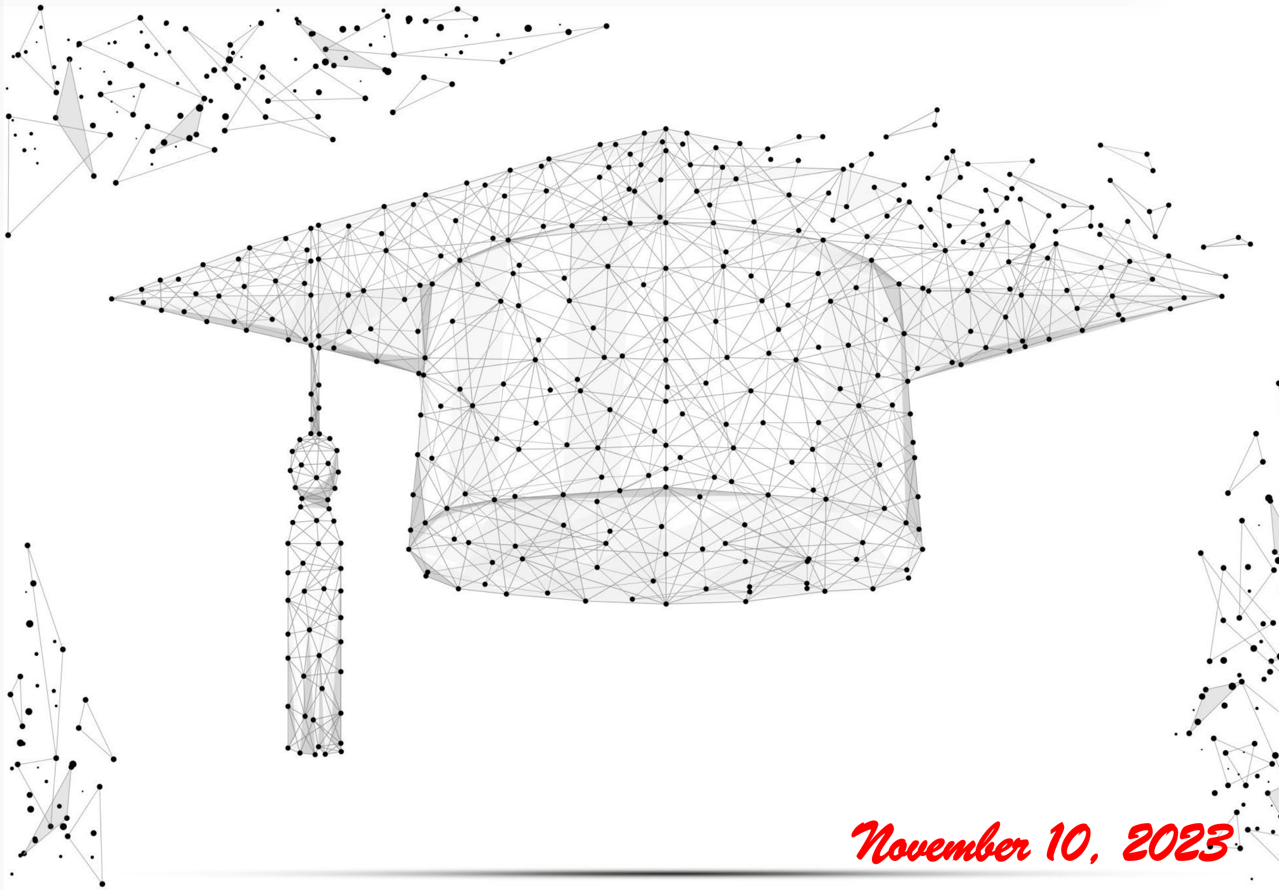
ChatGPT Plays the Treasures of Game Theory

