Strategy Shift in Prisoner's Dilemma through Utility Learning

Kwangsu Cho Christian Schunn
Learning Research and Development Center
Univ. of Pittsburgh

Prisoner's Dilemma (PD)

- Non zero-sum game
- Goal: Getting big payoffs
- Two players are involved.
- Strategy Choice without knowing each other's choice
  - In each trial, each player must choose between the *cooperate* (C) and the *defect* (D) strategy
- Players receive payoffs depending on both of the moves
  - Your payoffs depend on your partner's move
- In a typical study two players participate in multiple trial play of the game.

Prisoner’s Dilemma Payoff Matrix

Player 2

	Move	Defect2	Cooperate2
Player 1	Defect1	-1 -1	+10 -10
	Cooperate1	-10 +10	+1 +1

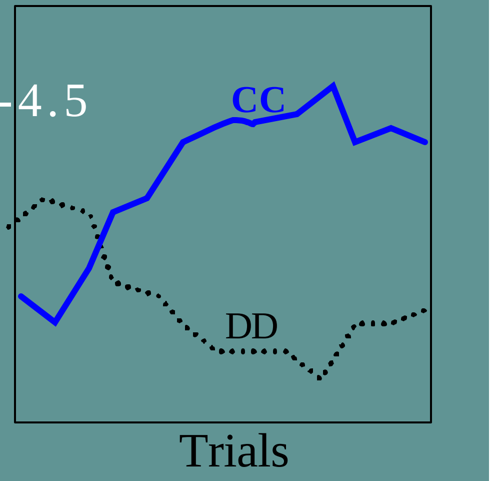
- Expected Payoff
  - Defect = (-1 + 10) / 2 = 4.5
  - Cooperate = - 4.5
  - Rational action = Defect
  - Irrational action = Cooperate
- A conflict between rational and irrational behavior
  - The loss from defect vs. the benefits from Coop.
- Strategy Shift = learning process
  - from the *Defect* to the *Cooperate*

Motivation & Goal

- Game theory assumes *Rationality*
  - Chaotic performance in the beginning is ignored.
  - Equilibrium state in games needs multi-hundreds of trials
  - Human cognition (learning and adaptation) is ignored
  - Lack of short-term prediction
- Simulation of Strategy shift in the PD
  - To consider human learning or adaptation process

Strategy Shift Phenomena

- Strategy Shift
  - From rational choice in the beginning to irrational choice later on
  - From Defect To Cooperate
- Conflicts between immediate payoff and goal
  - Immediate payoffs interfere with goal
    - Expected gain: Defect = 4.5 vs. Cooperate = -4.5



Lebiere, Wallach, & West (2000)

- Memory-based model
  - The most likely outcomes are determined by retrieving the most active of the possible move combinations
  - Retrieve most likely (most active) consequence of Cooperation and of Defection
  - Pick strategy with highest gain
- Winner takes all
  - Once a pattern of behavior is established, it seems not changeable
    - Strategy that's more common in the beginning tended to be stable
    - Self-reinforcing chunk strength
  - Inherent bias for defecting in the beginning
  - Strategy shift was hard to simulate

Our Model Flow

Retrieve Payoff Matrix

Calculate Expected Payoff (*EP*) per each strategy

Decide Strategy Choice Preference

If *EP*(D) > *EP*(C) or
If *EP*(D) < *EP*(C)

Make a Move

If *EP*(D) > *EP*(C)

D_Move_Defect D_Move-Cooperate

If *EP*(D) < *EP*(C)

C_Move-Defect ***C_Move-Cooperate***

Get Partner's Move

Receive Real Payoff (*RP*)

Compare RP with EP

Punish the rational choice if it fails (when RP < EP)
Reinforce the irrational one if it succeeds (when RP > EP)

