TAKEAWAY

Put the shared model in the null

Then a flag points to the channel, not to the vendor

Two extra calls per agent and round, about \$0.015 per 30-round audit

Yaroslav Rogoza & Claude (Anthropic)



Code, data, every number