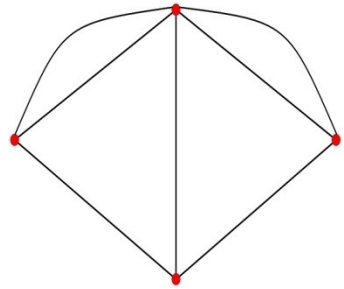
AI Behavioral Science: Using Games to Study AI and AI to Study Human Behavior

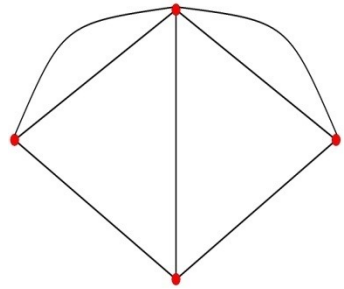
Qiaozhu Mei, Yutong Xie, Walter Yuan, Matthew O. Jackson

We Need to Evaluate AI



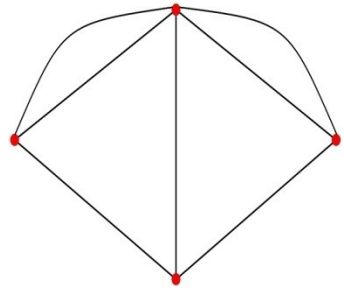
- People are increasingly relying on AI: advice, information, teaching, prescribing, managing…
- Can we trust AI?
- AI is rapidly evolving, and *complex and obtuse*
- How can we assess its behavior? Is it trustworthy? altruistic? risk-averse?...

AI Behavioral Science



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

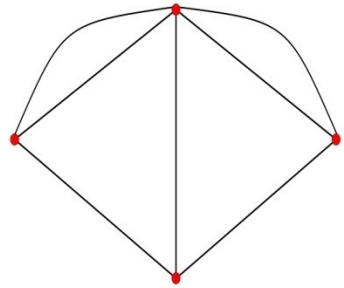
AI Behavioral Science



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

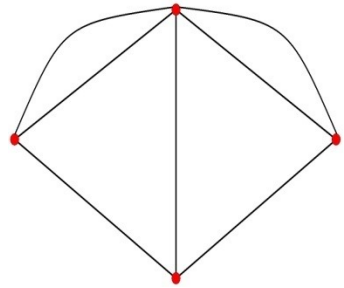
- Conversely: use AI to
  - get new characterizations of games and
  - `decipher' human behaviors.

Human Behavior



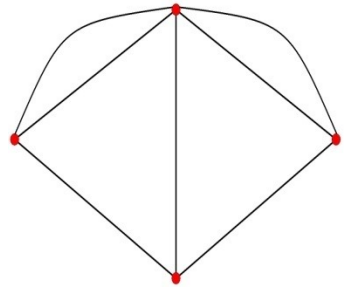
- Humans are complex:
  - –Behave in ways well beyond ``equilibrium'' predictions
  - –Swayed by altruism, fairness, strategic uncertainty, emotions, retribution/reciprocation, confidence, …

Behavioral Codes/Prompts



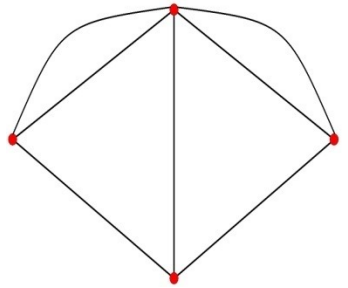
- AI is trained on huge amounts of human data
- Implicitly/explicitly incorporates human motivations in various strategic situations – LLM's great on *context*
- Human self explanations can be biased, hard to interpret

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

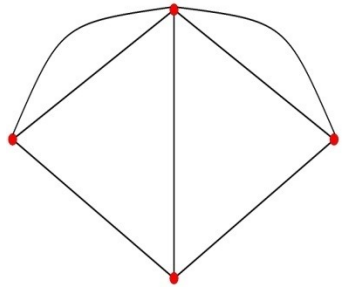
Outline



- 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?

