

# CS 886: Affective Computing Presentation

Affect Control Theory and the Prisoner's Dilemma  
Fall 2021



# Topic

- Discuss theoretical framework of affect control theory and the Prisoner's Dilemma
- Create a working model to investigate emotions in the prisoner's dilemma.

		Player 2	
		confess	don't confess
Player 1	confess	$(-6, -6)$	$(0, -10)$
	don't confess	$(-10, 0)$	$(-1, -1)$



# Motivation

- Determine a model for an AI-agent to participate in games like the prisoner's dilemma and account for human emotion.
- To help better understand the underlying attributes of human emotion in real-world decisions.
- Attempt to understand the rules and systems that humans use when engaging in decision making.
- Addressing the primary problem of the prisoner's dilemma: that two individualistically rational agents choose a poor solution for a problem rather than cooperate



# Related Works

- Direct tests using game-theoretical derivatives of cooperative situations have so far been unable to find a connection between the communication of emotion and cooperation (Wubben, 2009).
- However, direct evidence of communication and emotion going hand in hand include laboratory studies showing that participants played against the reciprocal strategy of tit-for-tat showed that communicated disappointment established more cooperation than did anger (Wubben et al., 2009).
- Further studies have shown that emotions help establish cooperation through indirect reciprocity (Wubben, 2011).



# Background

- Affect control theory seeks to explain behaviours in the context of social interactions, describe the routine and expected behaviours that people enact under normal circumstances, and the creative responses they generate when encountering noninstitutionalized or counter-normative situations (Robinson et al., 2006).
- The prisoner's dilemma is a thought experiment and game in game theory that demonstrates how two entirely rational individuals might not cooperate, even if it is in their mutual interests to do so (Kuhn 2019)
- A repeated prison dilemma is a prisoner's dilemma operating under the assumption that the game will repeat in the future, such that the credibility of the opponent and memory of the opponent's previous moves matter (Kuhn 2019)
- An infinite prisoner's dilemma is a repeated prisoner's dilemma that never ends (Kuhn 2019)



# Claims & Assumptions





- We are assuming that the scenarios modeled are built without any outside interferences.
- We assume that rational agents attempt to optimize actions based *only* on given premises.
- We assume emotions can be thought of as a given state in a machine
- We assume that agents will only model decisions based on the given scenarios





# Theoretical Results – Classic Prisoners Dilemma

- Purely rational agents choose suboptimal outcomes in the classic prisoner's dilemma (*The prisoner's dilemma*).
- It is easy to see how there is room for improvement in this situation.
- [Image Source: Encyclopedia Britannica (*The prisoner's dilemma*)]

Prisoners' dilemma		prisoner B			
		confess		remain silent	
prisoner A	confess	 5 years 5 years	 0 year 20 years		
	remain silent	 20 years 0 year	 1 year 1 year		

© 2010 Encyclopædia Britannica, Inc.



# Theoretical Results – Repeated Prisoners Dilemma

- Depending on how agents weight future outcomes – modeled by delta ( $\delta$ ) – their responses to the dilemma change as well (Bó, 2019).
- This gives us a starting point to untangle the dilemma to create a new set of actions the agents can follow.

	Cooperate	Defect		Cooperate	Defect		Cooperate	Defect	
Cooperate	3,3	0,5	Cooperate	3,3	0,5	.....	Cooperate	3,3	0,5
Defect	5,0	1,1	Defect	5,0	1,1		Defect	5,0	1,1
Repetition 1			Repetition 2				Repetition 10		