Matt Kovach
Hector Tzavellas
Sudipta Sarangi
Michael Wagnon

VIRGINIA TECH



Danube Conference 2023

Ethics, AI and Higher Education Management

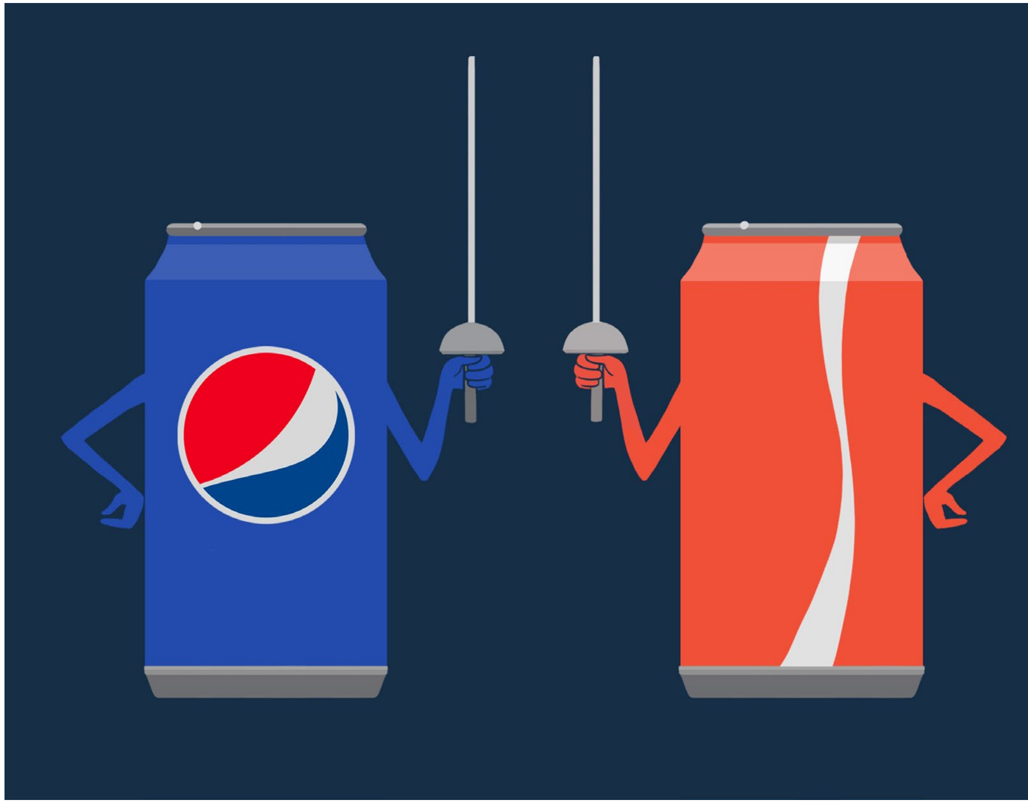
November 10, 2023

Motivation

- Understanding the decision-making ability of AI, especially Large Language Models (LLMs) is becoming increasingly important.
- There are growing literature studying different aspects of this – especially do LLMs behave like humans?
- Many human interactions involve strategic interactions: the outcome depends on the actions of more than one person!

Goal: *To test and understand the strategic capabilities of LLM*

Games we play...



Approach

GPT-3.5

- We adopt an experimental approach since this provides us with a controlled setting to gain insights.

- We use ChatGPT 3.5 released in 2022.

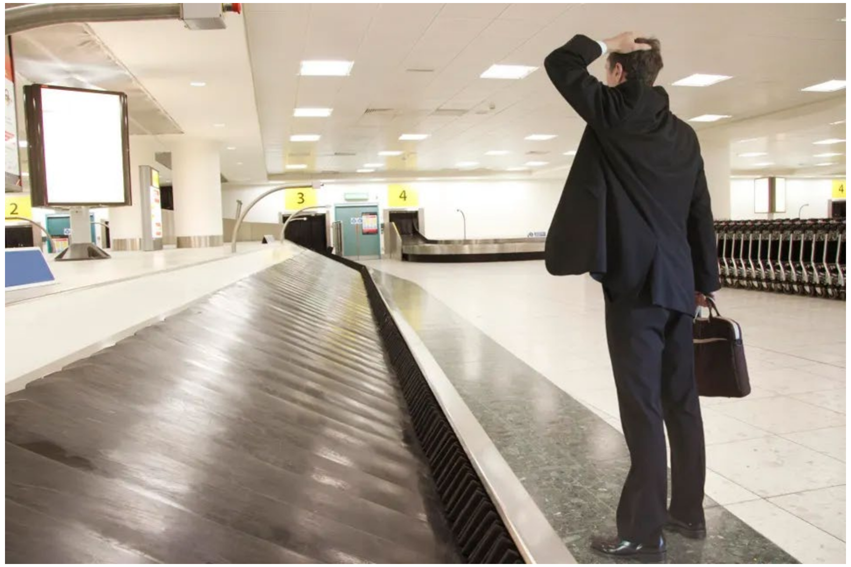
- We use the *Ten Little Treasures of Game Theory and Ten Intuitive Contradictions* (Goeree and Holt, AER (2001))

⇒ *This allows us to study how ChatGPT behaves in comparison to (human) experimental subjects.*

The Traveler’s Dilemma (AER, 1994)

The Traveler’s Dilemma (AER, 1994)

🙁



Traveler’s Dilemma

- Airline compensation rule: Value lies between [2, 100]