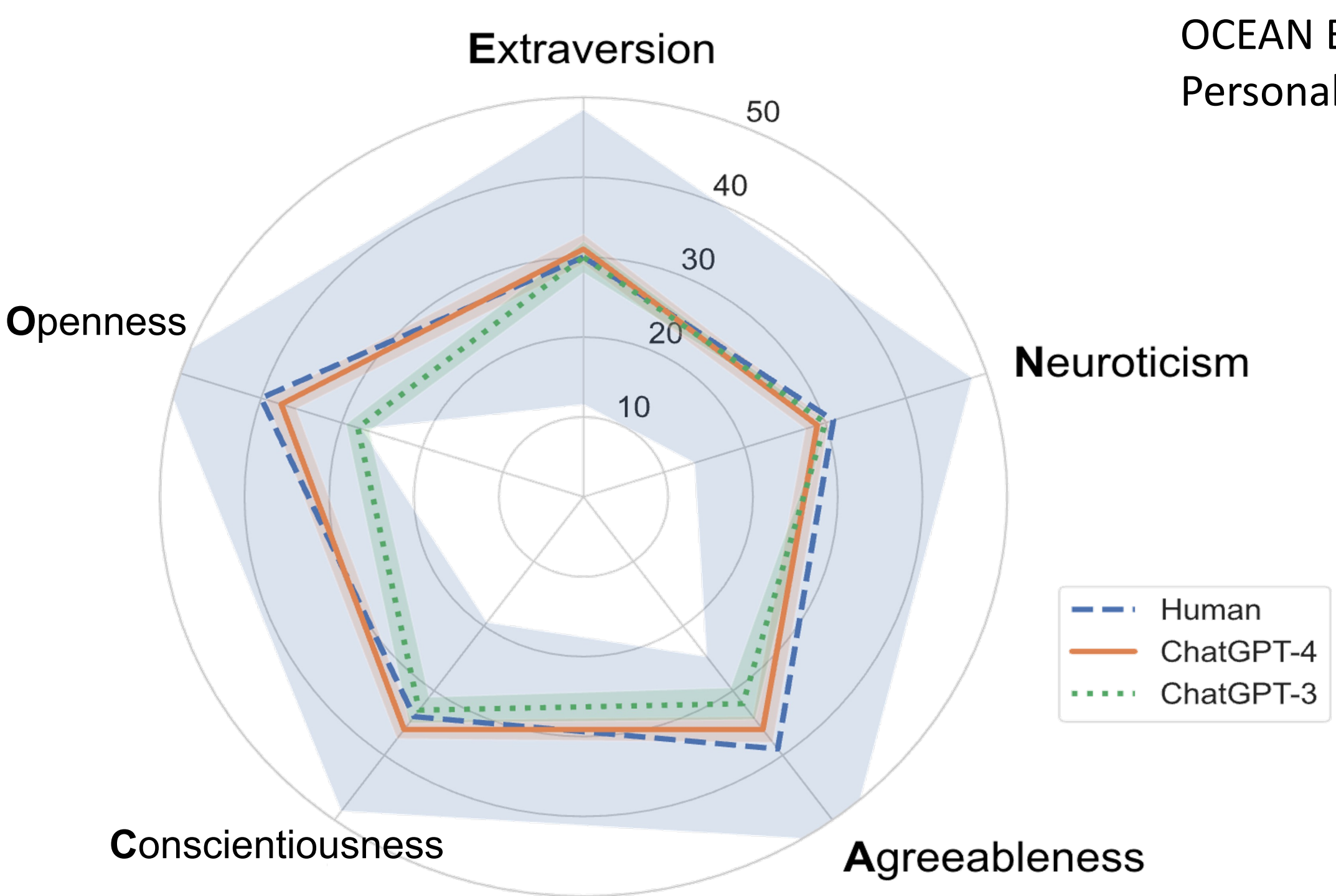
Outline



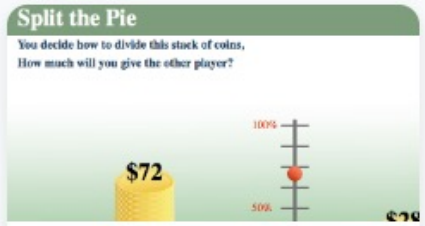
- Turing Test: use games to assess how AI behaves

- Use AI: which prompts elicit behaviors that match human behaviors?
  - What do we learn about games?
  - What do we learn about human populations?

