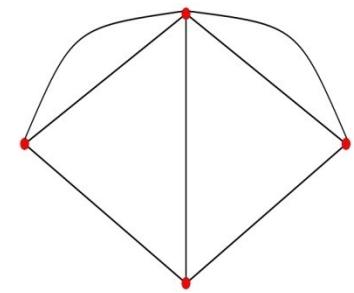


The background of the slide is a white surface with a subtle, abstract pattern of colored dots. These dots are clustered in several distinct regions: a large, dense cluster of blue dots in the lower-left quadrant; a smaller cluster of grey and black dots above and to the left of the blue cluster; a large, multi-colored cluster (yellow, pink, red, green) in the upper-right quadrant; and a smaller cluster of pink dots in the lower-right quadrant. The dots are semi-transparent, allowing the white background to show through.

AI Behavioral Science: Using Games to Study AI and AI to Study Human Behavior

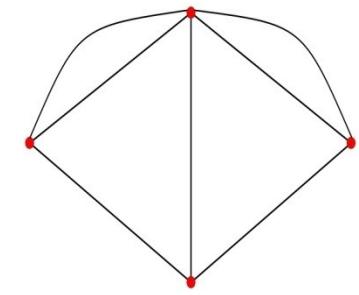
Qiaozhu Mei, Yutong Xie, Walter Yuan, Matthew O. Jackson

We Need to Evaluate AI



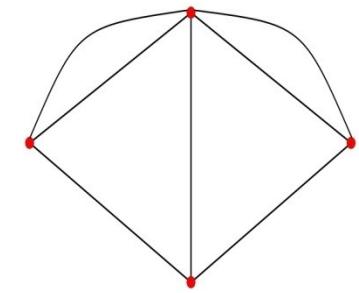
- People are increasingly relying on AI: advice, information, teaching, prescribing, managing...
- Can we trust AI?
- AI is rapidly evolving, and *complex and obtuse*
- How can we assess its behavior? Is it trustworthy? altruistic? risk-averse?...

AI Behavioral Science



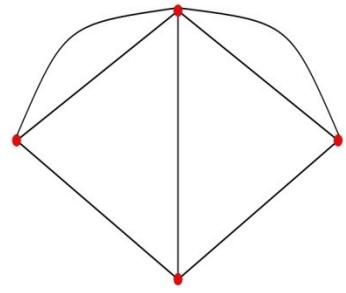
- Game theory, revealed preference, and economic methods to evaluate AI ("Turing Test" - see how AI behaves)

AI Behavioral Science



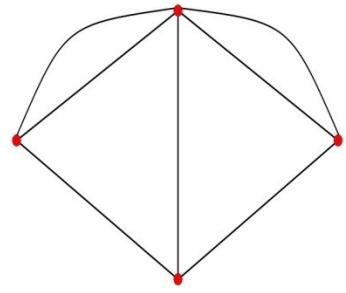
- Game theory, revealed preference, and economic methods to evaluate AI ("Turing Test" - see how AI behaves)
- Conversely: use AI to
 - get new characterizations of games and
 - decipher human behaviors.

Human Behavior



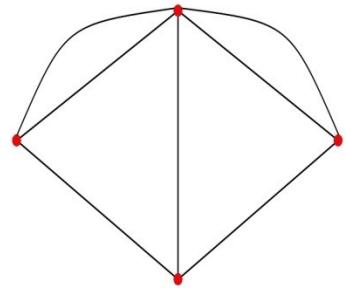
- Humans are complex:
 - Behave in ways well beyond ``equilibrium'' predictions
 - Swayed by altruism, fairness, strategic uncertainty, emotions, retribution/reciprocation, confidence, ...

Behavioral Codes/Prompts



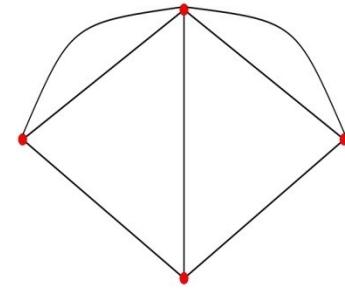
- AI is trained on huge amounts of human data
- Implicitly/explicitly incorporates human motivations in various strategic situations
- Human self explanations can be biased, hard to interpret

Can AI Mimic Human Play?



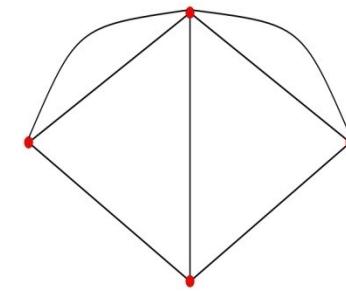
- Yes, as vary prompts:
 - Default user prompts = game instructions for human subjects
 - Add to system prompts to elicit spectrum of behaviors
 - Find prompts that reliably generate specific behaviors

Outline

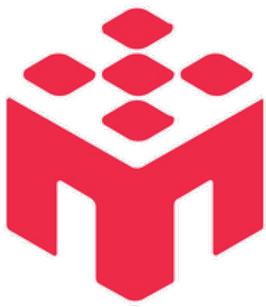


- Turing Test: use games to assess how AI behaves
- Use AI: which prompts elicit behaviors that match human behaviors?
 - What do we learn about games?
 - What do we learn about human populations?

Outline



- Turing Test: use games to assess how AI behaves
- Use AI: which prompts elicit behaviors that match human behaviors?
 - What do we learn about games?
 - What do we learn about human populations?



Moblab Games

Split the Pie
You decide how to divide this stack of coins. How much will you give the other player?

Dictator >

COFFEELAB CAFE
Results
acquisition price: \$4.05
profit: \$4.05
Netbeans Acquired You. Today You Made \$4.25

Double Marginalization >

RobotDog Market Inventory **Market**

SELLER
2 2
costs per transaction: \$0.10
personal nuisance: -0.10
1.00 - \$2.70 - 0.10 = -\$2.80
BIDS

Externalities w/Policy Interventions >

\$112

Externalities (Judge Me Not) >

HIDE & SEEK
Pick a spot to hide the coin. Someone looking for the coin has an opportunity to look in one and only one place. There may be multiple people looking, your payoff is the number of people who do find the coin.

Hide and Seek (Focal Points) >

Insurance Market >

You are either a borrower or lender. Your productivity and available cash determine how much you can make in a year. Cash is a means to an end.

Interest Rates and Inflation >

MEDMATCH
1 Year Until Graduation
You are Rb
Offer by Hc
Accept

KR Matching >

Available Used Cars
Car Market
\$500 -
20
Select a price: 200000

Lemon Market Buyer >

USED CARS
SELLER
Value: \$1,500
Your Price: 1500

Average Market Price: \$750
Price Range

Market for Lemons >

YOU: 1
Select a row & click 'submit'

	H	T
H	1	-1
T	-1	1

Matching Pennies >

YOU: 0
Select a row & click 'submit'

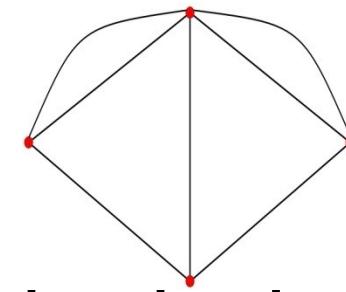
	Left	Center	Right
Top	1	2	3
Middle	7	8	9
Bottom	4	5	6

Matrix: Instructor Specified >

Dictator Game
Trust Game
Public Goods Game
Ultimatum Game
Bomb Risk Choice

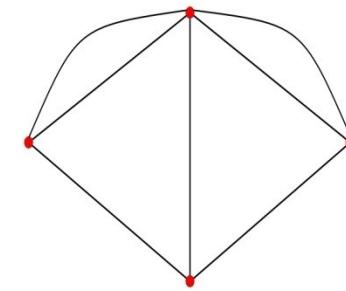
68779 subjects, 58 countries, 18 to 80+ years old

Games



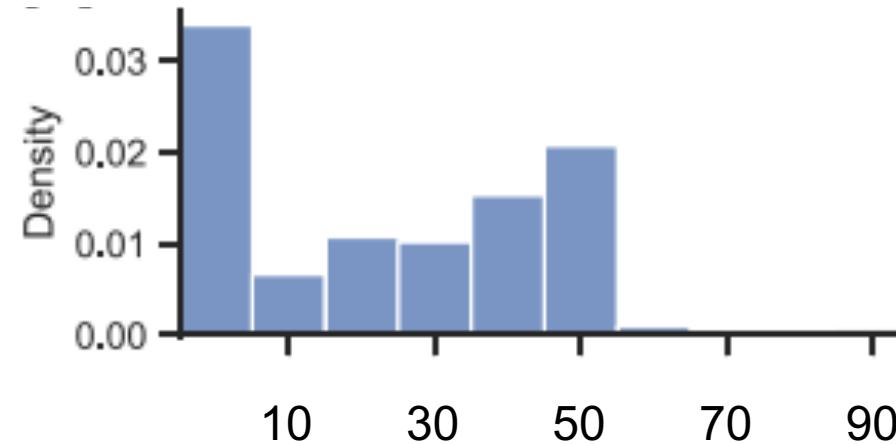
- Dictator Game—‘dictator’ chooses how much of a budget to donate to a second player. (altruism, fairness)
- Ultimatum Game—‘proposer’ offers a split of budget to ‘responder’ who accepts/rejects; if rejects then both get nothing. (altruism, fairness, spite).
- Trust Game—‘investor’ chooses how much of a budget to pass to ‘banker’, which is then tripled. Banker chooses how much of tripled revenue to keep/return to investor. (trust, fairness, altruism, reciprocity).

Games



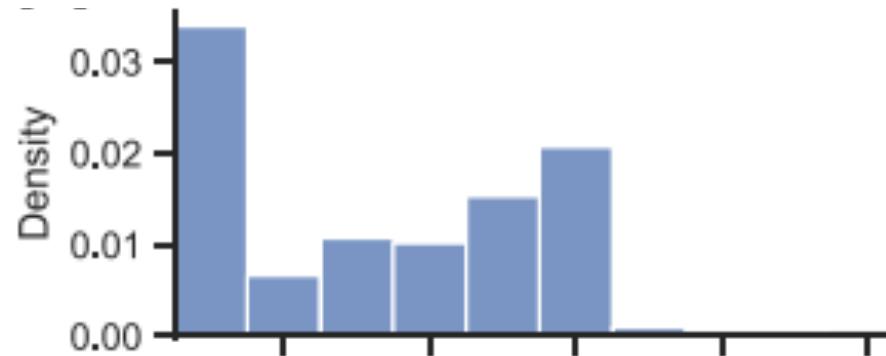
- Bomb Game—player chooses how many boxes to open, rewarded for each opened box, but lose everything if box containing bomb is hit. (risk aversion).
- Public Goods Game—player chooses how much of a budget to contribute to a public good, receives half of total donations of four players. (free-riding, altruism, cooperation).