- If they both announce the same number that is what they get ($x = y$).
- But if the numbers are $x < y$, then,

  the one who writes **x** gets x + **R**

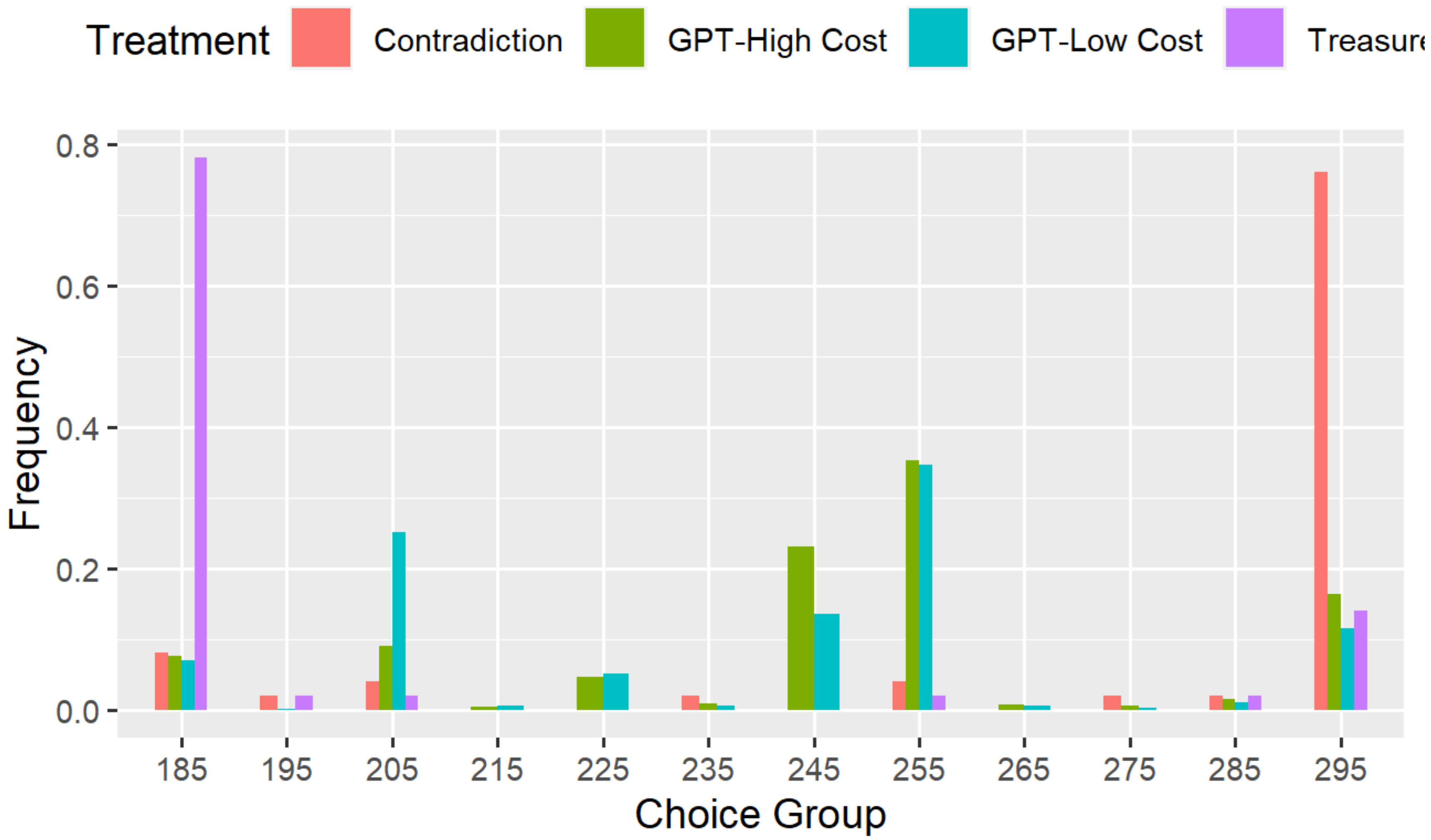
  the one who writes **y** gets y - **R**

Nash Equilibrium (NE)?

In the experiment: Value lies between **[180, 300]**

R is either **5** (Contradiction) or **180** (Treasure)

TD: Results



Treatment
Contradiction
GPT-High Cost
GPT-Low Cost
Treasur
0.8
0.6
0.4
0.2
0.0
Frequency
185
195
205
215
225
235
245
255
265
275
285
295
Choice Group

TD: Results

- Human behavior: In the *Treasure* treatment, experimental subjects find the NE but

  not in the *Contradiction* treatment.

- ChatGPT:
- 1. Cannot find the NE.
- 2. For ChatGPT it does NOT matter if **R** = 5 or 180

  ⇒ Cannot engage in strategic reasoning like humans.

“*We are in the phase of learning the secrets of AI*.”

- Both human beings and ChatGPT may fail to play the (theoretical) equilibrium strategy.

- The need to trust the other player may explain why ChatGPT does not play the equilibrium outcome in many cases.

- In many instances – especially in the Treasure treatments – human subjects do play equilibrium or close to it.

“*We are in the phase of learning the secrets of AI.*”

- When both fail to play the equilibrium behavior, ChatGPT’s behavior is <u>not aligned </u>with human behavior

  ⇒ *ChatGPT’s failure is NOT due emulation of human behavior.*

- Not surprisingly, this needs to be understood further and has implications for both Ethics and Higher Education.

ssarangi@vt.edu

@sudiptahere