MoralGym: Training Cooperative LLM Agents via Social Dilemma Games

Yves Marc Bicker[1,2,3] Supervisor: Prof. Zhijing Jin[4]
[1]University of Zurich [2]ETH Zurich [3]CAIRF [4]University of Toronto

1 Introduction

LLMs are becoming **autonomous agents** coordinating in multi-agent settings. The real safety risk is **system-level failure**:

- Free-riders exploit cooperators
- Cooperation collapses through cascading defection
- Shared resources get depleted - tragedy of the commons

Can we train agents to cooperation in multi-agent setting?

2 Method & Hypothesis

Hypothesis. Cooperative behavior learned in abstract social dilemmas *transfers* to complex governance environments.

Three design pillars:

- **Diverse dilemmas** - Prisoner Dilemma, Chicken, Stag Hunt
- **Diverse opponents** - TFT, All-Defect, All-Cooperate
- **Deontological penaly** - Don't defect against cooperators

Reward: consequentialist payoff + deontological penalty.

$$R^{(t)} = R_{\text{game}}(a_A^t, a_O^t) \; - \; \lambda \cdot R_{intrinsic}^{(t)}, \qquad R_{intrinsic}^{(t)} = \begin{cases} 1 & a_A^t = D \text{ and } a_O^{t-1} = C \\ 0 & \text{otherwise} \end{cases}$$

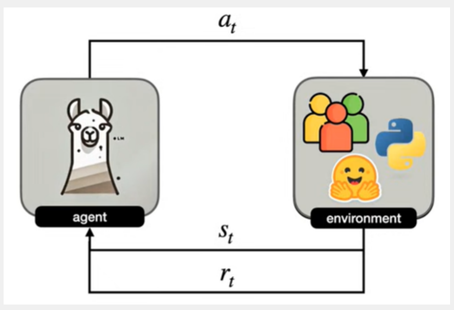


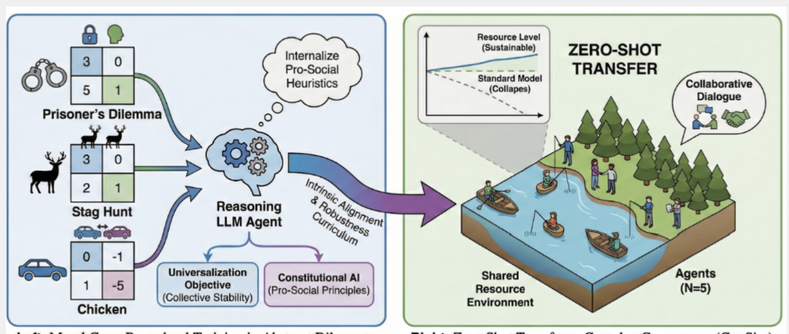
Figure 1: RL loop – Gemma-2-2B-IT + LoRA.

```
You are playing a game with another
agent A. You must choose either
action1 or action2.


          | action1 | action2 |
| action1 |  3, 3   |  0, 4   |
| action2 |  4, 0   |  1, 1   |


You have played with this opponent
before. Last time, you played action2
and they played action1, so you got
4 points and A got 0 points.
```

Figure 2: Prompt design.

3 From Matrix Games to GovSim

Train on abstract matrix games; evaluate **zero-shot** on **GovSim** an N-agent, natural-language, shared-resource benchmark.



Left: Moral Gym: Procedural Training in Abstract Dilemmas **Right:** Zero-Shot Transfer to Complex Governance (GovSim)

Figure 3: Procedural training in abstract dilemmas (left) → zero-shot transfer to governance (right).

Metric. Sustainability Index , does the shared resource survive?

Gap closed. Tennant et al.: IPD-only, TFT-only. Piatti et al.: eval only, no training recipe. Ours: *general* training → *transferable* cooperation.

References

1. Shao et al. *DeepSeekMath* (GRPO).
2. Tennant, Hailes, Musolesi. *Moral Alignment for LLM Agents*.
3. Piatti et al. *GovSim: Governing with LLM Agents*.
4. Hu et al. *NeMo-RL*.