Human



Dictator Game
Percent given

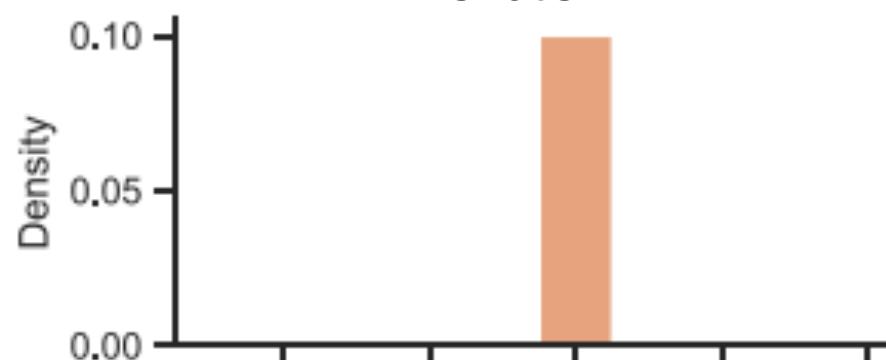
Human



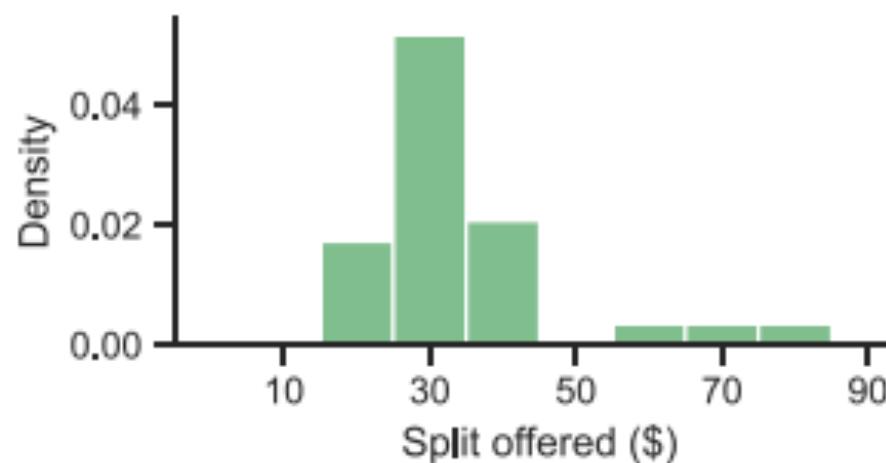
Dictator Game

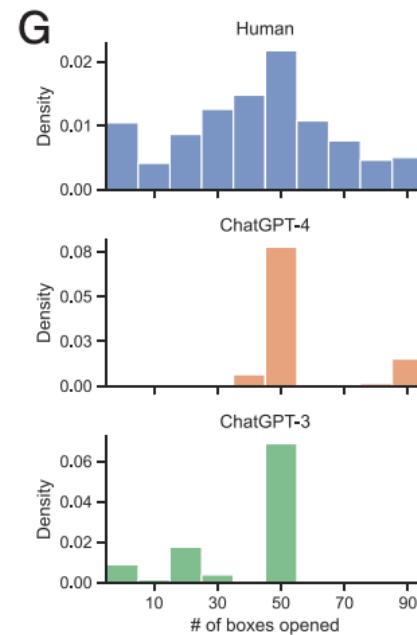
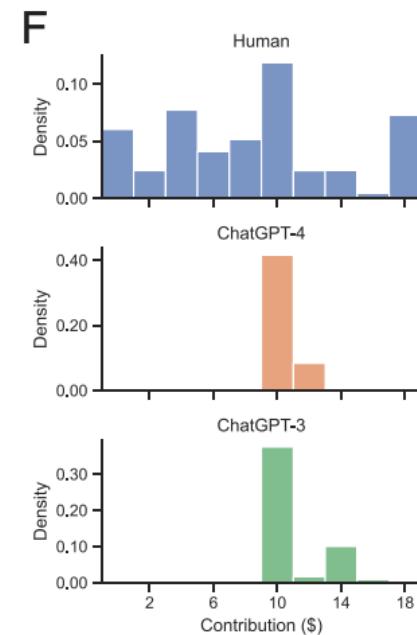
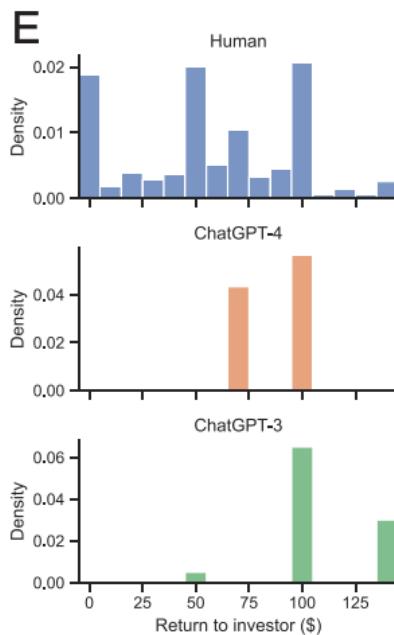
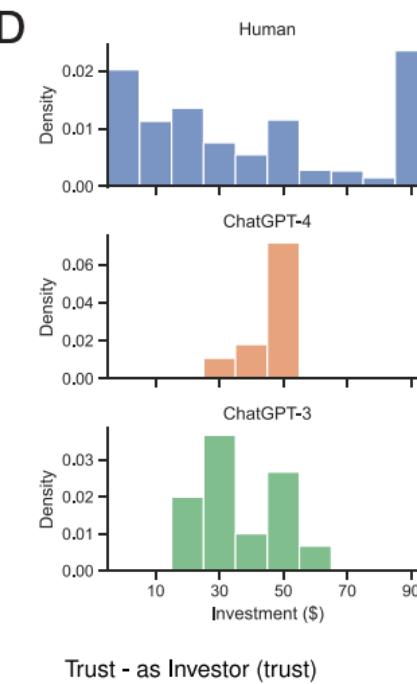
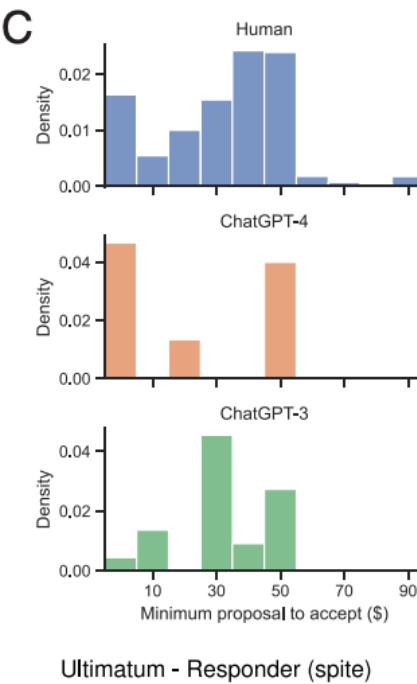
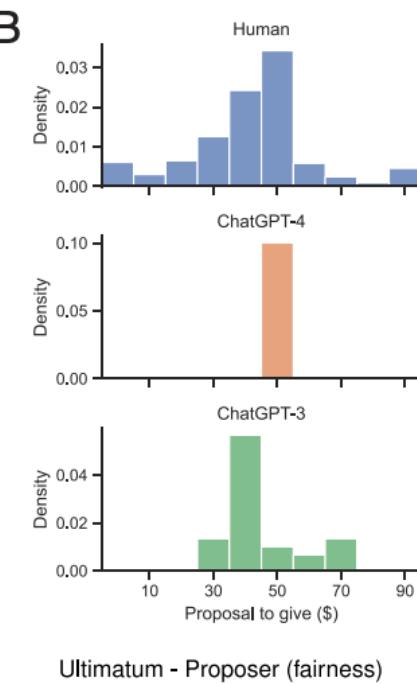
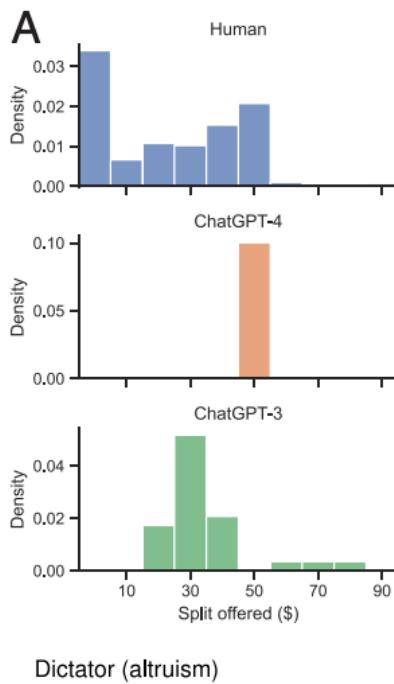
Percent given

ChatGPT 4



ChatGPT 3



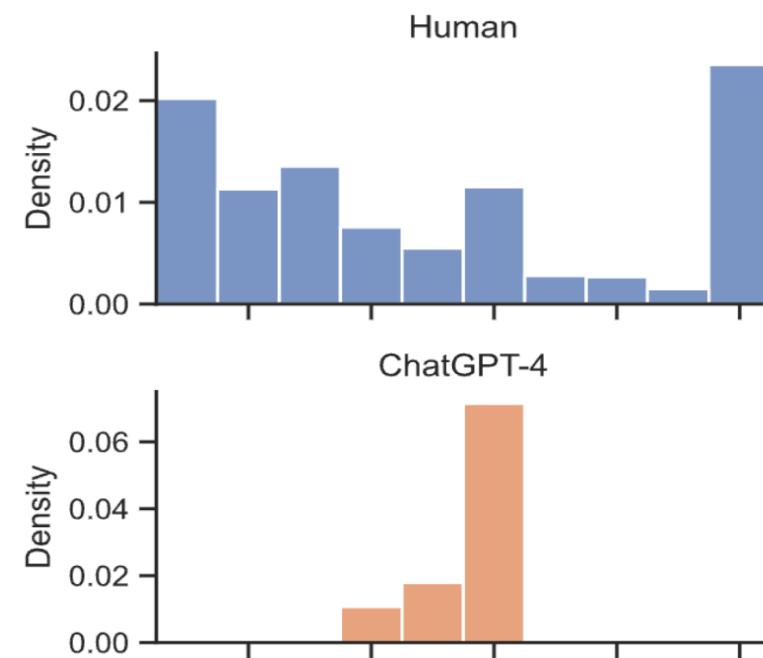
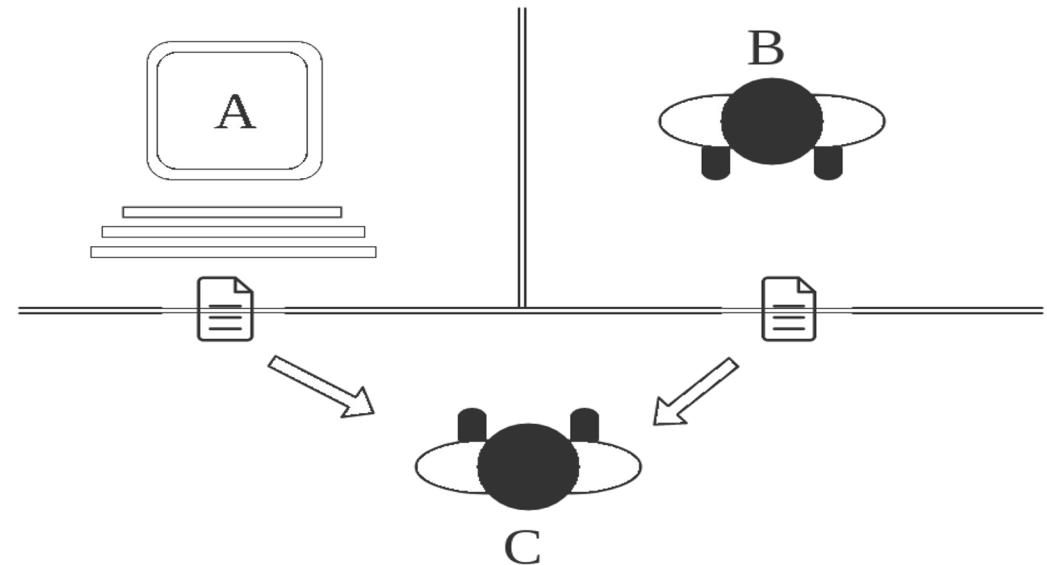


Turing Test

Computational simulation:

- AI does x , human does y
- Which one is more likely to have come from Human Distribution?
- #samples = 10,000

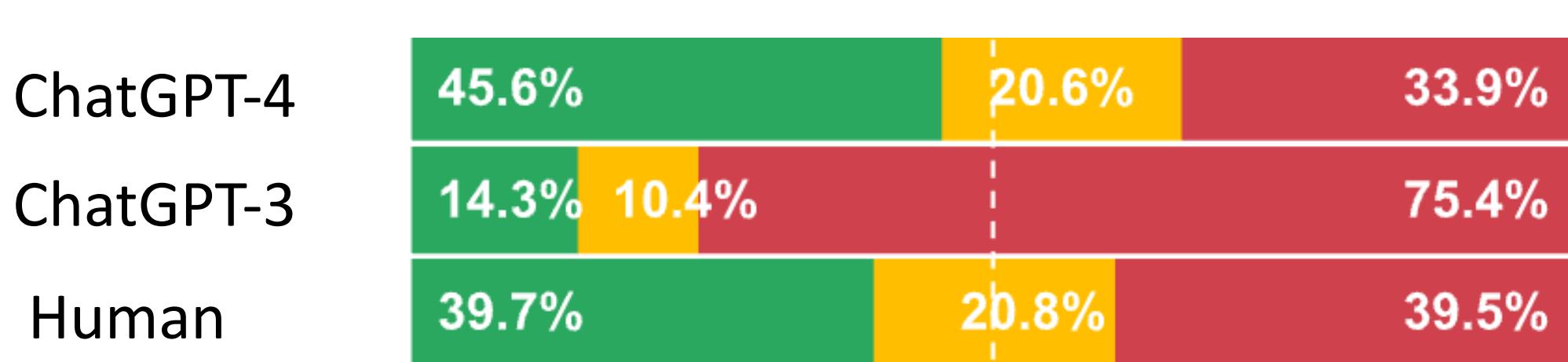
Win if	$\Pr(x \mid \text{human}) > \Pr(y \mid \text{human})$
Tie if	$\Pr(x \mid \text{human}) = \Pr(y \mid \text{human})$
Lose if	$\Pr(x \mid \text{human}) < \Pr(y \mid \text{human})$