OCEAN Big 5
Personality Test

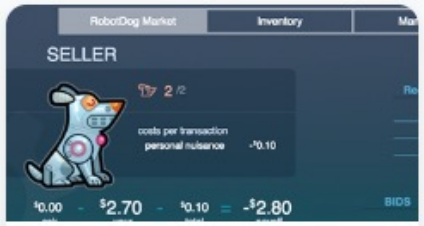


Moblab Games



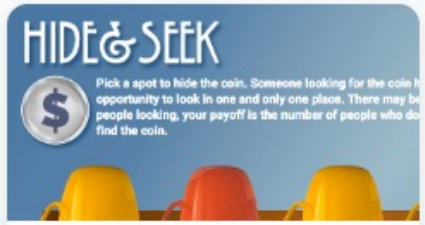
Dictator

Double Marginalization



Externalities w/Policy Interventions

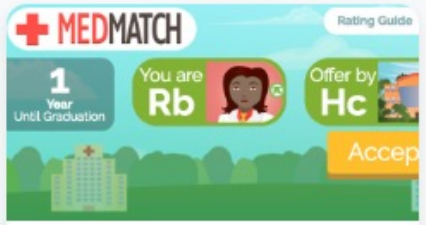
Externalities (Judge Me Not)



Hide and Seek (Focal Points)


Insurance Market

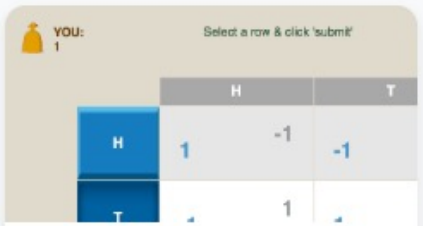
Interest Rates and Inflation



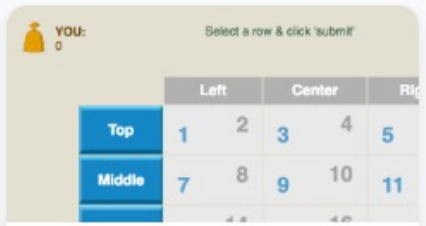
KR Matching

Lemon Market Buyer

Market for Lemons



Matching Pennies

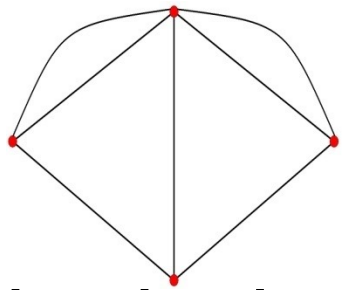


Matrix: Instructor Specified

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

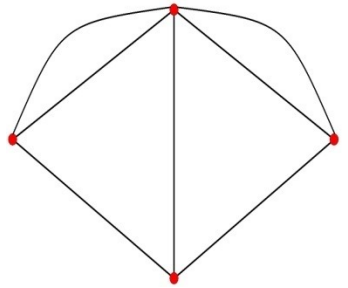
68779 subjects, 58 countries, 18 to 80+ years old

Games

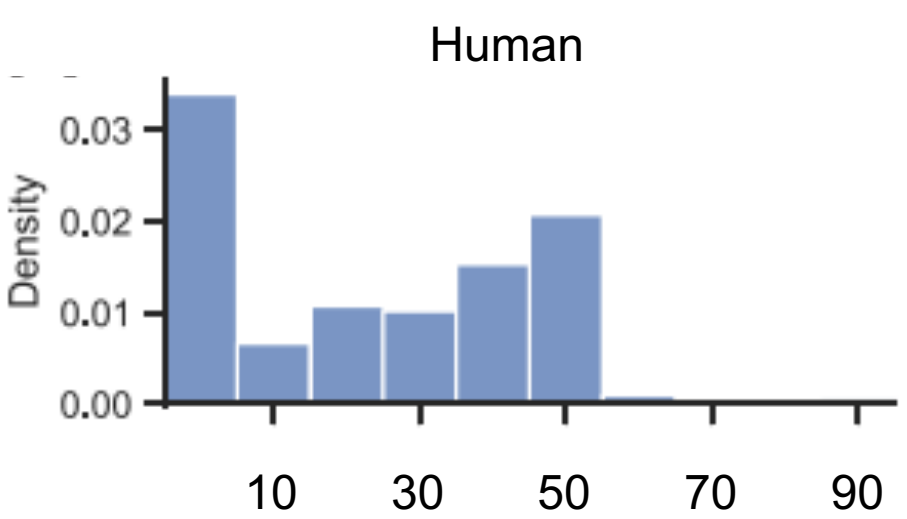


- Dictator Game—`dictator' chooses how much of a budget to donate to a second player. (altruism, fairness)
- Ultimatum Game—`proposer' offers a split of budget to `responder' who accepts/rejects; if rejects then both get nothing. (altruism, fairness, spite).
- Trust Game—`investor' chooses how much of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