Request New Goal

Utility Learning of the Model

- Production for rational choice is weighted in the beginning
  - When EP(D) > EP(C),
    - (spp D_Move-Defect :failures 0 :successes 20 :efforts 100)
    - (spp D_Move-Cooperate :failures 20 :successes 20 :efforts 100)
  - When EP(D) < EP(C),
    - (spp C_Move-Defect :failures 20 :successes 20 :efforts 100)
    - (spp C_Move-Cooperate :failures 0 :successes 20 :efforts 100)

Surprise-Based Utility Learning

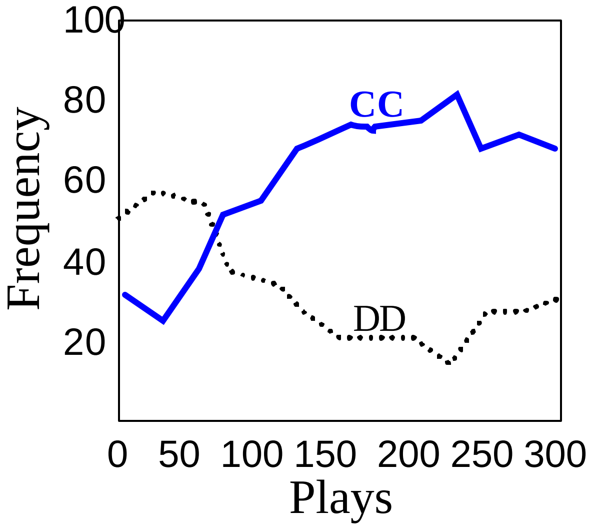
- Unbalanced Reinforcement of Strategy
  - Punish the rational choice if fails when RP < EP
    - e.g (spp Eval-Payoff-Poor-D :failure t)
  - Reinforce the irrational choice if succeeds when RP > EP
    - E.g. (spp Eval-Payoff-Good-C :success t)

Result 1. General Fit

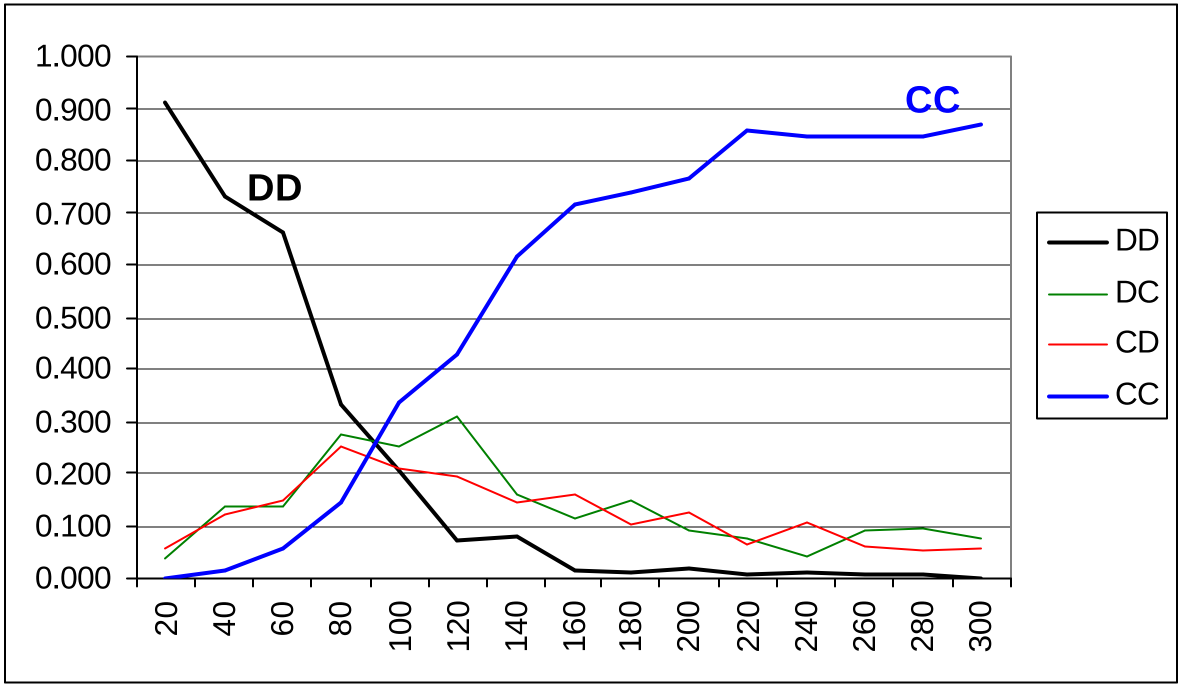
	DD	DC	CD	CC	r	Mean-Dev.
Human Data	30	7	8	55		
Lebiere et al	32	8	6	54	.99	.02
Our Model	20	13	12	55	.95	.06