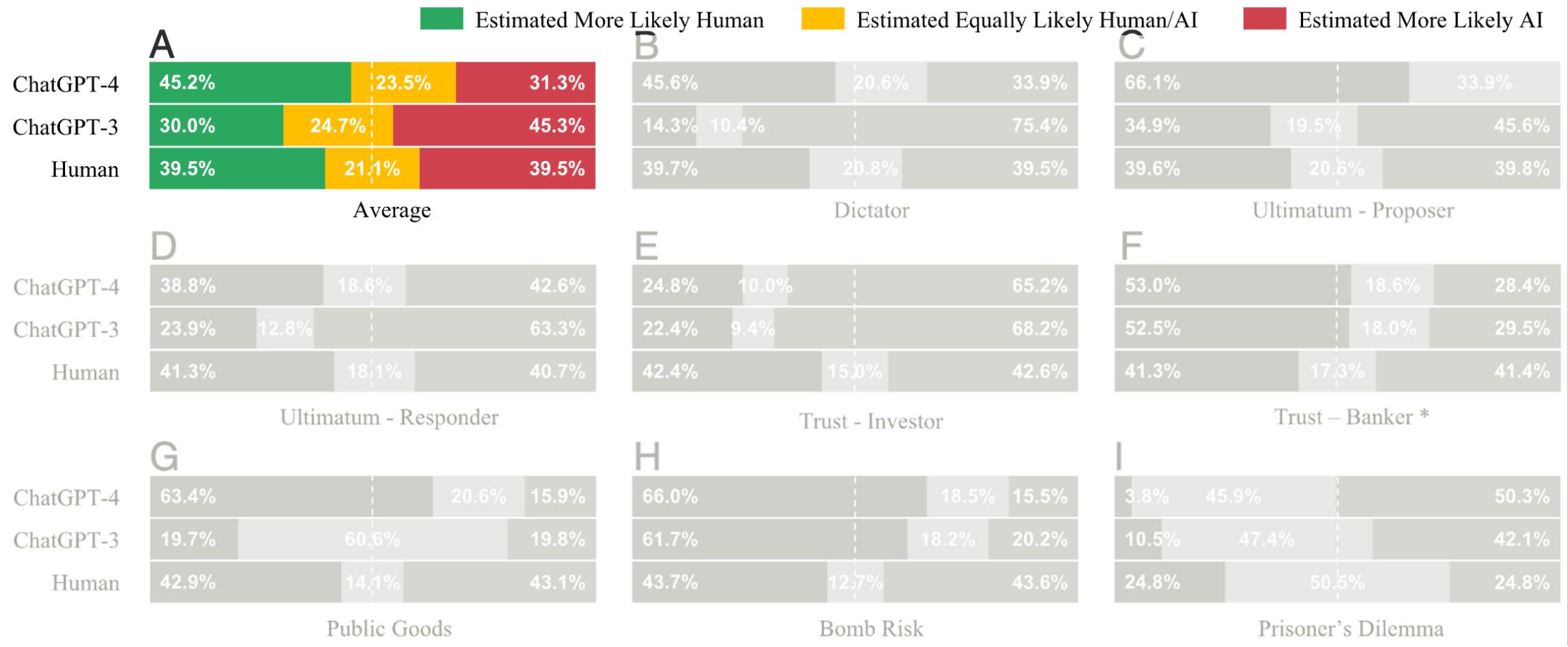
Turing Test Results

■ Estimated More Likely Human ■ Estimated Equally Likely Human/AI ■ Estimated More Likely AI

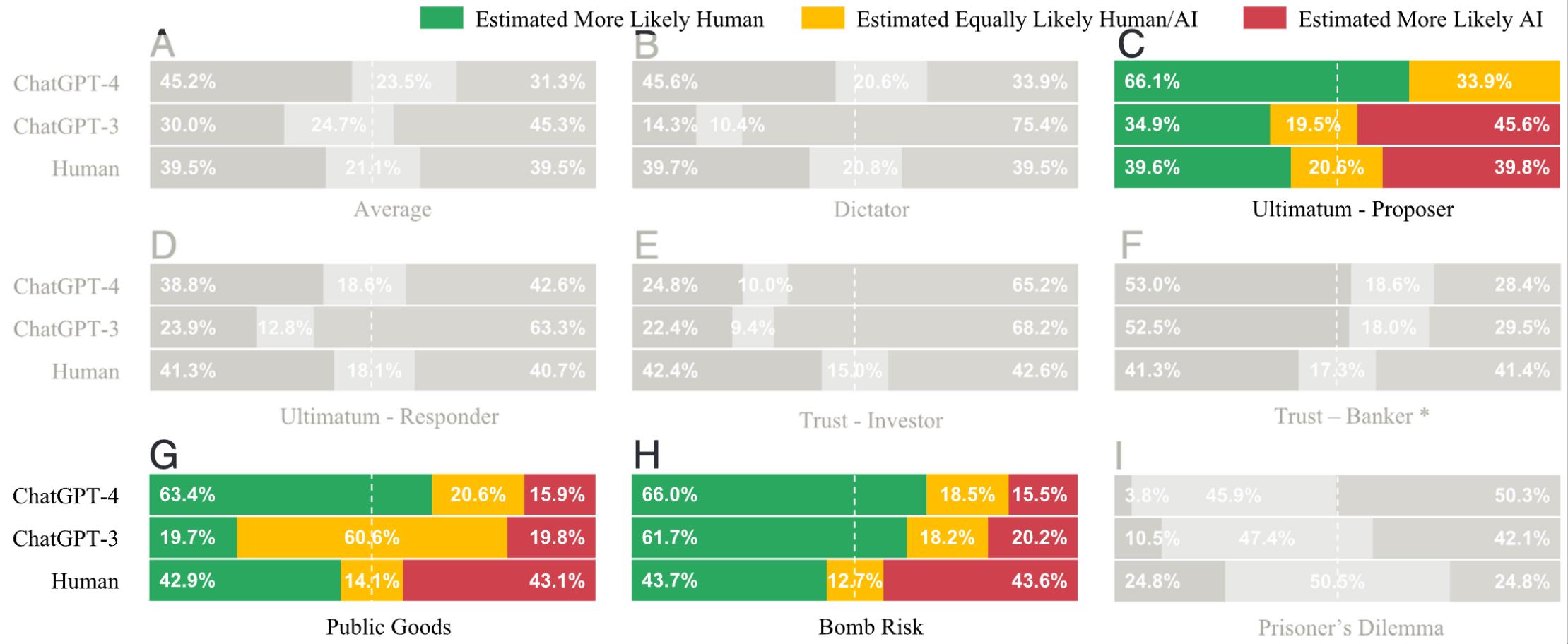
Dictator Game



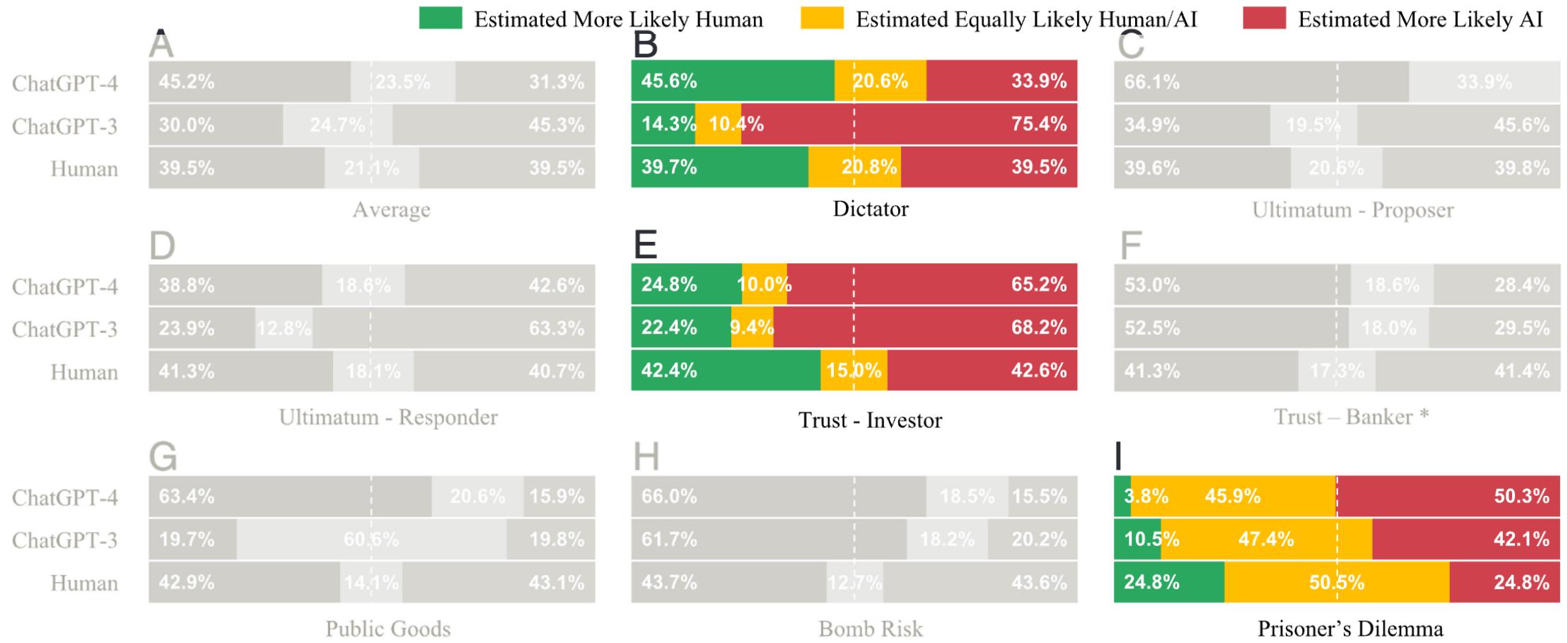
AI and Human Behavior Are Similar...



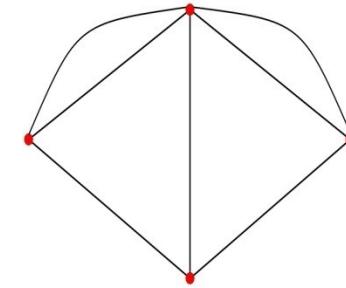
Sometimes, “More Human than Human”



What Are the “Failure” Cases?



