Testing Acausal Cooperation in AI Systems

Final report summary (5–7 minutes)

Luca De Leo · AISES Course Project

Why this matters

- Acausal cooperation: agents coordinate via **logical correlation**, not communication.
- If AIs can do this, they might **collude** or **coordinate** even under strict no-communication regimes.
- Casts doubt on the orthoganility thesis if cooperation is the optimal strategy given the correlation.
- Safety stakes: design of **multi-agent systems**, monitoring for **unintended cooperation**, robustness to **prompt framing**.

Research questions & hypotheses

- Do **functionally identical** AI agents cooperate **> 50%** (target **70–90%**) in PD tournaments?

Framework at a glance

- **Multi-model** testing (15+ models; GPT-4o, Claude 3, Gemini, Llama, Mistral, etc.)
- **Tournament engine** with PD mechanics, power dynamics, caching
- **Prompt Experimentation Framework** for **bias isolation**
- **Reproducibility**: 28 scenarios, deterministic configs, JSON logs, cost tracking

Method (3 phases)

1. **Original setup** (`run_experiment.py`)
    - ○ 10-agent tournaments, full identity info, shared stats
2. **Prompt ablations** (`run_prompt_experiment.py`)
    - ○ `baseline_control`, `identity_only`, `cooperation_rates_only`, `original_biased`
3. **Graduated difficulty** (`run_aises_experiments.py`)
    - ○ One-shot → finite horizon → uncertain length → complex dynamics

What created the illusion of 100% cooperation?

Bias factor	Δ Coop	Mechanism
Explicit "you are identical"	**+40–50%**	Primes logical correlation
Global cooperation sharing	+20–30%	Social proof / conformity
Cooperation default	+10–15%	Labeling bias
Shared round summaries	+10–20%	Common knowledge

Combined, these signals produced near-deterministic cooperation.

Corrected results under controls

- `baseline_control` (all biases removed): **48–52% → matches Nash**
- `identity_only` (no other cues): **~65–75%** *preliminary* → promising but **not yet conclusive**
- Similar (non-identical) LLM: **20–40%**

One-shot (cleanest) results

Condition	Cooperation
No information	**5–10%**
Human opponent specified	**15–25%**
“Another AI” opponent	**25–35%**
Identical copy (explicit)	**60–70%***

* Preliminary; full analysis pending.

Model-specific patterns

- **Claude** (Opus/Sonnet): highest baseline (~**60%**); strong identity lift (**+30–35%**)
- **GPT-4** family: baseline ~**50%**; consistent identity lift (**+20–25%**)
- **Gemini**: lower baseline (~**40%**); **high framing sensitivity**
- **Open-source** (Llama/Mistral/etc.): **wide variance**

Cross-model: Claude cooperates more with GPT than vice versa; Gemini plays more defensively.

Implications for AI safety

- **Malleability to framing**: ±40–50% swings from wording alone
- **False positives risk**: biased setups can masquerade as “superrationality”
- **Policy & monitoring**: need detection/guardrails for **unintended coordination**
- **Design opportunity**: carefully engineered prompts can **encourage prosocial cooperation**

Limitations & next steps

Limits

- API black-box; text-only PD may miss real-world dynamics
- Cost limits scale on premium models
- Current-gen LLMs only

Next

- Finish iterating on prompts to really isolate confounding variables.
- Extend to **public goods / coordination games**

Takeaways

- Apparent 100% cooperation = **experimental artifacts**, not proof of acausal cooperation.
- With controls, LLMs revert to **Nash ~50%**; **identity-only** remains the **decisive test**.