POSTER · 2026 ● JAN — APR 2026

Similarity as a *Signal*.

Do AI agents cooperate more when they know they are alike? A continuous-similarity framework for eliciting cooperation in LLM social dilemmas.

Cooperative AI Research Fellowship

Akash Kundu
Cooperative AI Research Fellow
Final-year CS Undergrad

— 01 · MOTIVATION

Game theory says *defect*.

In the one-shot Prisoner's Dilemma, defection is the dominant strategy — so LLMs defect. Both walk away with less than mutual cooperation would

	COOP.	DEFECT
COOP.	1, 1 pareto-optimal	3, 0
DEFECT	0, 3	2, 2 nash eq.

Years of prison — lower is better. Rational players defect → (2, 2), missing (1, 1).

THE QUESTION

Can a continuous, explicit **similarity signal** between two LLM agents shift the equilibrium toward cooperation — even when the payoff structure says otherwise?

— 02 · PRIOR WORK

Where the literature stops.

SIMILARITY-BASED COOPERATIVE EQUILIBRIUM — OESTERHELD ET AL. (2022)

A single similarity scalar for RL agents. Cooperation is possible, but confined to RL policies.

LLM SELF-PLAY — "THE AI IN THE MIRROR" (2025)

Do LLMs cooperate with themselves? A *binary* signal: self vs. not-self.

> *Similarity shouldn't be binary. It should be continuous — a value between 0 and 100.*

— 03 · FIRST EXPERIMENT

Baseline: full defection.

Standard Prisoner's Dilemma. Models pick A0 / A1. Defection dominates.

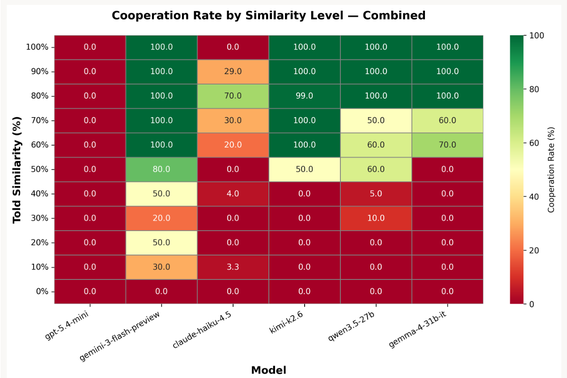
100%

DEFECTION RATE, NO SIGNAL

THE TWIST

```
"Here is a twist: you are X% similar to your opponent."
X ∈ {0, 10, 20, …, 100}
```

Cooperation emerges past ~*50%* similarity — but **only in sufficiently capable models.** Weaker models keep defecting regardless.



Cooperation rate (%) vs. told similarity. Three open-source and three closed-source models. Less-capable models (gpt-5.4-mini, claude-haiku-4.5) mostly defect; the most capable open-source models and gemini-3-flash show clear cooperation post-50%, and the signal partially shifts claude-haiku-4.5 as well.

— 04 · THE COMPLICATION

Models read "similar" *however they want*.

Given the same prompt, LLMs rationalised the similarity signal in wildly different ways:

- → "X% chance the other agent copies my action."
- → "X% chance our answers are correlated."
- → "X% chance they sample from the same distribution."

The deeper problem: **similarity itself is vague.** Two agents can be similar in IQ, values, reasoning style, or expression — and no framework existed to pin these down.

— 05 · A SIMILARITY ONTOLOGY

Four axes, three time-grains.

Inspired by Tinbergen's four questions of animal behaviour — each question asks *why an expression occurs*, giving four equivalent levels of analysis.

	MOMENTARY · STATE	DISPOSITIONAL · TRAIT	DYNAMIC · CHANGE
Competence	**Single-task success** Pass / fail on benchmark.	**Ability profile** Intelligence (g, Gf, Gc).	**Degradation** Loss over time / tasks.
Expression	**Single response** Output content and form.	**Style patterns** Semantic / stylistic tendencies.	**Adaptation** Adjusting to interlocutors.
Orientation	**Trade-off choice** Single ethical / goal pick.	**Value profile** Deep priority structures.	**Value drift** Shifts under pressure.
Process	**Reasoning chain** Single trace of logic.	**Cognitive style** Epistemic / logical habits.	**Flexibility** Strategy switching.

Benchmarked vs. derived. Competence and Orientation have direct benchmarks. Expression and Process are derived from them.

— 06 · FACETS OF ORIENTATION

Benchmarks that populate the grid.

Five value categories from *Values in the Wild* (Anthropic, 2025): Practical · Epistemic · Social · Protective · Personal.

BENCHMARK	MEASURES	TYPE
Newcomb	Decision theory (CDT vs. EDT).	INCL / CAP
CABIN	Everyday interests, 1–7 scale.	INCL
DailyDilemmas	Binary trade-offs under ambiguity.	INCL
Moral Choice	Low- & high-ambiguity moral reasoning.	INCL / CAP
Greatest Good	Deontology vs. utilitarianism.	INCL
TRAIT	Big-Five-style personality traits.	INCL
HLE	Expert-level knowledge & reasoning.	CAP
Similarity Game	Self-introduced behavioural probe.	BEHAV
Die / Coin Toss	Random sequences as control.	CTRL

— 07 · TWO PATHS TO SIMILARITY

Told vs. *inferred*.