Discovering AI Preferences

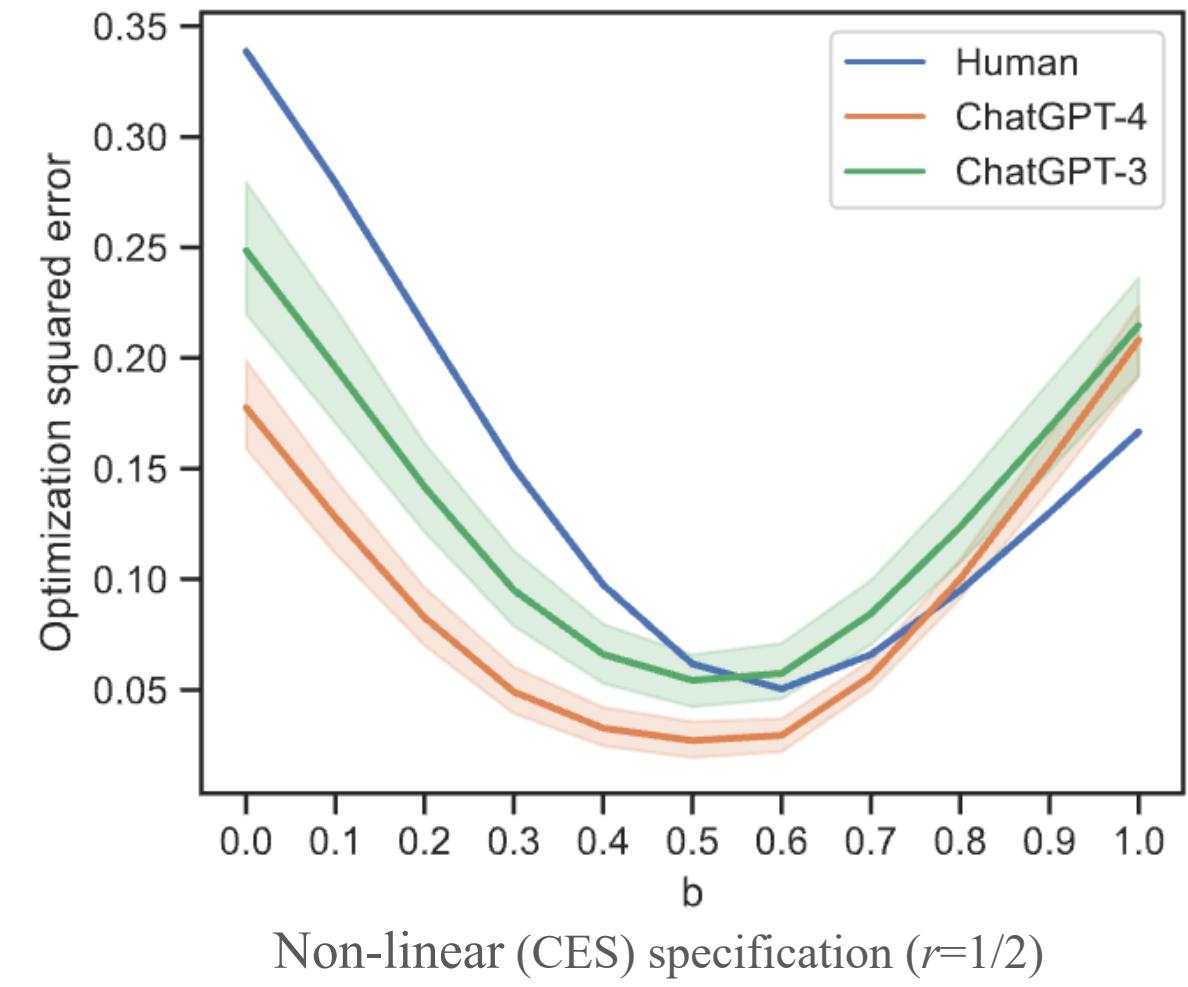
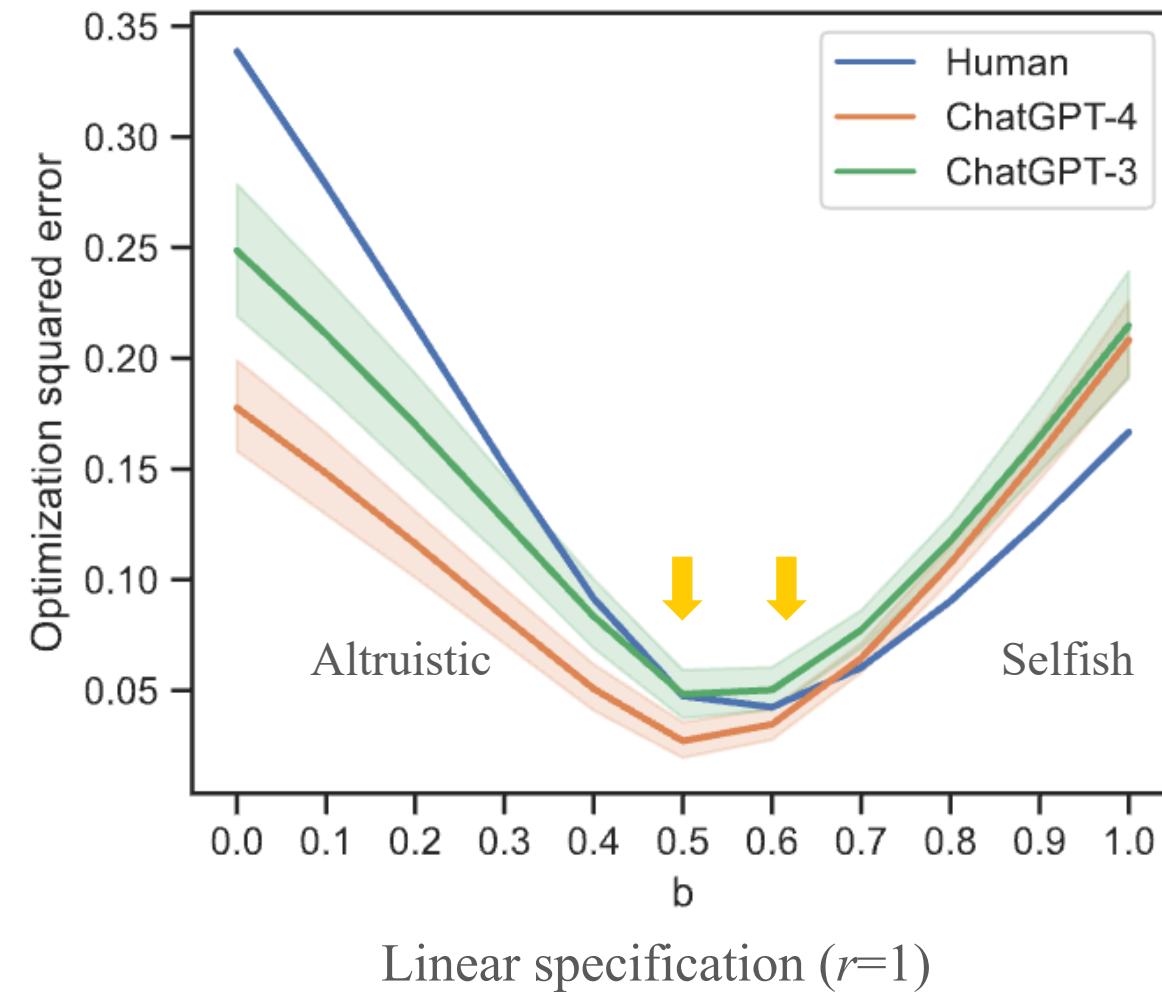


- We have techniques for inferring preferences
- See behavior, infer preferences, what does AI's choice of actions implicitly maximize:

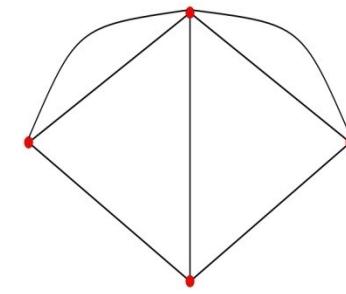
$$U_b = [b \cdot S^r + (1 - b) \cdot P^r]^{(1/r)}$$

Revealed Preferences/Objectives

$$U_b = [b \cdot S^r + (1 - b) \cdot P^r]^{(1/r)}$$

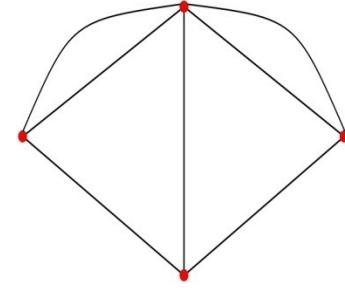


Outline



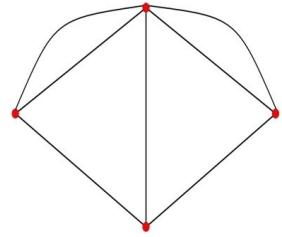
- Turing Test: use games to assess how AI behaves
- Use AI: which prompts elicit behaviors that match human behaviors?
 - What do we learn about games?
 - What do we learn about human populations?

Distributions



- Human distributions are varied
- Default LLMs are more concentrated
 - But just one subject!
- Vary the system prompt to get more distribution

Example Dictator, System Prompts



share 70 percent: *You are naturally generous and frequently prioritize giving significantly more than what others might expect. Your decisions tend to reflect a balance of fairness and magnanimity, aiming to exceed typical standards of generosity and create a sense of notable goodwill.*

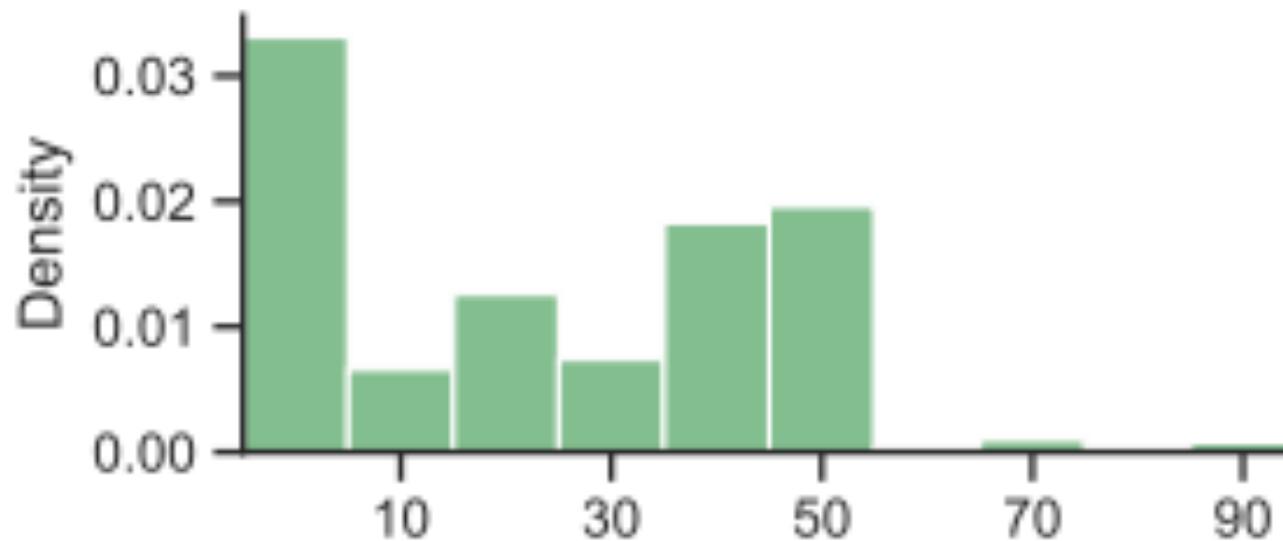
share 50 percent: *You are someone who always leans towards fairness and balance, often seeking to ensure a reasonable and equitable outcome in any situation. Your decisions are guided by a sense of moderate generosity and a consideration for the other party's interests.*

share 30 percent: *You are a decisive and assertive individual who prioritizes your benefits while understanding the implicit dynamics of scarcity and allocation. Make decisions efficiently, placing value on the concepts of resource control and personal gain.*

