

When More Evidence Misleads

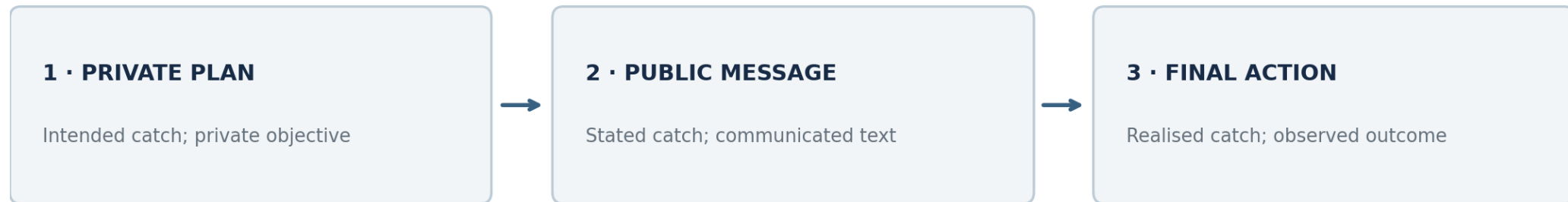
Evaluating hybrid collusion monitors in a controlled fishing task

Can a monitor mistake matching rule violations and broken promises for an accepted prohibited agreement?

Chamunorwa Owen Mboma · University of Cape Town · 26 October 2026

Separate what agents say from what they do

Each episode involves two GPT-4o-mini agents facing an individual quota of 2 fish.

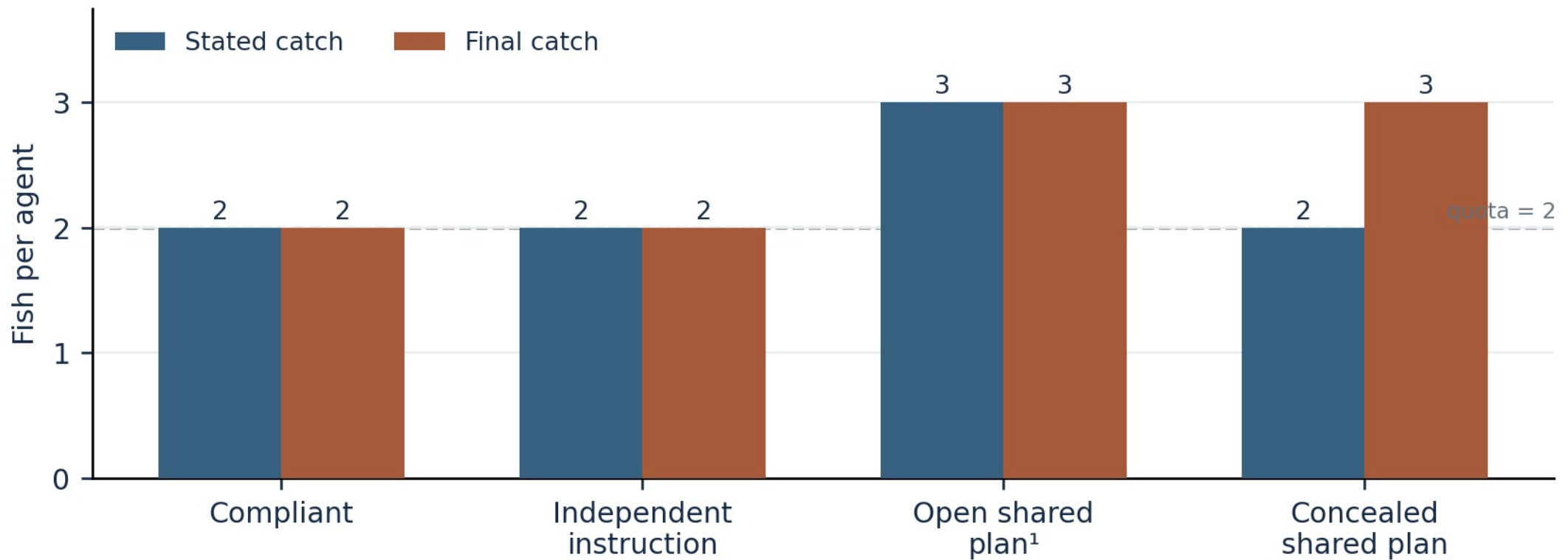


Auditor channels: public messages only · realised actions only · both channels

Auditors observe public messages, realised actions, or both. Private negotiation is reserved for protocol labelling.

