Human Behavior patterns in the Stochastic Prisoner's Dilemma

Evangelia Chalioti Himnish Hunmna

Yale University

CRETE 2025

What if it is possible to isolate behavioral patterns that matter for emulating the human play of the Prisoner's dilemma & instill them in a computation model via training?

What if it is possible to isolate behavioral patterns that matter for emulating the human play of the Prisoner's dilemma & instill them in a computation model via training?

Such a model will

- generate more data on the human play of games, while cutting costs,
- be a more accurate representation of human behavior in games,
- can inform better predictions about real situations with human agents.

What if it is possible to isolate behavioral patterns that matter for emulating the human play of the Prisoner's dilemma & instill them in a computation model via training?

Such a model will

- generate more data on the human play of games, while cutting costs,
- be a more accurate representation of human behavior in games,
- can inform better predictions about real situations with human agents.

This paper focuses on identifying one such behavioral pattern:

- finding the determinants of cooperation in an infinitely repeated Prisoner's Dilemma with a stochastic twist,
- creating the training conditions for a learning algorithm to exhibit it.

Literature

Kloosterman (2020), Rojas (2012), Vespa & Wilson (2019): human agents are more likely to cooperate in period 0 when p_A is lower than a threshold. p_A: probability that the less cooperative payoffs are realized in the future.

- This paper uses Reinforcement learning agents instead of humans,
- varies p_A & the discount rate.

Literature

Kloosterman (2020), Rojas (2012), Vespa & Wilson (2019): human agents are more likely to cooperate in period 0 when p_A is lower than a threshold. p_A: probability that the less cooperative payoffs are realized in the future.

- This paper uses Reinforcement learning agents instead of humans,
- varies p_A & the discount rate.

Sgroi & Zisso (2007), Spiliopoulos (2011), Fudenberg & Karreskog (2021), Lin, Bouneffouf & Cecchi (2020): study human behavior patterns in games

- This paper adopts a teacherless training algorithm.

Literature

Kloosterman (2020), Rojas (2012), Vespa & Wilson (2019): human agents are more likely to cooperate in period 0 when p_A is lower than a threshold. p_A: probability that the less cooperative payoffs are realized in the future.

- This paper uses Reinforcement learning agents instead of humans,
- varies p_A & the discount rate.

Sgroi & Zisso (2007), Spiliopoulos (2011), Fudenberg & Karreskog (2021), Lin, Bouneffouf & Cecchi (2020): study human behavior patterns in games

- This paper adopts a teacherless training algorithm.

Sandholm & Crites (1996), Harper et al (2017): models are trained to become hyper specialized.

- This paper enables the algorithm to have general knowledge about the game; to emulate human biases & inconsistencies.

Iterated Stochastic Prisoner's Dilemma

Prisoner's Dilemma (PD): 2 agents choose to either cooperate or defect.

- Playing Defect leads to higher payoff regardless of opponent's action.

Iterated Stochastic Prisoner's Dilemma

Prisoner's Dilemma (PD): 2 agents choose to either cooperate or defect.

- Playing Defect leads to higher payoff regardless of opponent's action.

Iterated *Prisoner's Dilemma (IPD)*: repeating the dilemma sequentially.

- If infinite time horizon: cooperation can emerge during play.

Iterated Stochastic Prisoner's Dilemma

Prisoner's Dilemma (PD): 2 agents choose to either cooperate or defect.

- Playing Defect leads to higher payoff regardless of opponent's action.

Iterated *Prisoner's Dilemma (IPD)*: repeating the dilemma sequentially.

- If infinite time horizon: cooperation can emerge during play.

Iterated Stochastic *Prisoner's Dilemma (ISPD)*: same game as IPD except that the payoff matrix is subject to change.

- Uncertainty about which payoff matrix will be used to calculate the reward from the outcome of an iteration of the dilemma.

The 2 Dilemmas. C: *Cooperate*, D: *Defect*.