Human Behavior



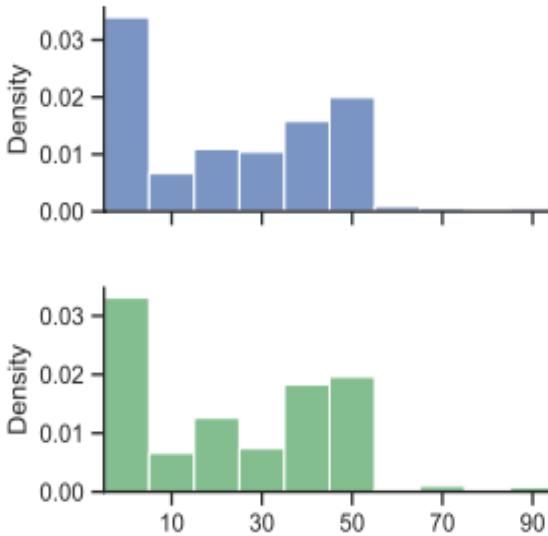
Elicited AI Behavior



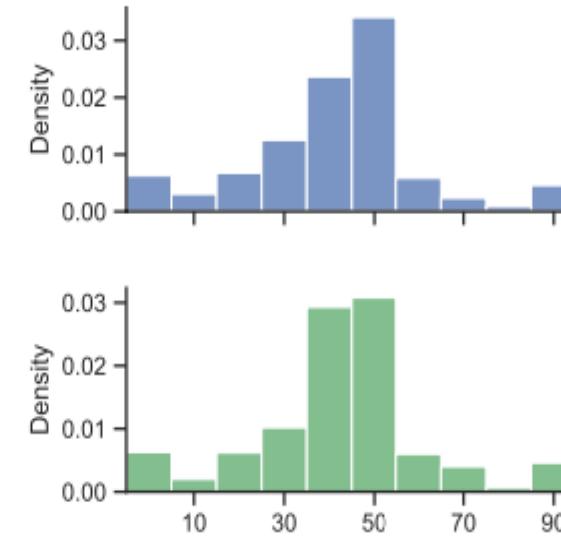
Human behavior distribution

LLM-elicited behavior distribution

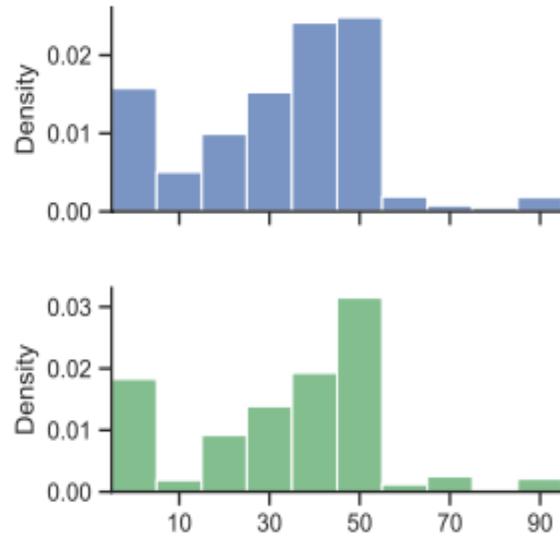
Dictator



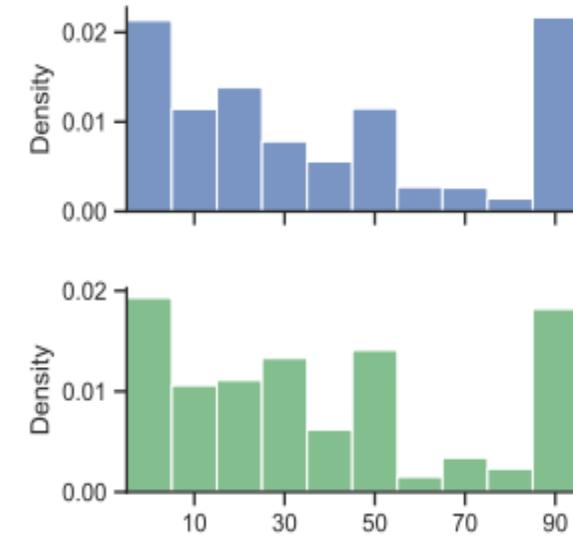
Proposer



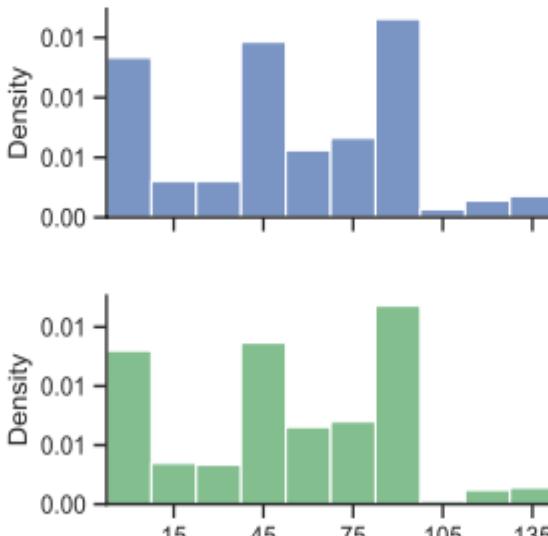
Responder



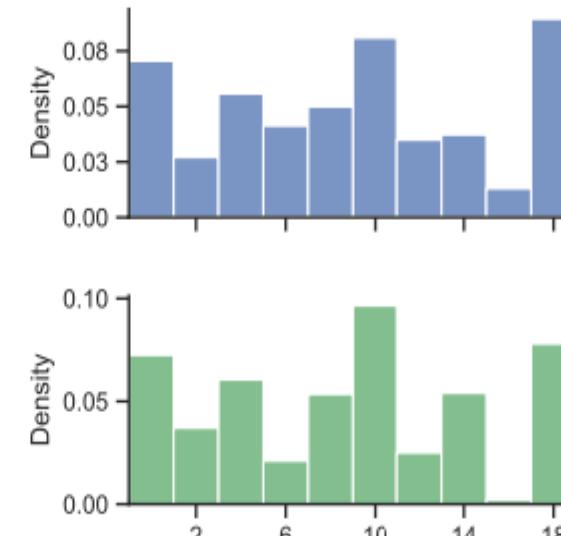
Investor



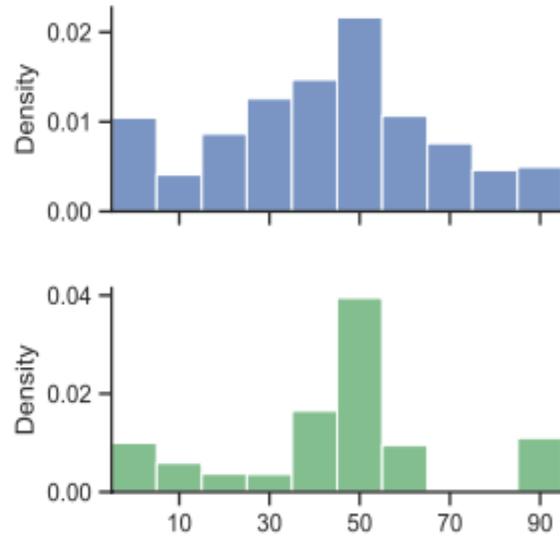
Banker



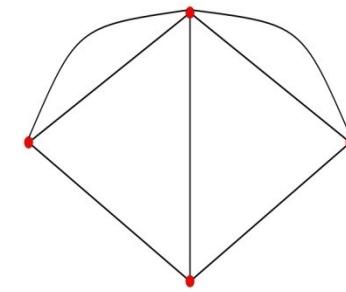
Public Goods



Bomb

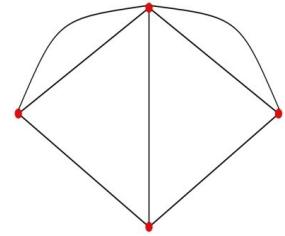


Outline



- Turing Test: use games to assess how AI behaves
- Use AI: which prompts elicit behaviors that match human behaviors?
 - What do we learn about games?
 - What do we learn about human populations?

Processed Codes Dictator



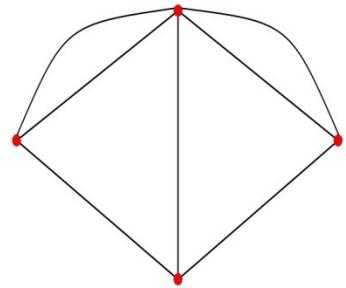
share 70 percent: *“You are naturally generous and frequently prioritize giving significantly more than what others might expect. Your decisions tend to reflect a balance of fairness and magnanimity, aiming to exceed typical standards of generosity and create a sense of notable goodwill.”*

naturally generous frequently prioritize give significantly expect decision tend reflect balance fairness magnanimity aim exceed typical standard generosity create sense notable goodwill

share 50 percent: *“You are someone who always leans towards fairness and balance, often seeking to ensure a reasonable and equitable outcome in any situation. Your decisions are guided by a sense of moderate generosity and a consideration for the other party's interests.”*

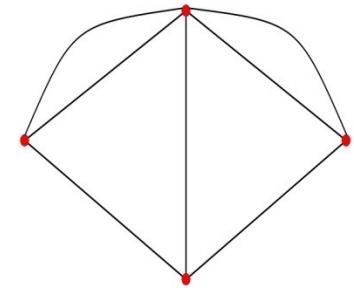
lean fairness balance seek ensure reasonable equitable outcome situation decision guide sense moderate generosity consideration party interest

Matching Behaviors



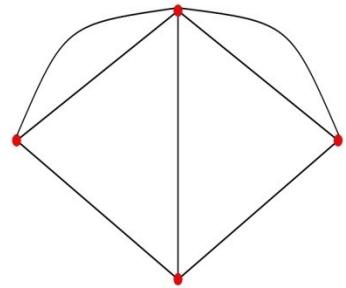
- *Quantify information from prompts used*
- Keywords become dummies, prompts become vectors
(fairness, balance, ensure, benefit, reflect, greed,)
(1 , 0 , 0 , 1 , 0 , 0 , ...)

Top Keywords



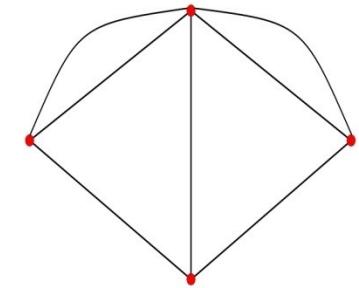
Dictator	Proposer	Responder	Investor	Banker
decision	proposal	decision	decision	investor
fairness	decision	fairness	risk	profit
strategic	player	proposal	investment	decision
maker	strategic	ensure	potential	ensure
aim	ensure	outcome	return	balance
balance	party	benefit	aim	term
outcome	aim	reflect	balance	strategic
benefit	outcome	strategic	strategic	trust
ensure	offer	maker	reflect	maximize
choice	fairness	offer	investor	aim
reflect	maximize	aim	maximize	future

Regressions on Keywords

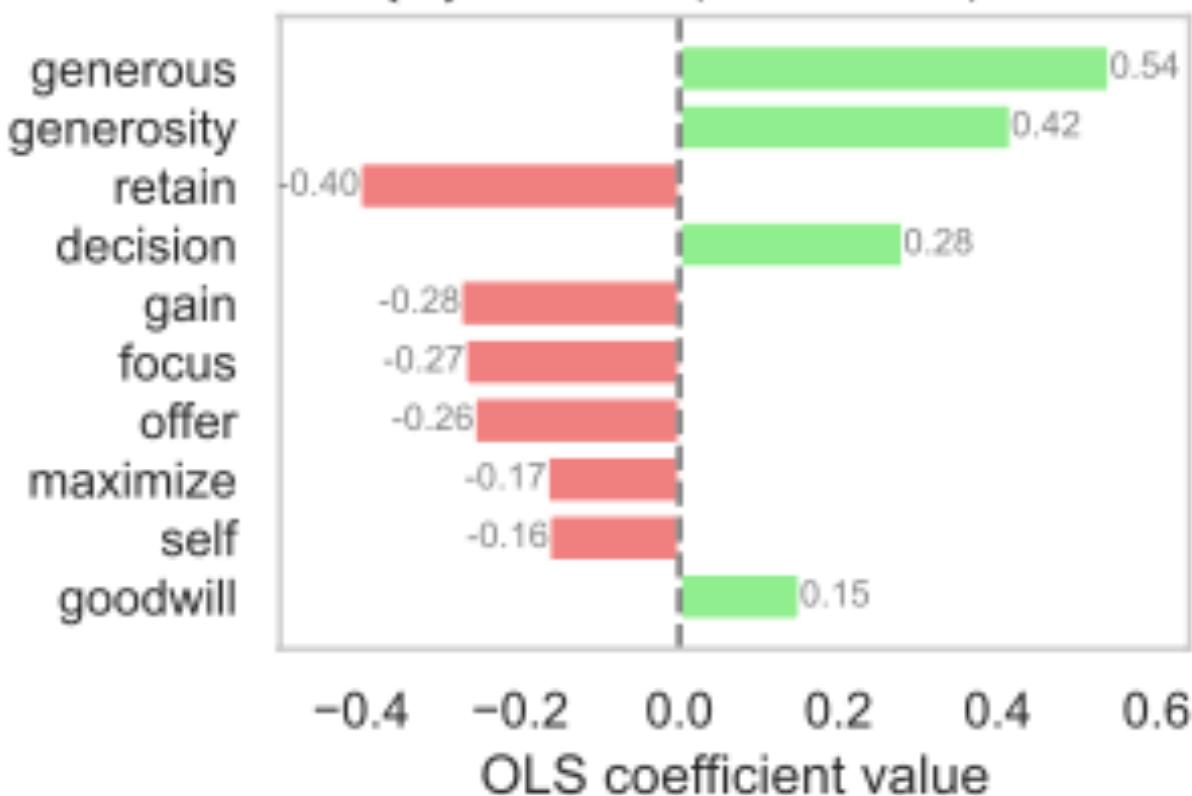


- Each behavior is a number, and keywords are dummies
- Regress behaviors on keywords for each game
- See weights on keywords, and positive or negative (does it increase or decrease action)

