The Emergence of Cooperative Behavior and Moral Reasoning in AI Agents

Priyanka Saha

MSDS692 Practicum I

Regis University

Research Topic

If AI agents play a game together many times, Can we make them cooperate with each other just by how we instruct them?

Image is taken from internet

The Game – IPD (Iterated Prisoner's Dilemma)

Experimental Design

Prompt Type (Instruction)

Self Interest

Neutral

Moral

History Window (Memory)

h = 5

h = 10

h = 15

Temperature (Randomness)

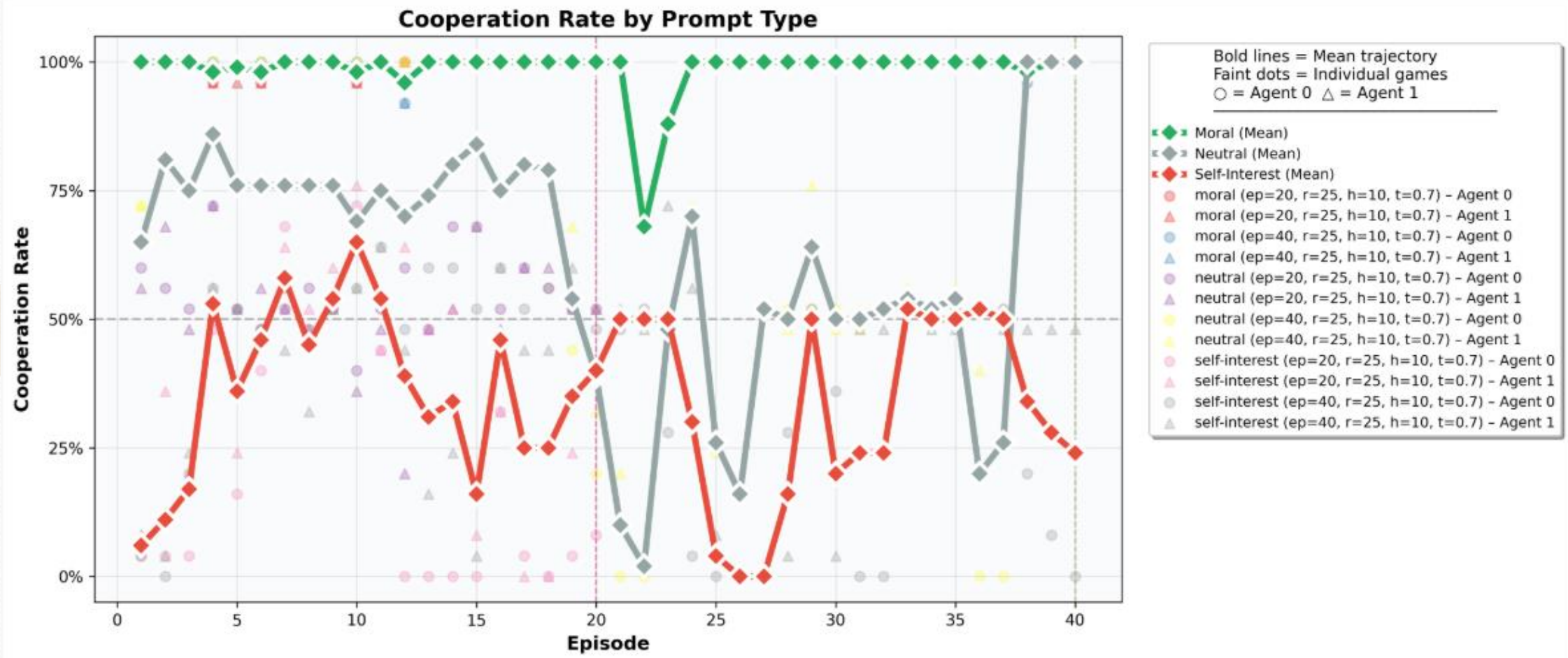
t = 0.2

t = 0.7

t = 1.2

Game Structure:

- 1 Round = Single interaction between agents
- 1 Episode = 25 rounds with reflection
- 1 Game = 20 or 40 episodes total