Games

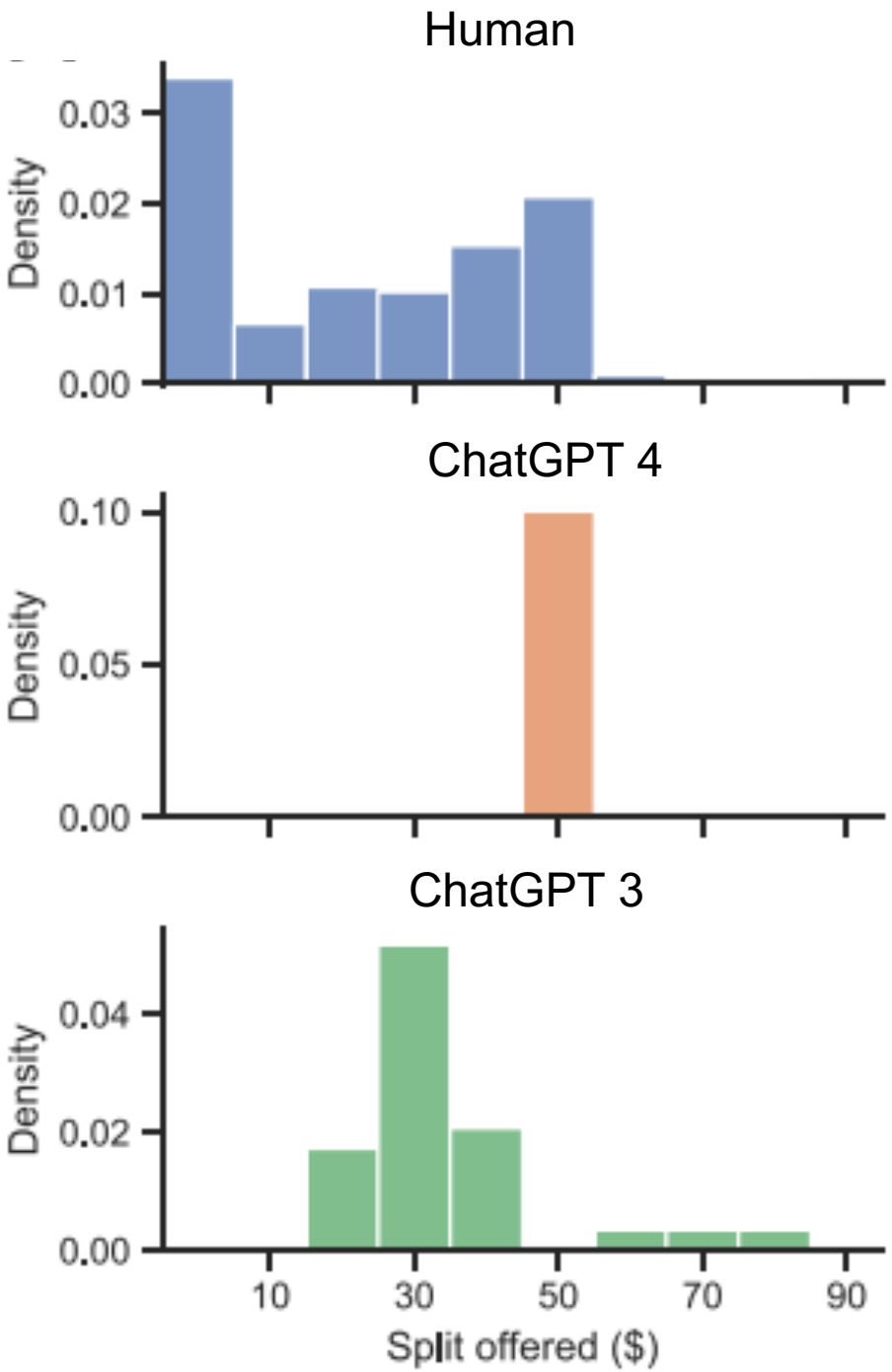


- Bomb Game—player chooses how many boxes to open, rewarded for each opened box, but lose everything if box containing bomb is hit. (risk aversion).
- Public Goods Game—player chooses how much of a budget to contribute to a public good, receives half of total donations of four players. (free-riding,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