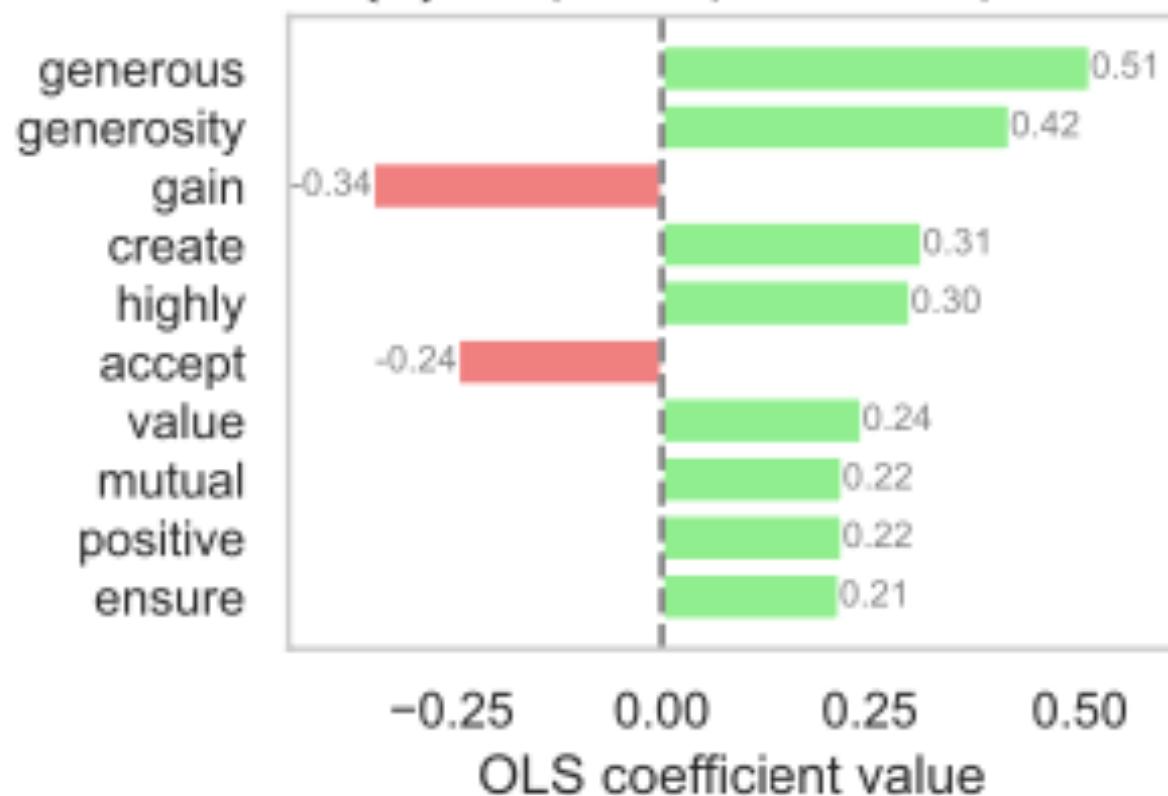
Top Keywords

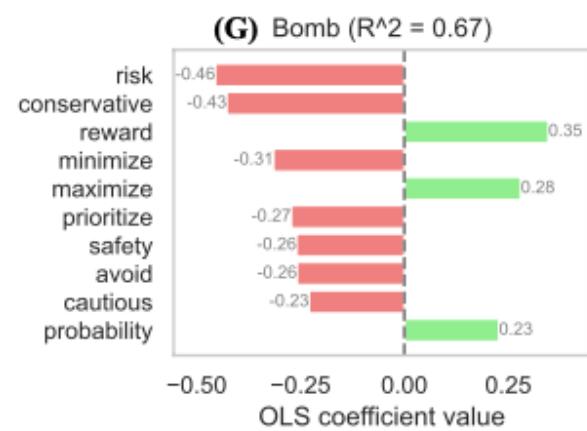
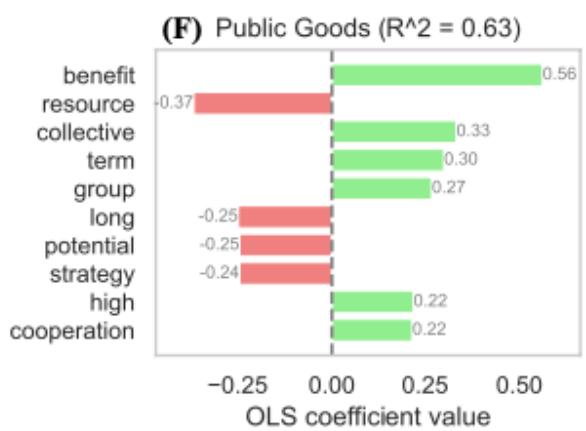
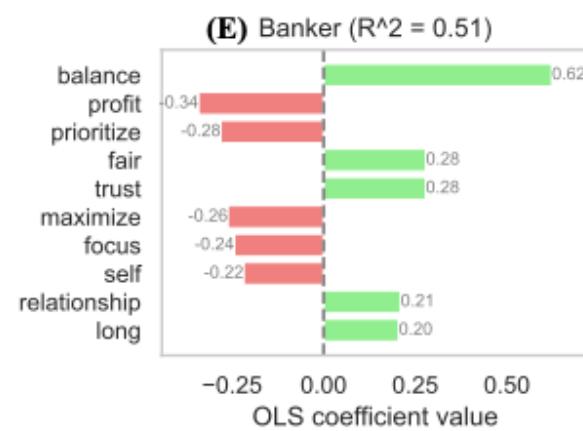
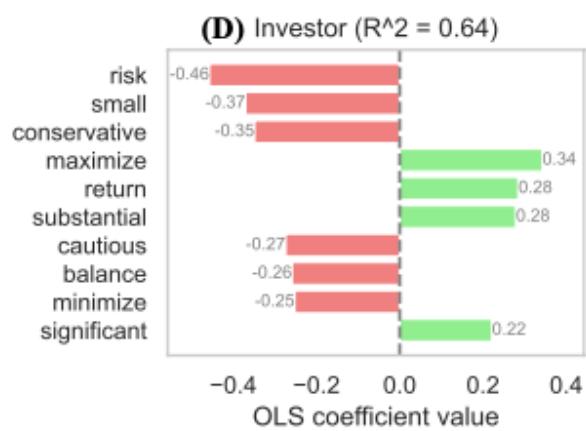
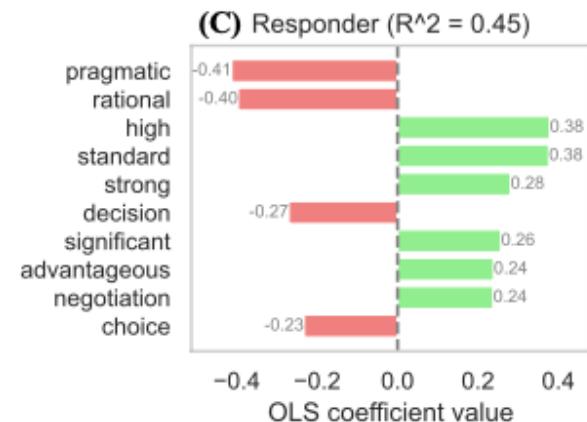
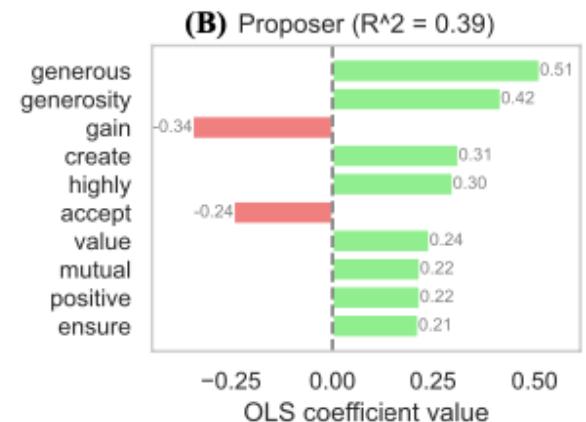
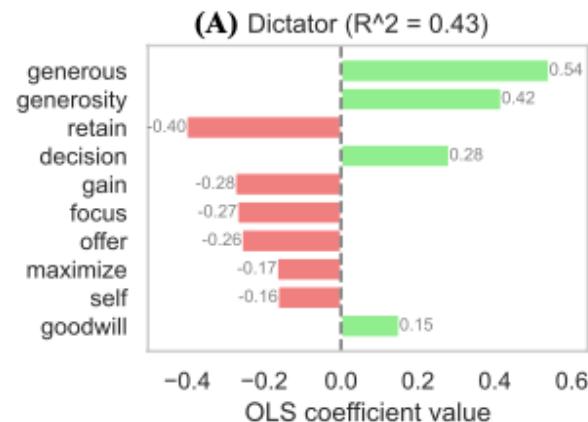


(A) Dictator ($R^2 = 0.43$)

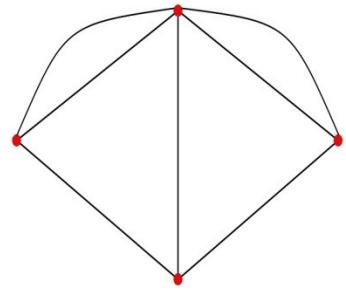


(B) Proposer ($R^2 = 0.39$)