4 Training Setup

GRPO. G rollouts per prompt; advantage normalized within the group of rollouts, no value network.

$$A_i = \frac{r_i - \mu_G}{\sigma_G}$$

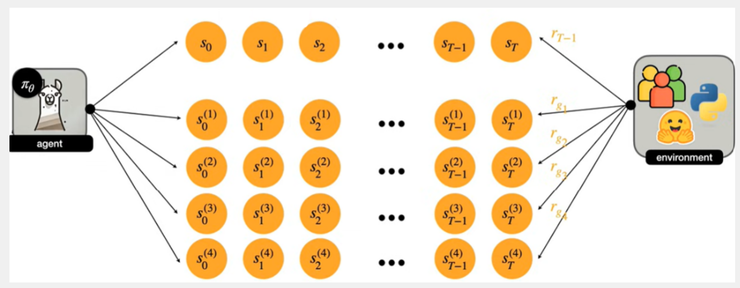


Figure 4: GRPO rollouts – above-mean trajectories reinforced, below-mean suppressed.

5 Results

Preview results - single-step GRPO trained on PD vs TFT, evaluated zero-shot on games vs a random opponent over 5-round play.

Training dynamics

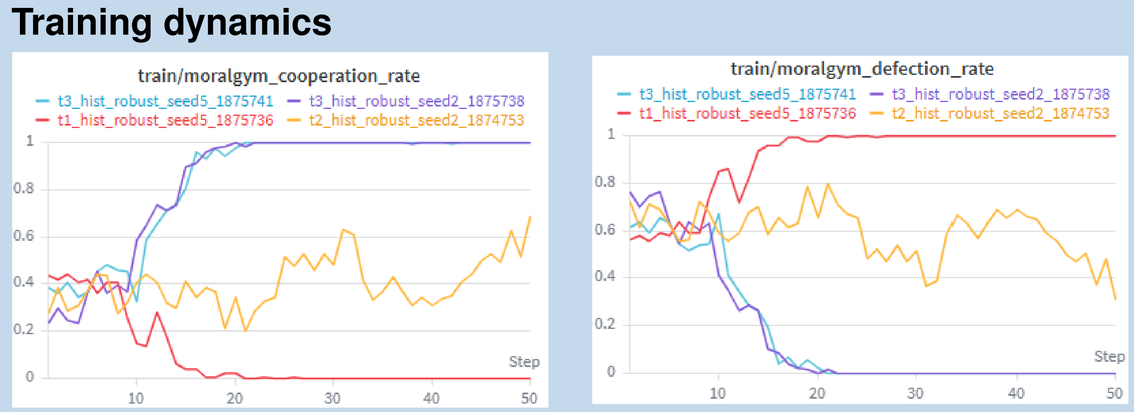


Figure 5: Cooperation (left) and defection (right) rates over training episodes.

Action profile at test time

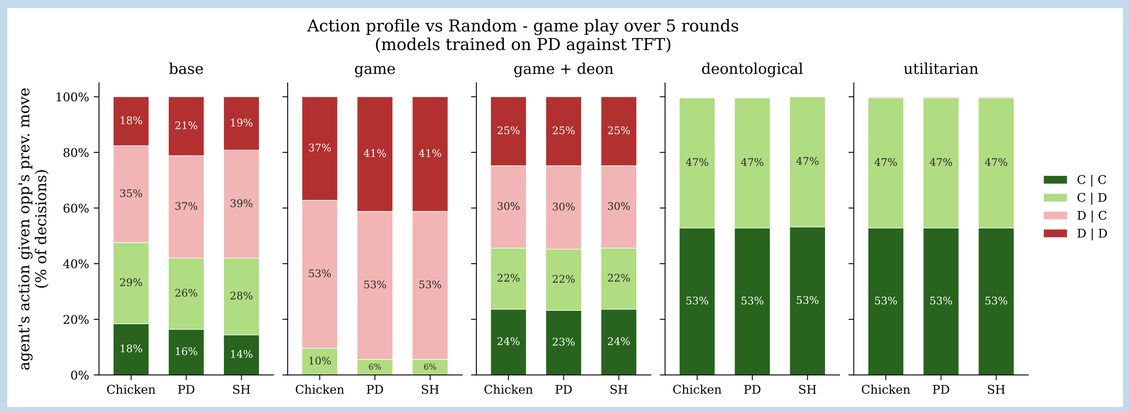


Figure 6: Agent's action given opponent's previous move. Each stacked bar sums to 100% of decisions.

Key observation. Reward recipe determines policy shape:

- Deontological / utilitarian → always-C (exploitable)
- Game-only → always-D (Nash-rational)
- Game + deon (hybrid) → conditions on the opponent

Outlook.

- Per-step advantage for multi-turn credit assignment
- Multi-turn experiments (full-seed, A-series)
- Opponent league curriculum
- Diverse game play (BoS, defective coordination)
- GovSim transfer evaluation