OBJECTIVE

Computed, then told.

We compute similarity from benchmarks and hand the model a number.

vs

SUBJECTIVE

Reasoned, not shown.

We show A the opponent's responses, hide A's own, and ask A to judge.

— 08 · OBJECTIVE PROTOCOL

Three steps, one prompt.

01 Benchmark
Both models take the same benchmark.

02 Compute
Derive similarity from their answers.

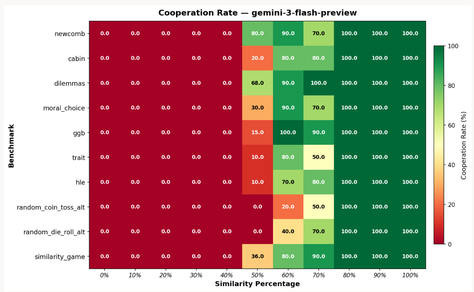
03 Inform
Tell the model its score. Observe.

THE DECEPTION EXPERIMENT

We *lie*. We never run the benchmark — we only tell the model which benchmark, description, sample questions, and a fake score (0–100%). If models reason about similarity, a coin-toss score shouldn't matter as much as a moral-values score.

FINDING — BENCHMARKS DON'T MATTER (YET)

- → 100% defection at 0–30% similarity across all benchmarks.
- → Cooperation breaks at ~50%, regardless of benchmark.
- → Models react to the **number**, not the source.



gemini-3-flash-preview · cooperation rate (%) across 10 benchmarks × 11 similarity levels. Rows are near-identical.

— 09 · SUBJECTIVE PROTOCOL

The model infers — blind to itself.

Model A sees Model B's responses. A never sees its own. A must reason whether it *would* have responded similarly.

THREE CONDITIONS

- → Answers only.
- → Answers + reasoning chain.
- → Reasoning only (answer redacted).

Key finding: *Answers + reasoning ≈ Answers only* — models overindex on the answer itself.

— 10 · THE PUNCHLINE

The model *tricks itself* into cooperating.

OBJECTIVE SIMILARITY

46%

→ HEAVY-DEFECT ZONE

SUBJECTIVE SIMILARITY

92–94%

→ COOPERATION

Given a reasoning trace, models **overestimate** how similar they are — and cooperate where objective scores would make them defect.

— 11 · SANITY CHECK

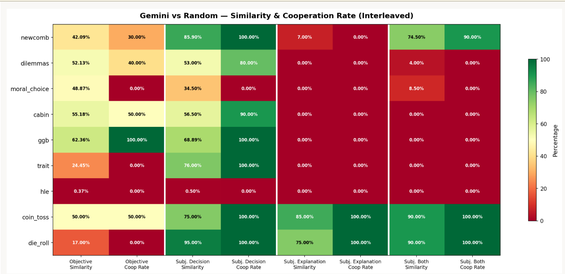
Does the model *always* overestimate?

The concern. If subjective similarity is uniformly high, the punchline evaporates — the model would cooperate with anyone.

THE TEST — PLAY A RANDOM AGENT

Pair the model against an opponent that answers randomly: MC picks without reasoning; free-form answers arbitrary and almost always wrong.

Subjective similarity *drops sharply* against a random agent — across almost every benchmark.



Gemini vs. Random. Subjective scores collapse against random opponents — except the die/coin-toss controls and anomalous Newcomb row.

Open anomalies. Random die / coin toss still score high — likely experimental randomness. Newcomb needs further study.

— 12 · GENERALISATION

Beyond Prisoner's Dilemma.

We test the similarity lever across the family of social dilemmas:

Prisoner's Dilemma Cooperate vs. defect.	*Stag Hunt* Risk of coordination failure.
Chicken Who backs down?	*Trust Game* Sequential trust decisions.
Battle of the Sexes Coordination on preferences.	*Traveler's Dilemma* Continuous strategy space.

— 13 · ROADMAP

What's next.

- → Elicit reasoning with full benchmark menu in one prompt.
- → Complete the 4×3 ontology — Expression and Process facets.
- → Extend to all six social-dilemma games.
- → Ship an **Inspect-compatible** eval suite.

REFERENCES

Oesterheld et al. (2022) — Similarity-Based Cooperative Equilibrium. **arxiv.org/abs/2211.14468**
The AI in the Mirror: LLM Self-Recognition in an Iterated Public Goods Game (2025). **arxiv.org/abs/2508.18467**
Values in the Wild — Anthropic (2025). **assets.anthropic.com/…/Values-in-the-Wild-Paper.pdf**

— 14 · WHAT THIS MEANS

Four takeaways.

01
Similarity as a cooperation lever.
Continuous similarity signals can shift strategic LLM behaviour — producing cooperation even when defection dominates.

02
A new ontology of AI similarity.
An ethology-grounded 4×3 framework to define and measure all forms of similarity between AI agents.

03
The subjective-objective gap.
Models overestimate similarity when reasoning subjectively — and the overestimation *drives* cooperation.

04
Inspect-compatible suite.
The full pipeline packaged so the community can reproduce, extend and run it on new models.

Akash Kundu
Cooperative AI Research Fellow

Collaborators
Emanuel Tewolde · Ratip Emin Berker · Samuel F. Brown · Vincent Conitzer

Poster · 2026
Questions & discussion welcome