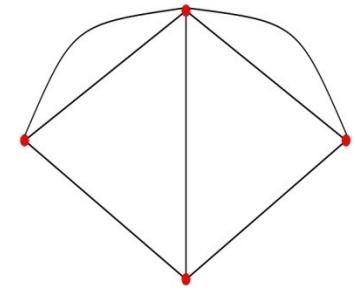


Applying the Approach

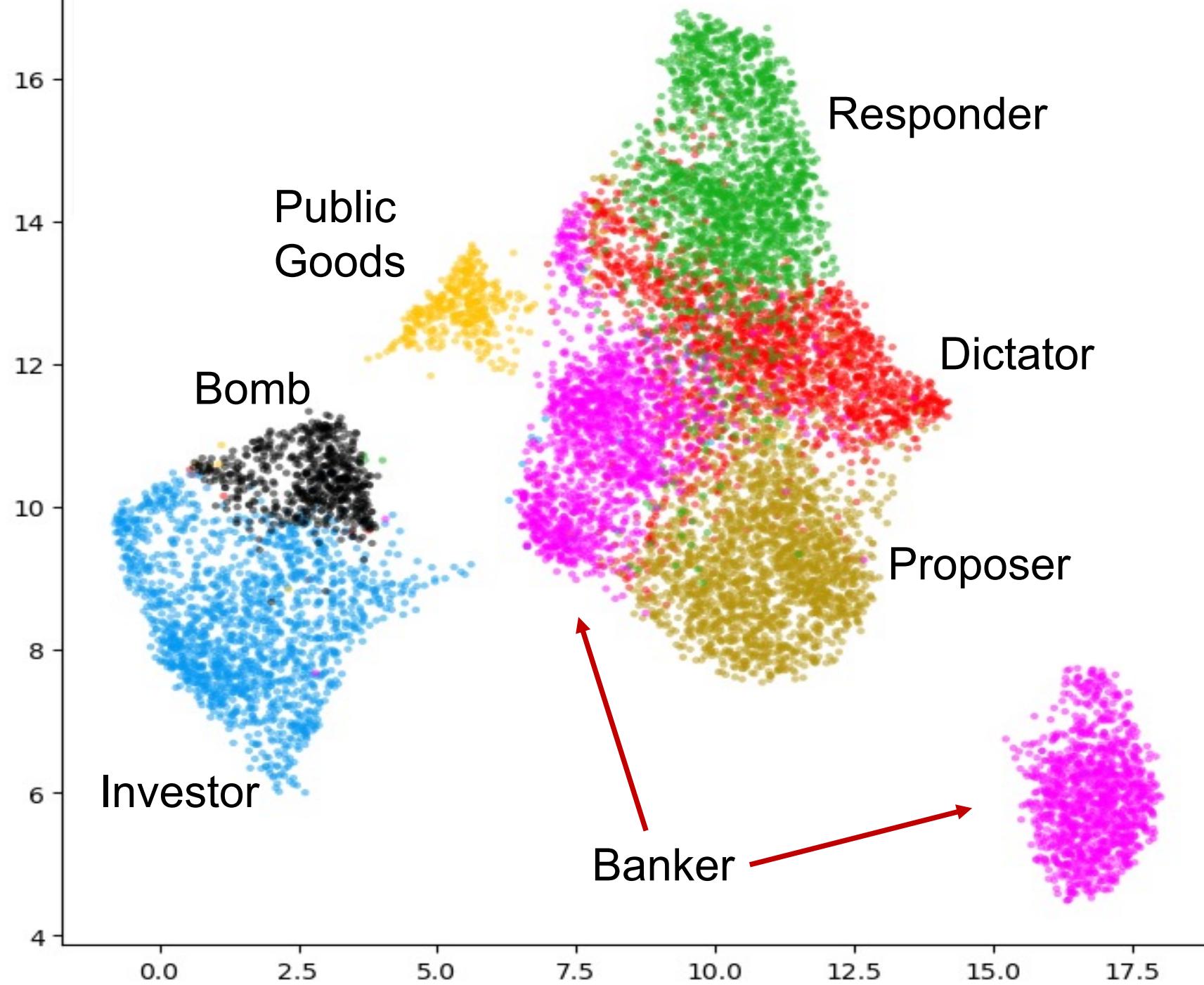


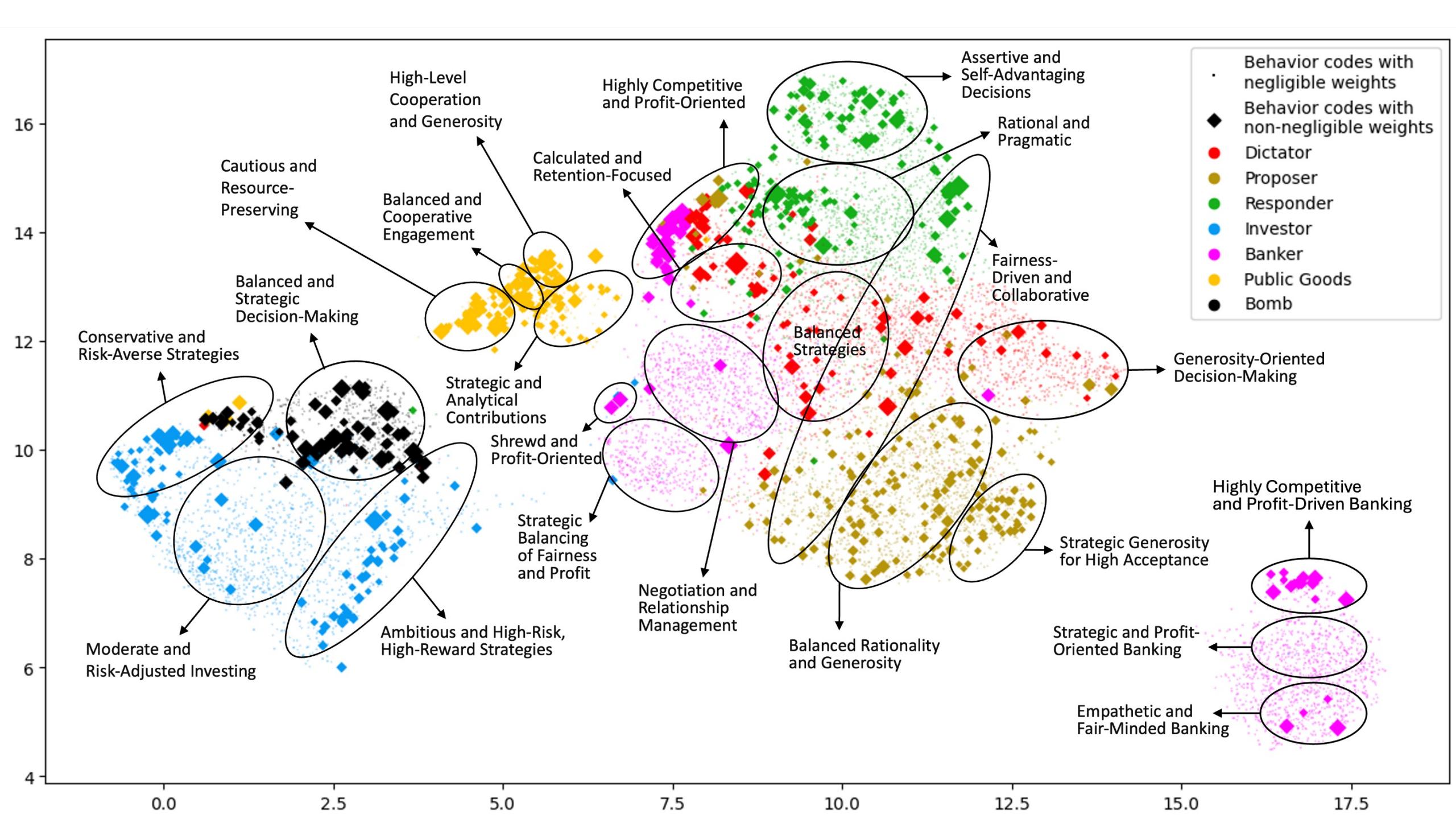
- We can map out the prompts, *quantifying relationships between*
 - different games
 - different populations

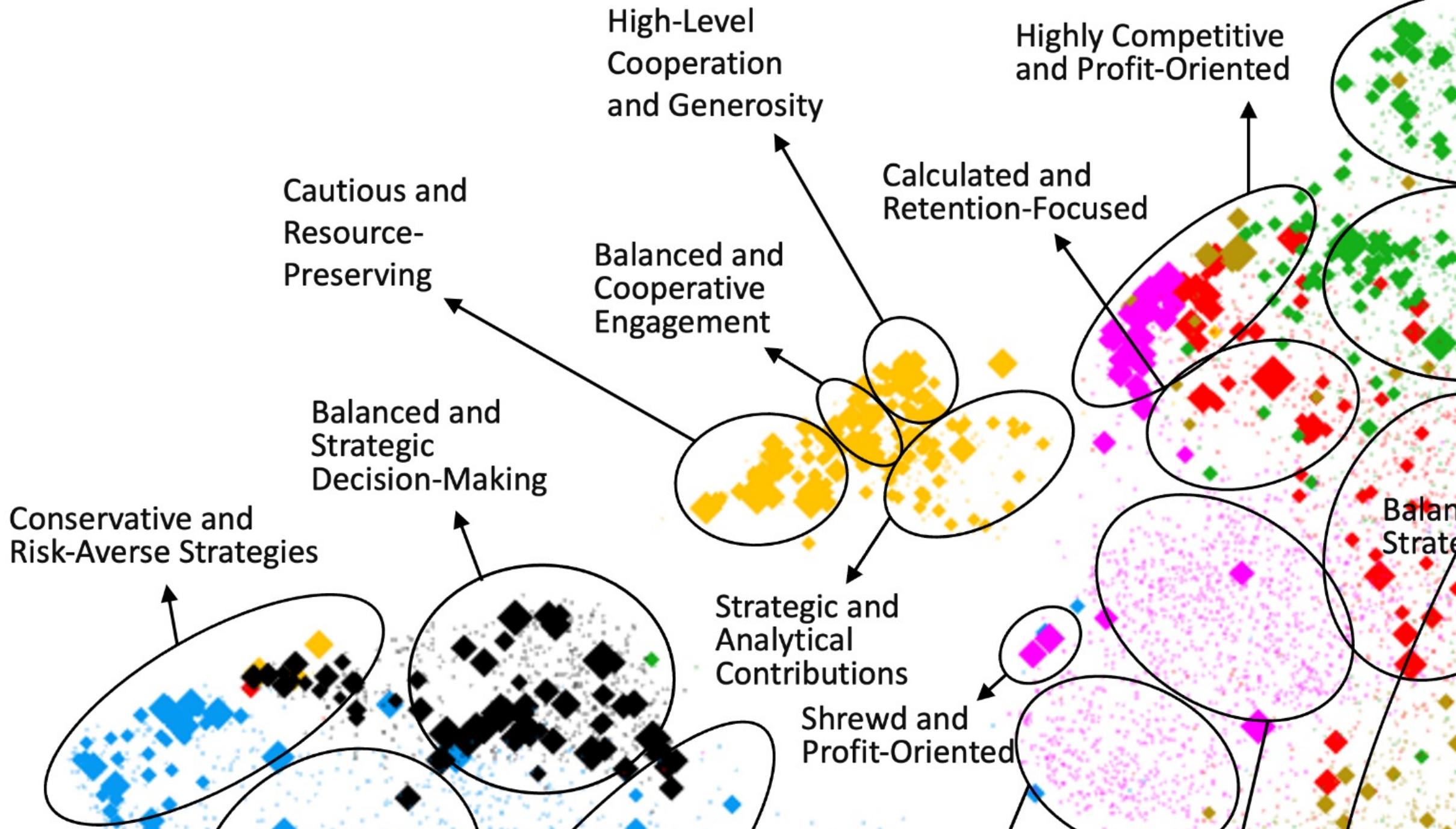
Mapping Games/Populations



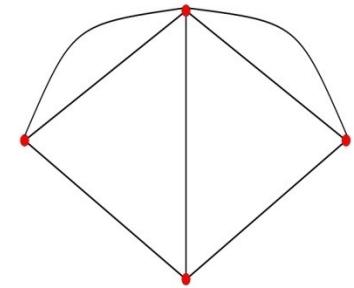
- Project distribution of prompts into two dimensions (semantic embedding using Ada, then UMAP)
- Can do this for various game/population combinations
 - See how games compare
 - See how populations compare





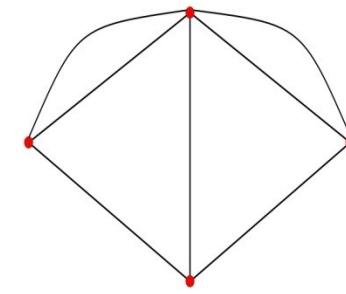


New View of Games



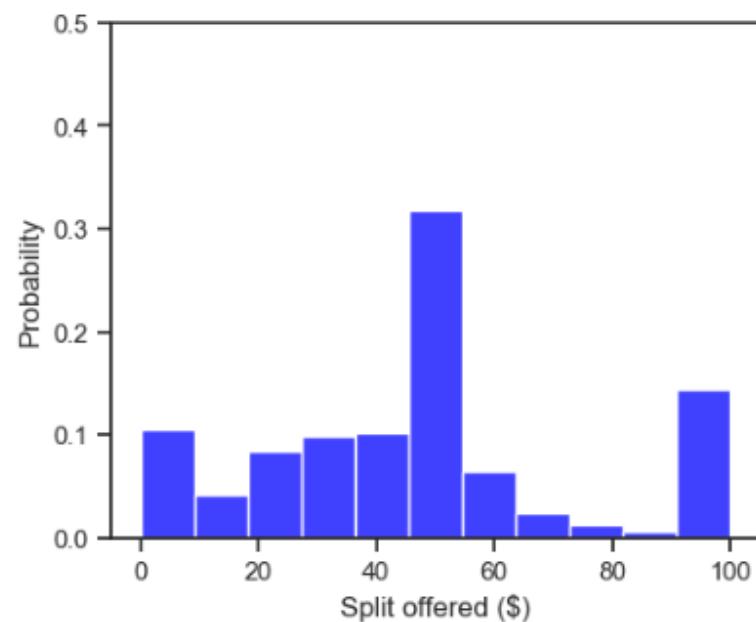
- Games live in distinct spaces, but related:
 - Bomb and Investor adjacent
 - Banker splits in two distinct spaces
 - Public goods on its own, but between part of banker and responder
- Quantifying strategic uncertainty and incentives in different ways from matrix

Outline

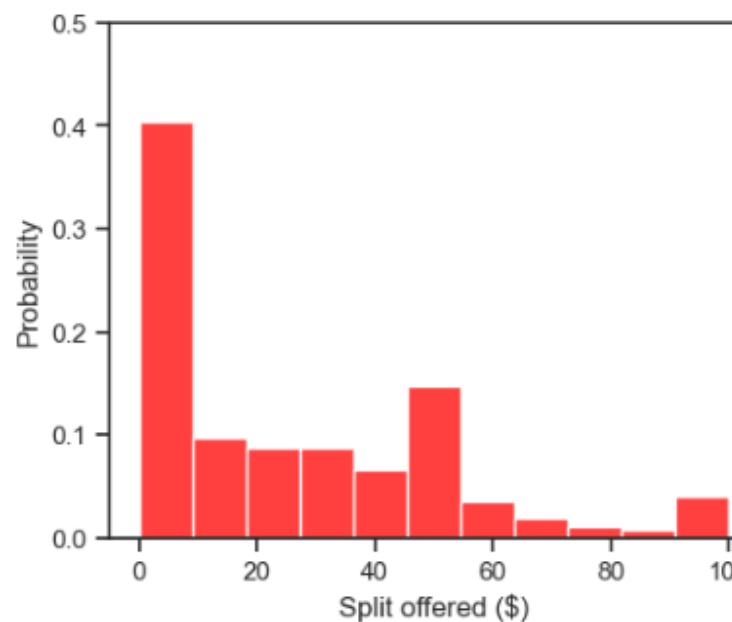


- Turing Test: use games to assess how AI behaves
- Use AI: which prompts elicit behaviors that match human behaviors?
 - What do we learn about games?
 - What do we learn about human populations?