# Methods – Preliminaries

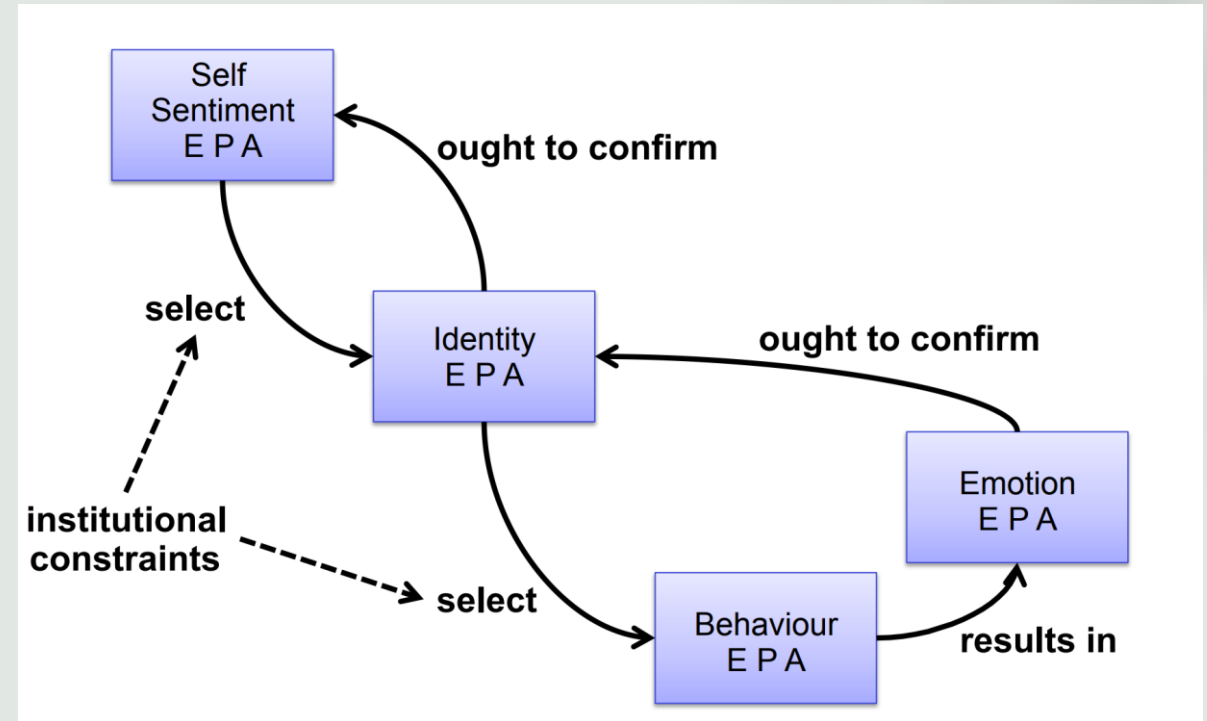
- There is now considerable evidence that humans are not purely self-interested and do not always behave according to the predictions of game theory (Melo, 2011).
- Emotional displays by foreign parties can be used to determine optimal decision-making policies in social dilemmas that consider other parties' emotional displays (Melo, 2011).
- Recent discovery of extortion and generous strategies has renewed interest in the role of strategy in shaping behavior in the repeated prisoner's dilemma (Melo, 2020).



# Methods – Affect Control Theory

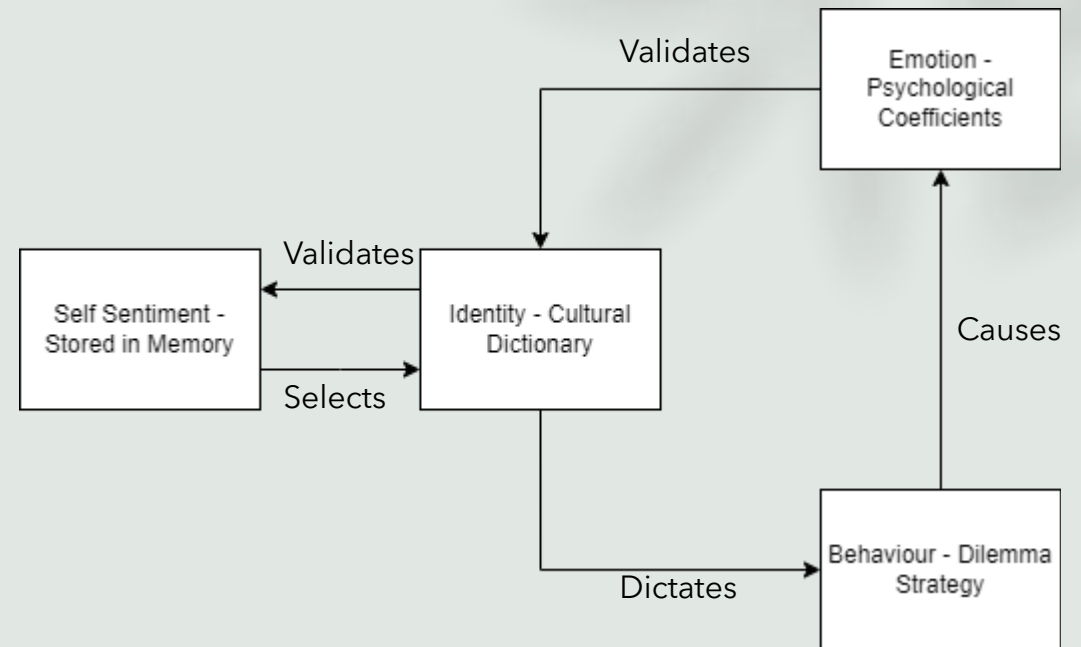
## Affect control theory argues

1. Actors react to social situations in terms of symbols and the meanings that those symbols carry for them (Robinson et al., 2006).
2. The meanings that symbols have are largely shared within a culture, leading actors to be able to role take, viewing the situation from the position of other actors and anticipating their reactions to the interaction (Robinson et al., 2006).
3. Actors are motivated to maintain the meanings associated with the self (Robinson et al., 2006).
4. Meanings can shift within situations as a result of one's own or others' actions (Robinson et al., 2006).
5. Emotions act as signals about how events are maintaining or failing to maintain self-identities within an interpersonal situation (Robinson et al., 2006).