	C	D
C	$32, 32$	$10, 52$
D	$52, 10$	24,24

Dilemma A

	C	D
C	$62, 62$	$10, 82$
D	$82, 10$	24,24

Dilemma B

In Dilemma B,

- higher payoff when both agents cooparate,
- & higher payoff when an agent defects.

Also δ: discount factor, p_A: defective probability - the probability that the less cooperative Dilemma A gets realized.

The 2 Dilemmas. *C*: *Cooperate*, *D*: *Defect*.

	C	*D*
C	32, 32	10, 52
D	52, 10	24,24

Dilemma A

	C	*D*
C	62, 62	10, 82
D	82, 10	24,24

Dilemma B

The realization of the stochastic process is revealed to players at the start of every period.

- A game is played in period 0,
- & the game realized in period 1 is played in every period onwards.

The 2 Dilemmas. *C*: *Cooperate*, *D*: *Defect*.

	C	*D*
C	32, 32	10, 52
D	52, 10	24,24

Dilemma A

	C	*D*
C	62, 62	10, 82
D	82, 10	24,24

Dilemma B

The realization of the stochastic process is revealed to players at the start of every period.

- A game is played in period 0,
- & the game realized in period 1 is played in every period onwards.

Key reasons to use this stochastic process:

- It exaggerates the role of p_A (we can observe effects from altering it),
- the number of different realized games is manageable,
- the early periods matter more anyway in infinite horizon games.

The 2 Dilemmas. C: *Cooperate*, D: *Defect*.

	C	D
C	$32, 32$	$10, 52$
D	$52, 10$	24,24

Dilemma A

	C	D
C	$62, 62$	$10, 82$
D	$82, 10$	24,24

Dilemma B

Cooperation in periods $t \geq 1$ is viable iff B is realized in period 1.

- Grim trigger is an equilibrium when B is infinitely repeated but not when A is infinitely repeated.
- Mutual cooperation in future periods & so punishment is possible when B is realized in period 1.

The 2 Dilemmas. *C*: *Cooperate*, *D*: *Defect*.

	C	*D*
C	32, 32	10, 52
D	52, 10	24,24

Dilemma A

	C	*D*
C	62, 62	10, 82
D	82, 10	24,24

Dilemma B

Constraints for cooperation to be viable in period 0:

$$32 + p_A\left(24\delta + 24\delta^2 + ...\right) + (1 - p_A)\left(62\delta + 62\delta^2 + ...\right) \geq$$

$$52 + \left(24\delta + 24\delta^2 + ...\right)$$

$$62 + p_A\left(24\delta + 24\delta^2 + ...\right) + (1 - p_A)\left(62\delta + 62\delta^2 + ...\right) \geq$$

$$82 + \left(24\delta + 24\delta^2 + ...\right)$$

The cutoff of p_A is independent of the game played in period 0.

The 2 Dilemmas. *C*: *Cooperate*, *D*: *Defect*.

	C	*D*
C	32, 32	10, 52
D	52, 10	24,24

Dilemma A

	C	*D*
C	62, 62	10, 82
D	82, 10	24,24

Dilemma B

Constraints for cooperation to be viable in period 0:

$$32 + p_A\left(24\delta + 24\delta^2 + ...\right) + (1 - p_A)\left(62\delta + 62\delta^2 + ...\right) \geq$$

$$52 + \left(24\delta + 24\delta^2 + ...\right)$$

$$62 + p_A\left(24\delta + 24\delta^2 + ...\right) + (1 - p_A)\left(62\delta + 62\delta^2 + ...\right) \geq$$

$$82 + \left(24\delta + 24\delta^2 + ...\right)$$

The cutoff of p_A is independent of the game played in period 0.

The 2 Dilemmas. C: *Cooperate*, D: *Defect*.

	C	D
C	$32, 32$	$10, 52$
D	$52, 10$	24,24

Dilemma A

	C	D
C	$62, 62$	$10, 82$
D	$82, 10$	24,24

Dilemma B

Both cooperation & defection can occur on the equilibrium path.