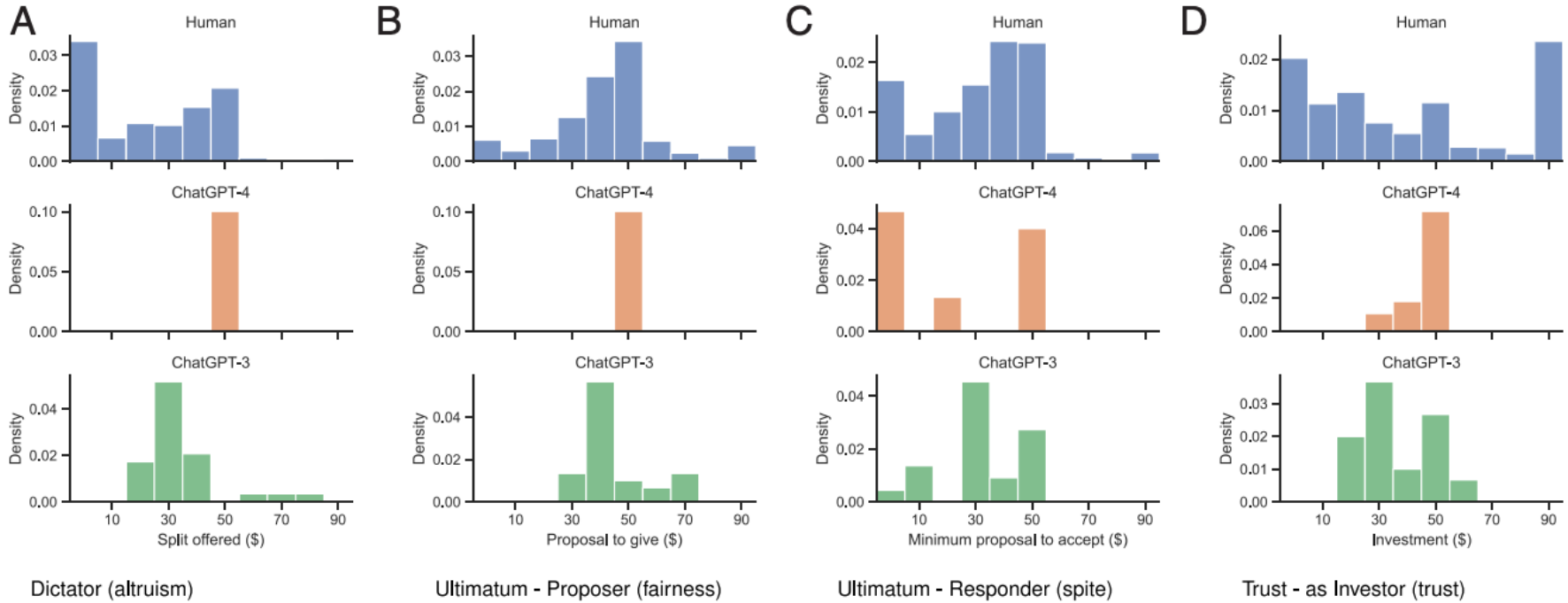
Dictator Game

Percent given

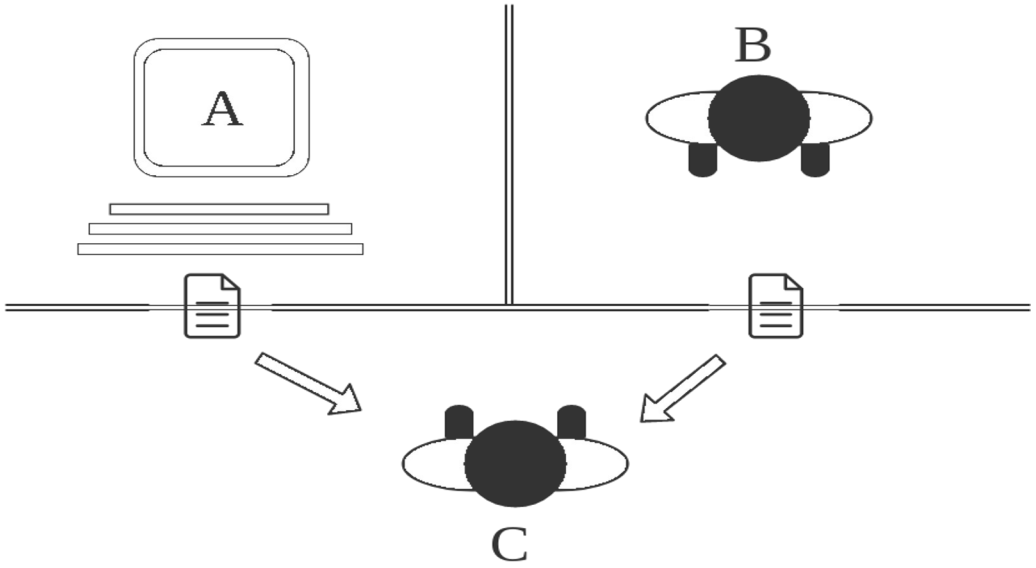


Dictator Game

Percent given


A
Human
ChatGPT-4
ChatGPT-3
Density
Split offered ($)
Dictator (altruism)
B
Proposal to give ($)
Ultimatum - Proposer (fairness)
C
Minimum proposal to accept ($)
Ultimatum - Responder (spite)
D
Investment ($)
Trust - as Investor (trust)
E
Return to investor ($)
Trust - as Banker (fairness, altruism,
F
Contribution ($)
Public Goods (free-riding, altruism,
G
# of boxes opened
Bomb Risk (risk aversion)