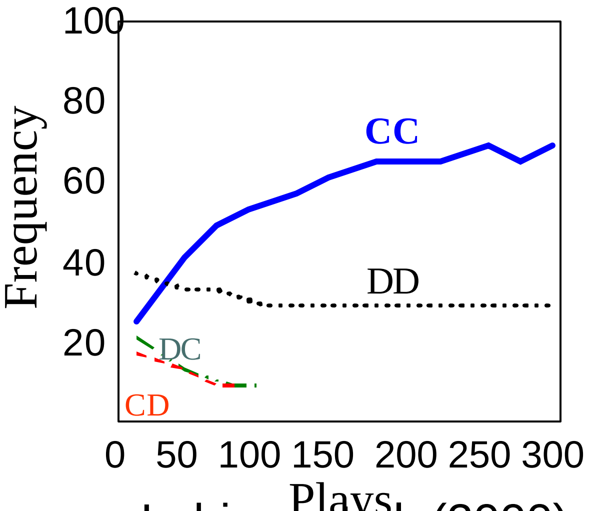
- Method
  - 10 groups of two players
  - 300 trials per group

Result 2. Strategy Shift



Rappoport et al. (1976)





Lebiere et al. (2000)

Result 3. Individual difference

Human Data				
Run	DD	DC	CD	CC
1	1	1	1	97
2	7	1	1	92
3	14	1	2	83
4	4	5	5	86
5	21	4	3	72
6	24	5	5	66
7	54	12	7	27
8	34	2	52	11
9	58	25	5	12
10	83	9	4	3
	30	7	8	55

Cho & Schunn				
Run	DD	DC	CD	CC
1	20	4	8	68
2	23	7	6	64
3	9	9	19	63
4	20	9	12	59
5	20	14	10	56
6	21	17	7	55
7	20	18	8	54
8	16	16	18	50
9	32	11	17	40
10	22	23	15	40
	20	13	12	55

Lebiere et al.				
Run	DD	DC	CD	CC
1	1	0	2	97
2	1	1	2	96
3	2	9	2	87
4	5	4	10	81
5	4	19	12	65
6	10	13	12	65
7	13	21	18	48
8	92	4	3	1
9	93	3	3	1
10	95	3	2	0
	32	8	6	54

Conclusion

- Our model captures features not previously captured
  - model captures both the asymptotic behavior and the strategy shift
- The model doesn't assume any altruistic assumption
  - considering partner's gains as general solutions in the Game theory. Instead, the model seeks moves for its own maximal gain.
- Surprise based learning
  - Unbalanced or weighted reinforcement learning
  - Reinforcing each strategy as either good or poor
    - the natural defecting strategy is reinforced negatively when it fails, but not positively even when it succeeds.
    - the cooperative is reinforced only positively when it's successful

Limitations and Difficulties

- Dominant preference for defecting in the beginning
- Sometimes human players start with the irrational choice, cooperation
  - We don’t model it

- Learning too slow
  - Utility learning unit is limited to 1 in/decrement per experience
- Turning off surprise-based learning
  - Habituation process?
  - Once a behavior is set, it doesn’t need to be strengthen or weaken

Acknowledgement

Christian Lebiere

Dan Bothell

Raluca Budiu

John Anderson

Especially

You, ACT-R fellows