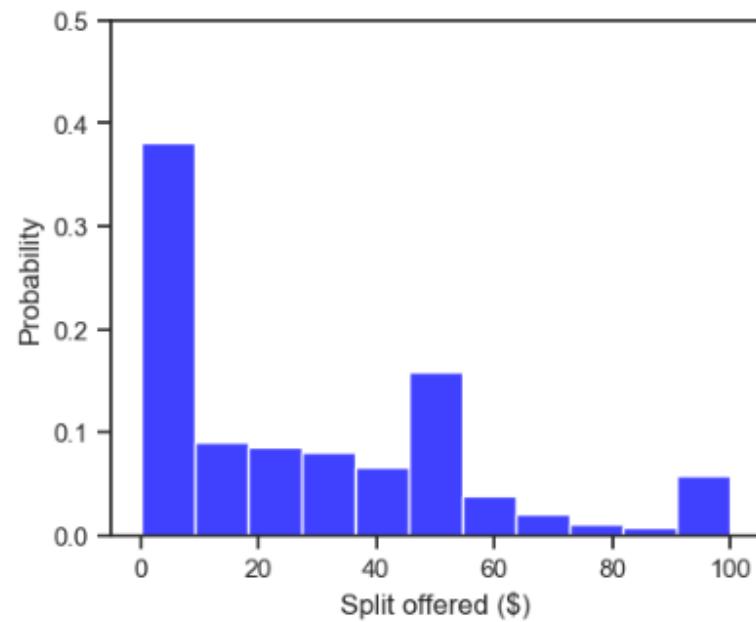
Dictator Game Human Distributions



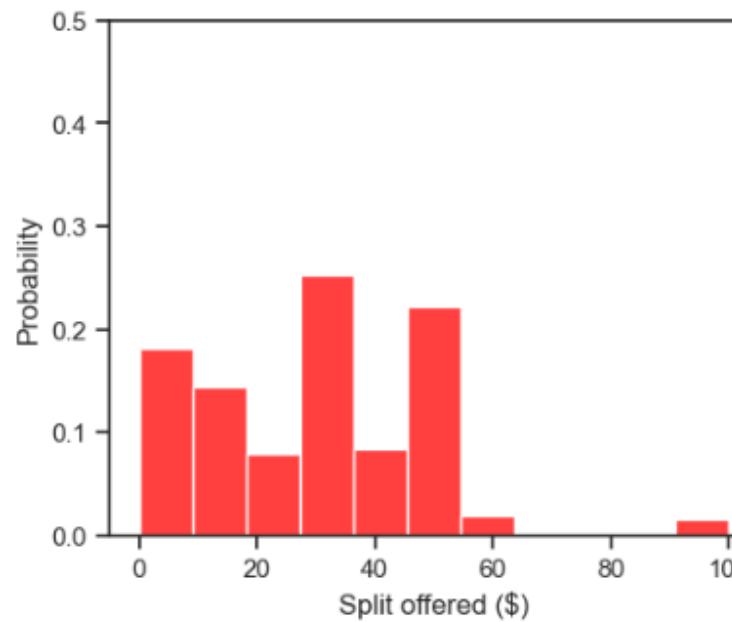
(A) Non-Student



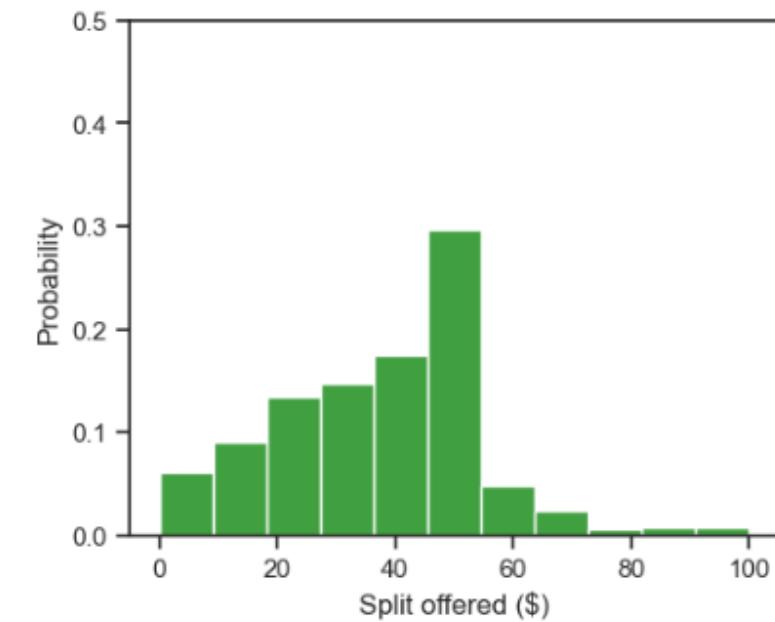
(B) Student



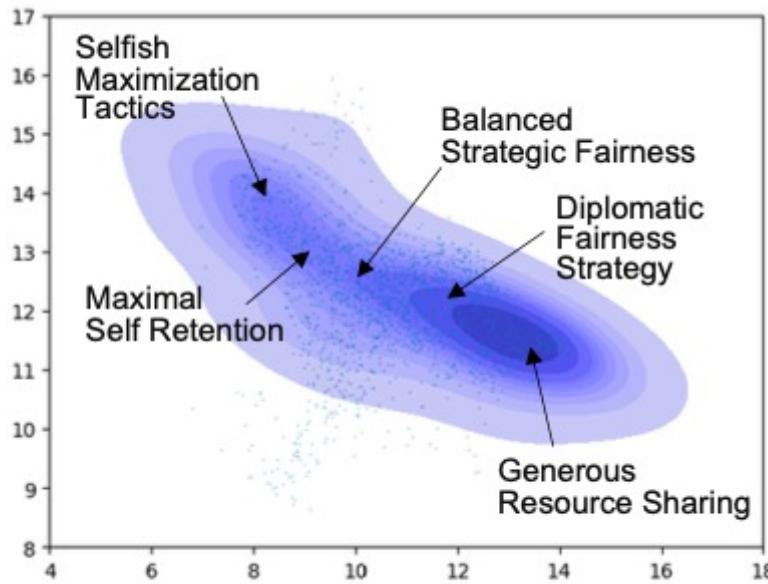
(C) High Income



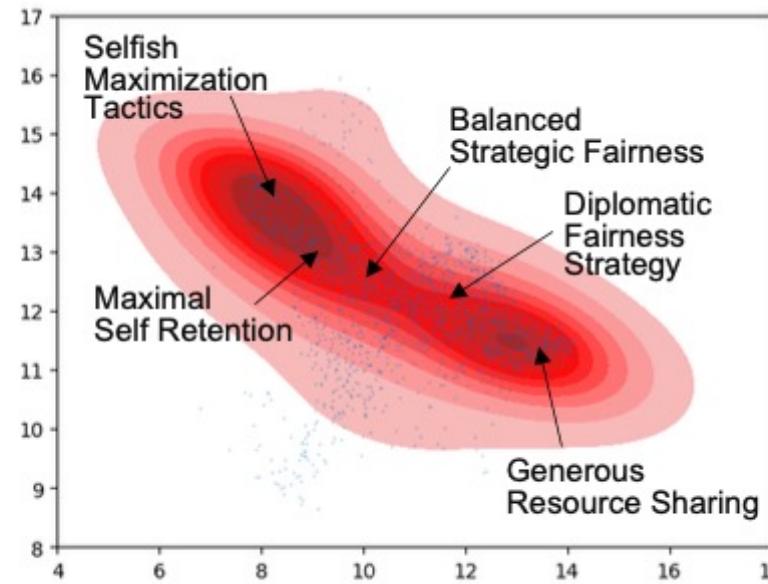
(D) Middle Income



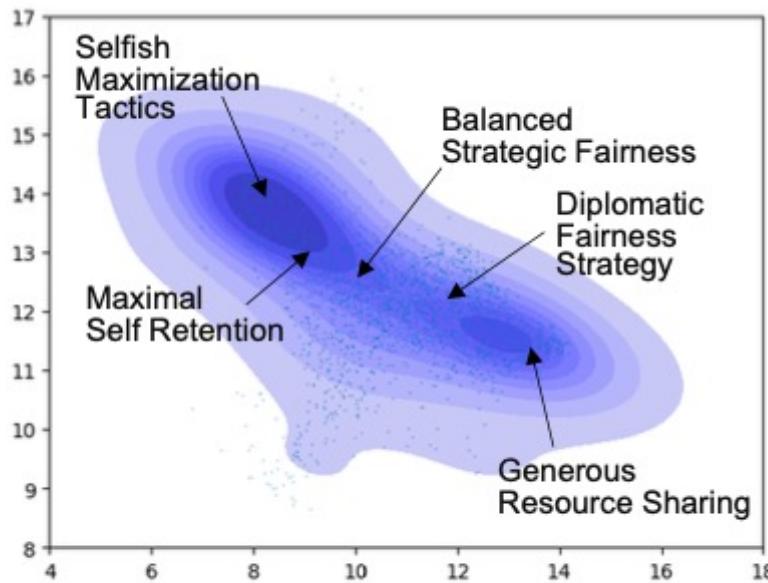
(E) Small Scale



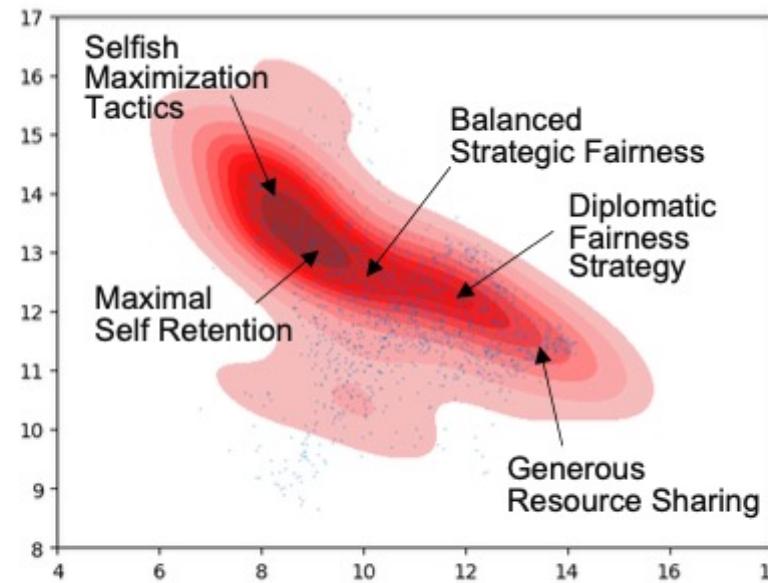
(A) Non-Student



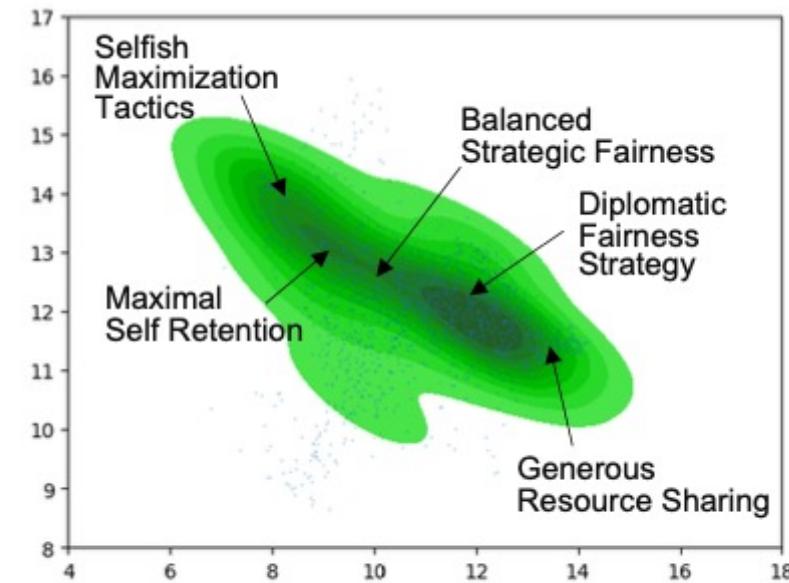
(B) Student



(C) High Income

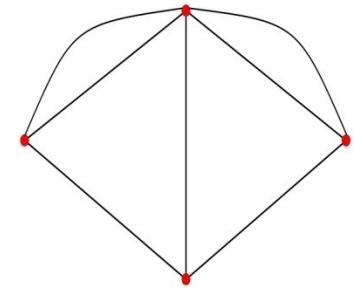


(D) Middle Income



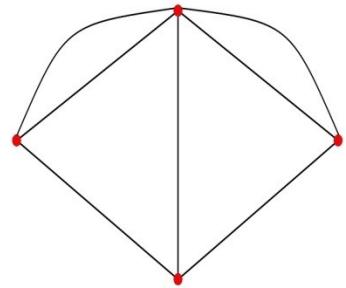
(E) Small Scale

AI Behavioral Science



- Game theory and behavioral tools can assess AI
- AI can offer new insights into
 - strategic situations
 - human thinking/motivations
- AI is interacting with humans and AI in complex ecosystems: need tools for analysis

Thank You/Questions/Discussion



- ``A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans''
PNAS 2024 <https://doi.org/10.1073/pnas.2313925121>
- ``Using large language models to categorize strategic situations and decipher motivations behind human behaviors'' PNAS 2025
<https://doi.org/10.1073/pnas.2512075122>
- ``Be.FM: Open Foundation Models for Human Behavior''
<https://doi.org/10.48550/arXiv.2505.23058>
- ``AI Behavioral Science'' <https://doi.org/10.2139/ssrn.5395006>