- If p_A small: the players start with C but switch to D if A is realized in period 1.
- If p_A large: the players start with D but switch to C if B is realized in period 1.

Reinforcement Learning

Any task is characterized by the reward provided given a state & an action:

- the actions the agent can take to affect the environment,
- the states of the environment that the task takes place in.

In this paper,

- the reward is the cumulative payoff aggregated over all iterations of the game against the opponent.
- the action any player can take is either C or D.
- the states are the 4 possible outcomes (C, C), (C, D), (D, C), (D, D).

Q Learning algorithm (QLearner) uses an update rule:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \; max_a \; Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

- Q: action-value function
- R_t: reward in period t
- S_t: state in period t
- a, A_t: actions at period t
- γ: QLearner discount rate
- α: learning rate
  - It shows to what extent the QLearner's estimated rewards at a point get influenced by the expectation encountered in the next state following an action & the reward stemming from that action.

Methodology

Use of an open-source codebase containing over 230 IPD strategies.

The codebase and the match logic are modified to enable

1. the dilemma being played in period 0 to be random.
   - At period 0, either Dilemma A or B is played with probability 0.5.
   - p_A changes among a fixed collection of values: to avoid introducing a bias towards playing A or B.
   - If Dilemma A is realized, then in period 2 and onwards, p_A is set to 1.
   - If Dilemma B is realized, then in period 2 and onwards, p_A is set to 0.

Methodology

Use of an open-source codebase containing over 230 IPD strategies.

The codebase and the match logic are modified to enable

1. the dilemma being played in period 0 to be random.
   - At period 0, either Dilemma A or B is played with probability 0.5.
   - p_A changes among a fixed collection of values: to avoid introducing a bias towards playing A or B.
   - If Dilemma A is realized, then in period 2 and onwards, p_A is set to 1.
   - If Dilemma B is realized, then in period 2 and onwards, p_A is set to 0.
2. infinite play, by changing the probability the game ends at any period.

Methodology

Use of an open-source codebase containing over 230 IPD strategies.

The codebase and the match logic are modified to enable

1. the dilemma being played in period 0 to be random.
   - At period 0, either Dilemma A or B is played with probability 0.5.
   - p_A changes among a fixed collection of values: to avoid introducing a bias towards playing A or B.
   - If Dilemma A is realized, then in period 2 and onwards, p_A is set to 1.
   - If Dilemma B is realized, then in period 2 and onwards, p_A is set to 0.
2. infinite play, by changing the probability the game ends at any period.
3. the QLearners to sense some aspects of the environment:
   - the past 12 moves of their opponent,
   - the proportion of cooperation in the match so far.

Training: the updated values for exp. total rewards from each state & action are feed back to QLearner at the start of a new match.

- One iteration of training involves putting the QLearner with no knowledge through 20 matches against 20 random strategies.
- QLearner retains knowledge it gained from each match.
- QLearner goes through 15 iterations before testing.

Results for Period 0

- A cooperation majority persisted at most value of p_A.
- Flips from cooperative to defective majority tend to persist at higher p_A but cooperative majorities remain the norm.

Results for Period 0

- A cooperation majority persisted at most value of p_A.
- Flips from cooperative to defective majority tend to persist at higher p_A but cooperative majorities remain the norm.

Results for Period 1

- "Switching C to D" happens more often as p_A increases.
- If A gets realized in period 1 onwards, the number of cooperation in period 0 will also decrease.
- Actions in period 0 do not influence this trend.

Conclusion

The value of defective probability leads players to make optimistic or pessimistic assumptions about the future, impacting their current behavior.

- Under high learning rates, the Q-Learner is capable of exhibiting the same behavior after following randomized training.

Human agents update their expectations about the future once a game is established for the rest of the foreseable future.

- The Q-Learner is capable of updating its beliefs like human agents when either A or B gets realized in period 1.

Thank you!