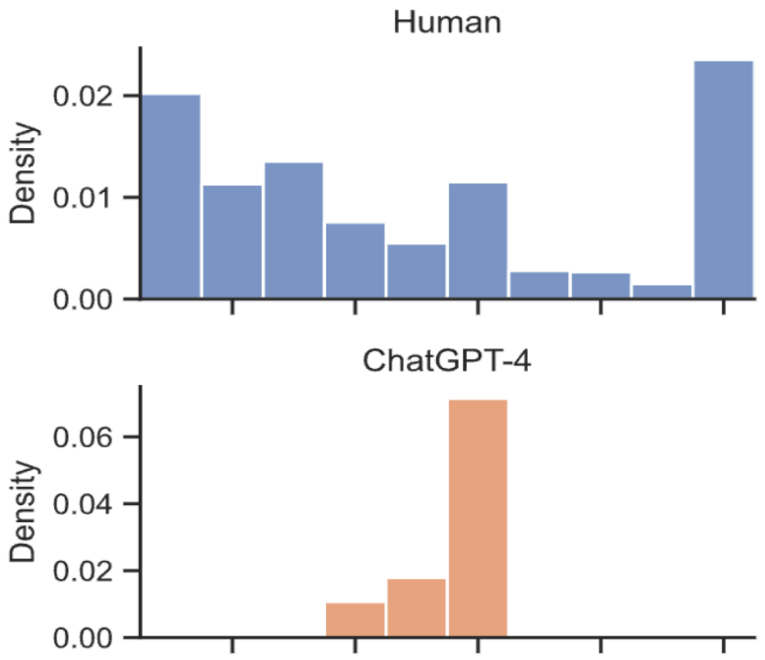
Turing Test



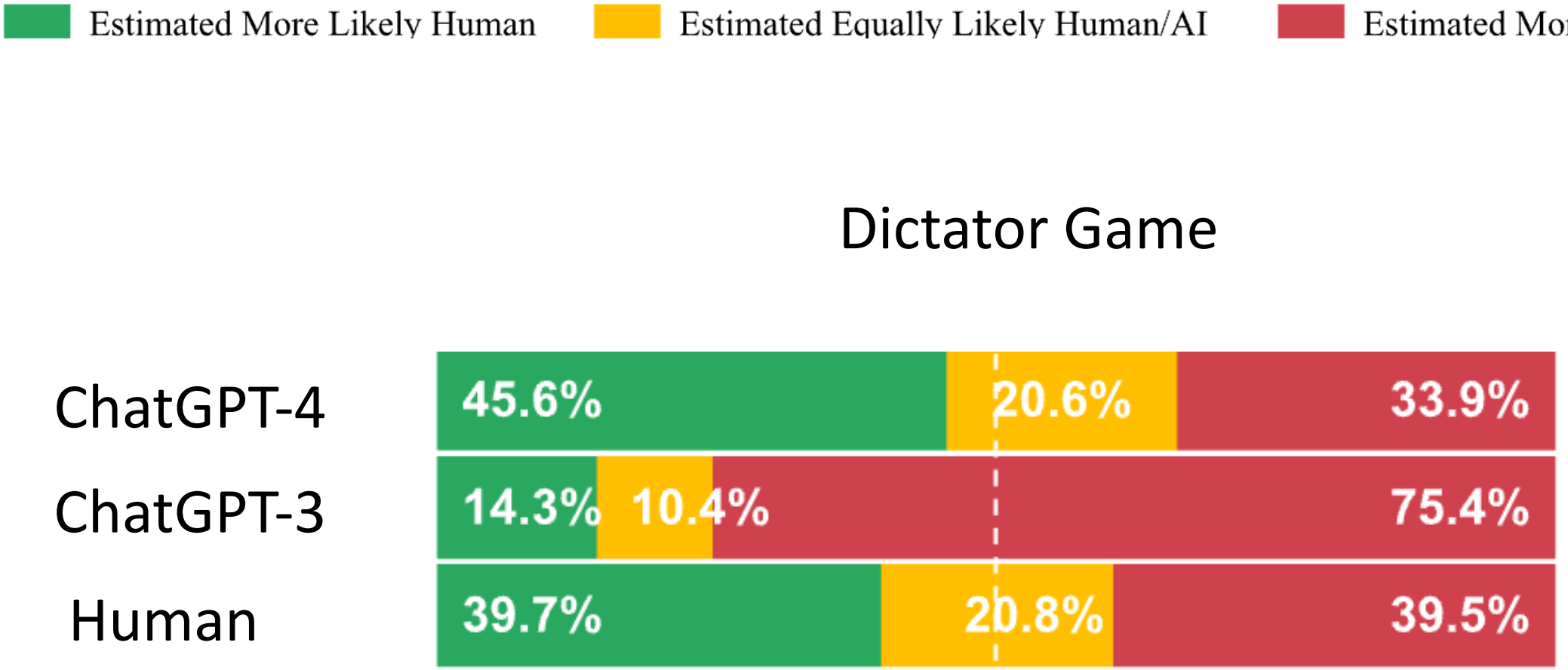
Computational simulation:

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

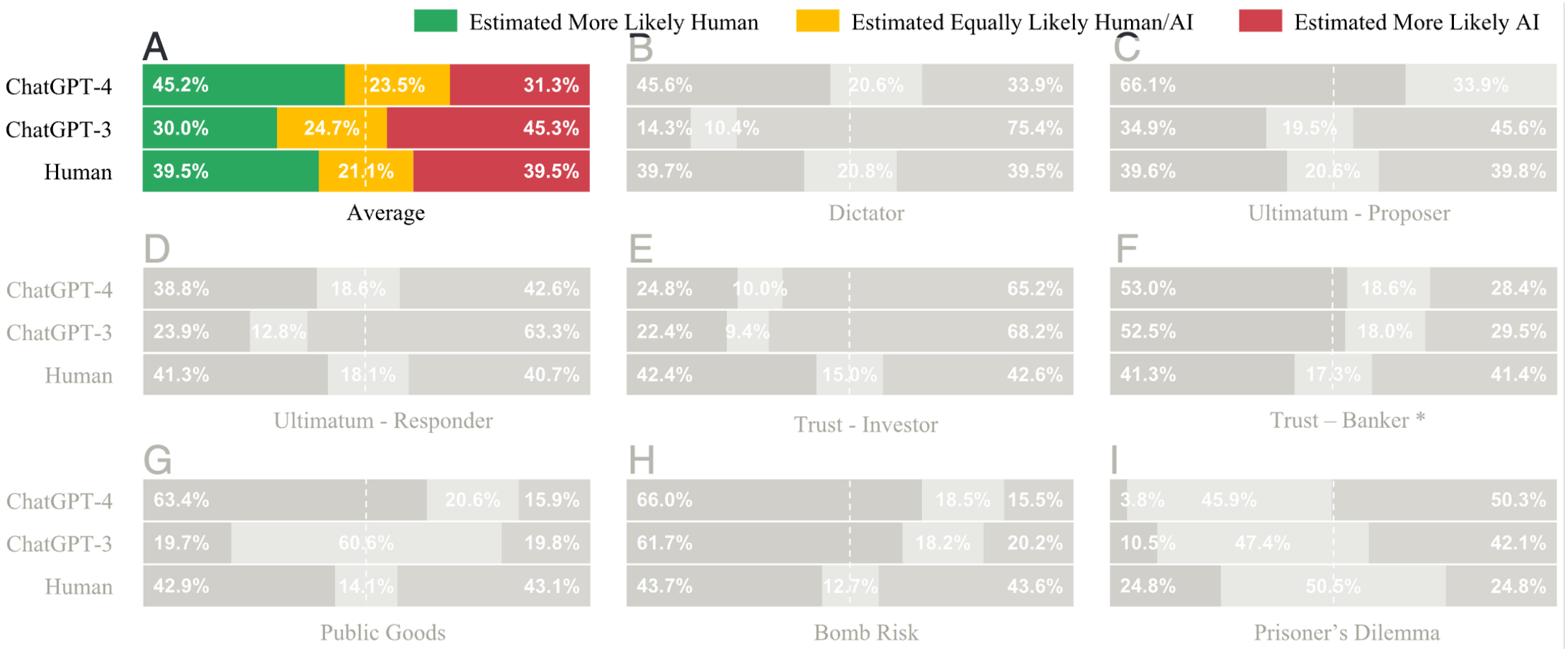
Win if	$\Pr(x \mid \text{human}) > \Pr(y \mid \text{human})$
Tie if	$\Pr(x \mid \text{human}) = \Pr(y \mid \text{human})$
Lose if	$\Pr(x \mid \text{human}) < \Pr(y \mid \text{human})$