# Experimental Results – Affect Control Model

- Following in line with affect control theory, we can model agents in the Prisoners Dilemma and Iterated Prisoners Dilemma as follows:
- Self sentiment – this represents the current state in memory for an agent
- Identity – this represents the cultural representation of the agent
- Behavior – this represents the strategy an agent takes in the dilemma
- Emotion – this represents how the agent feels as a result of the dilemmas outcome



# Affect Control Model – Parameters

- Self sentiment represents the behavioral state stored in an agent from past events or over multiple iterates of a game.
  - Example: An agent who feels close to the other agent may choose a friend or sibling identity. An agent who feels betrayed from a previous iteration or game, may choose a grim trigger strategy as revenge.
- Identity represents the dictionary of cultural roles that an agent can participate in
  - Example: a conman identity causes an agent to try and build trust and confidence in other agents to betray them such that they have more points than is possible through pure cooperation.
- Behavior represents the dilemma strategy that an agent partakes in
  - Example: An agent plays tit-for-tat when playing against new opponents but plays grim trigger against opponents upon betrayal.
- Emotion represents the feeling an agent gets upon finishing the dilemma
  - Example: based on intrinsic psychological coefficients, which can be defined in terms of constant attributes agents have.



# Strengths and Weaknesses

- Strengths
  - Provides a framework that a reinforcement learning model could use to train agents
  - Exact parameters can be custom defined for different scenarios
- Weaknesses
  - Oversimplifies emotions and loses some of the qualitative nuance that comes with different emotional states
  - Hard to extend models past games which can be defined rigorously



# Discussion

- Philosophical Questions
  - Can emotional states be quantified
  - Can all possible emotional states be rigorously defined into a dictionary
  - Emotion and reason may not be strictly distinct.
- Ethical Questions
  - Is it ok to let machines learn human states
  - Can a machine truly understand what is happening in the game
- Practical Questions
  - Is it feasible to define all possible emotional states and identities
  - Can an agent learn all possible states for games such that it behaves like a human
- Future Applications
  - Training an agent to engage in social reciprocity like people do
  - Recreating the evolutionary conditions which causes human psychological properties to exist in the first place





# Conclusion

- Many possible Affective Control Models exists for the Prisoner's Dilemma
- Emotional states in the Prisoner's Dilemma are related to past results and future motives
- Emotions intrinsic to individuals changes the strategy an agent uses in the Prisoner's Dilemma
- Emotion and reason are closely related. A purely rational agent likely does not achieve good long-term results compared to an agent with emotional modelling



# References

1. De Melo, Celso M., and Kazunori Terada. "The Interplay of Emotion Expressions and Strategy in Promoting Cooperation in the Iterated Prisoner's Dilemma." *Scientific Reports*, vol. 10, no. 1, 2020, <https://doi.org/10.1038/s41598-020-71919-6>.
2. De Melo, Celso M., et al. "A Computer Model of the Interpersonal Effect of Emotion Displayed in a Social Dilemma." *Affective Computing and Intelligent Interaction*, 2011, pp. 67–76., [https://doi.org/10.1007/978-3-642-24600-5\\_10](https://doi.org/10.1007/978-3-642-24600-5_10).
3. J., Wubben Maarten J. *Social Functions of Emotions in Social Dilemmas*. Erasmus University Rotterdam , 2009.
4. Kuhn, Steven. "Prisoner's Dilemma." *Stanford Encyclopedia of Philosophy*, Stanford University, 2 Apr. 2019, <https://plato.stanford.edu/entries/prisoner-dilemma/>.
5. Robinson, Dawn T., et al. "Affect Control Theory." *Handbooks of Sociology and Social Research*, 2006, pp. 179–202., [https://doi.org/10.1007/978-0-387-30715-2\\_9](https://doi.org/10.1007/978-0-387-30715-2_9).
6. Wubben, Maarten J.J., et al. "How Emotion Communication Guides Reciprocity: Establishing Cooperation through Disappointment and Anger." *Journal of Experimental Social Psychology*, vol. 45, no. 4, 2009, pp. 987–990., <https://doi.org/10.1016/j.jesp.2009.04.010>.
7. Wubben, Maarten J.J., et al. "The Communication of Anger and Disappointment Helps to Establish Cooperation through Indirect Reciprocity." *Journal of Economic Psychology*, vol. 32, no. 3, 2011, pp. 489–501., <https://doi.org/10.1016/j.joep.2011.03.016>.
8. Bó, Pedro Dal, and Guillaume R. Fréchette. "Strategy Choice In The Infinitely Repeated Prisoners' Dilemma." *American Economic Review*, vol. 109, no. 11, Nov. 2019, pp. 3929-52., <https://www.aeaweb.org/articles?id=10.1257/aer.20181480>.
9. "The prisoner's dilemma," *Encyclopædia Britannica*. [Online]. Available: <https://www.britannica.com/science/game-theory/The-prisoners-dilemma>. [Accessed: 02-Dec-2021].