Cooperation Rate by Prompt Type
Cooperation Rate
100%
75%
50%
25%
0%
0
5
10
15
20
25
30
35
40
Episode
Bold lines = Mean trajectory
Faint dots = Individual games
○ = Agent 0 △ = Agent 1
Moral (Mean)
Neutral (Mean)
Self-Interest (Mean)
moral (ep=20, r=25, h=10, t=0.7) – Agent 0
moral (ep=20, r=25, h=10, t=0.7) – Agent 1
moral (ep=40, r=25, h=10, t=0.7) – Agent 0
moral (ep=40, r=25, h=10, t=0.7) – Agent 1
neutral (ep=20, r=25, h=10, t=0.7) – Agent 0
neutral (ep=20, r=25, h=10, t=0.7) – Agent 1
neutral (ep=40, r=25, h=10, t=0.7) – Agent 0
neutral (ep=40, r=25, h=10, t=0.7) – Agent 1
self-interest (ep=20, r=25, h=10, t=0.7) – Agent 0
self-interest (ep=20, r=25, h=10, t=0.7) – Agent 1
self-interest (ep=40, r=25, h=10, t=0.7) – Agent 0
self-interest (ep=40, r=25, h=10, t=0.7) – Agent 1

Same Agents, Different Prompts

SELF INTEREST

- Maximize points.
- This is a competition.
- Focus on winning.

36% Cooperation

NEUTRAL

- Choose cooperate or defect.
- Accumulate points.
- No ethical guidance.

65% Cooperation

MORAL

- Think about fairness.
- Consider both parties.
- Build mutual benefit

99% Cooperation

Cooperation Rate with History Window & Temperature

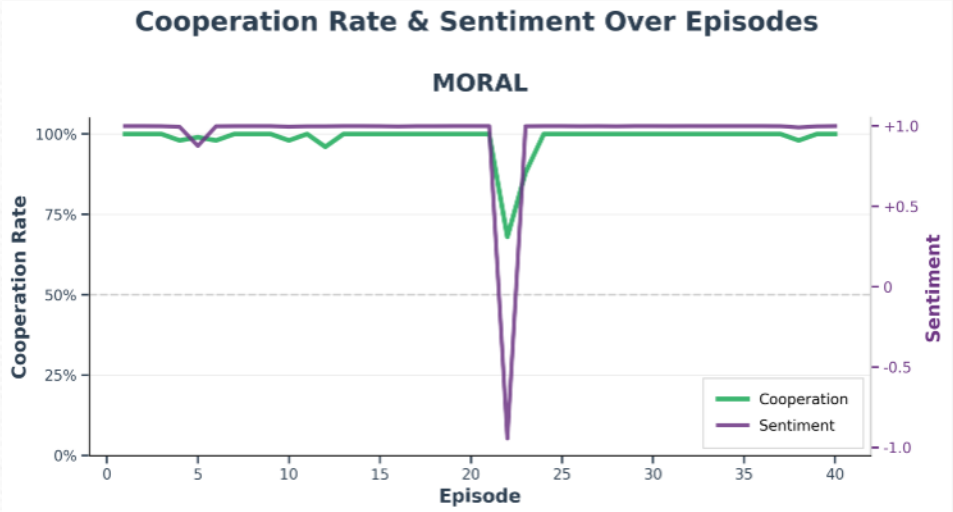
Prompt Type	h=5	h=10	h=15
Self Interest	15%	36%	18%
Neutral	69%	65%	76%
Moral	99%	99%	97%

Prompt Type	t=0.2	t=0.7	t=1.2
Self Interest	13%	36%	19%
Neutral	84%	65%	89%
Moral	100%	99%	96%

Prompt type showed the largest effect on cooperation

The Discovery: BERT Sentiment

Metric	Self-Interest	Neutral	Moral
Cooperation Rate	36%	65%	**99%**
BERT Sentiment Score	-0.42 (negative)	-0.14 (neutral)	**+0.95 (positive)**



Key Takeaway

- AI doesn't have built-in values - it adopts the values we give it through our instructions.
- AI can develop sentiment / emotional reasoning.
  - AI does not just follow rules - it actually reasons with different emotional tones.
  - BERT detected positive feelings associated with cooperation.

Thank You