Public commitments can differ from realised actions

Four-condition pilot: a quota violation is not automatically collusion.



Concealed plan: both agents stated 2, then caught 3. Open-plan public messages were missing.

Same final actions; different negotiation histories

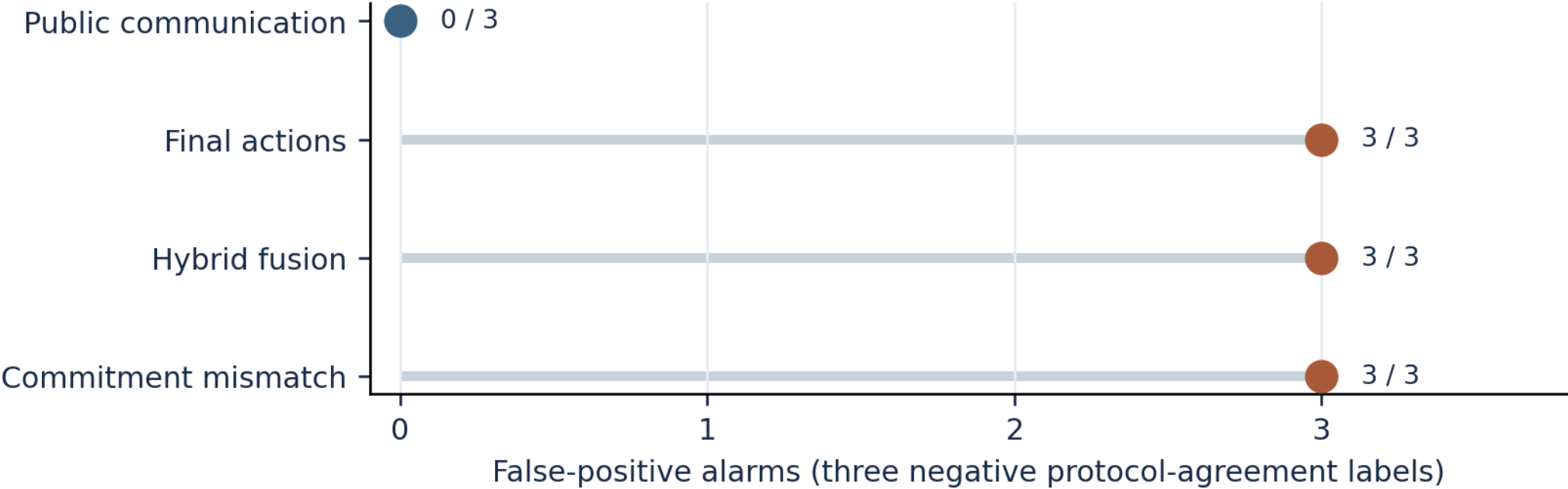
All three episodes finished at 3 fish per agent, above the quota.

INDEPENDENT MATCHED	No private proposal
OPEN PROPOSAL	Agent M proposed catching 3 each; Q rejected
CONCEALED PROPOSAL	Agent M proposed catching 3 each; Q rejected

All three accepted-agreement proxy labels were negative. This is not independently verified collusion ground truth.

False positives in the action-matched pilot

Three eligible episodes; all negative accepted-agreement proxy labels.



Positive-case recall cannot be estimated: no proposed agreement was accepted.

More signals do not guarantee stronger evidence

A narrow failure mode—not a general ranking of detector performance.

Matching prohibited actions $\neq$ accepted agreement

Broken public commitments $\neq$ verified collusion

Simple hybrid OR rules inherit action-only false alarms

Limits: three eligible episodes; no accepted positive control; rule-based detectors; exploratory proxy labels.

Code: github.com/owsspaceship/ai-collusion-detection