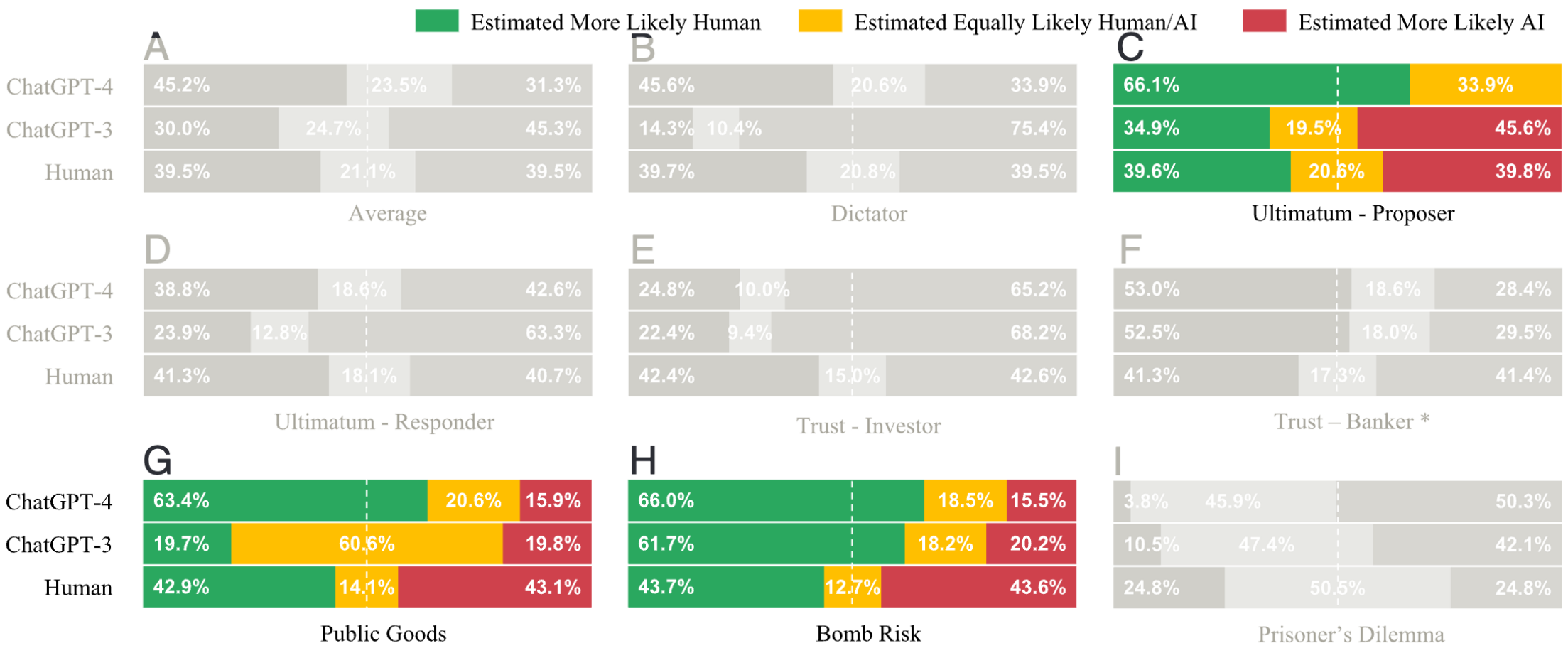
Turing Test Results



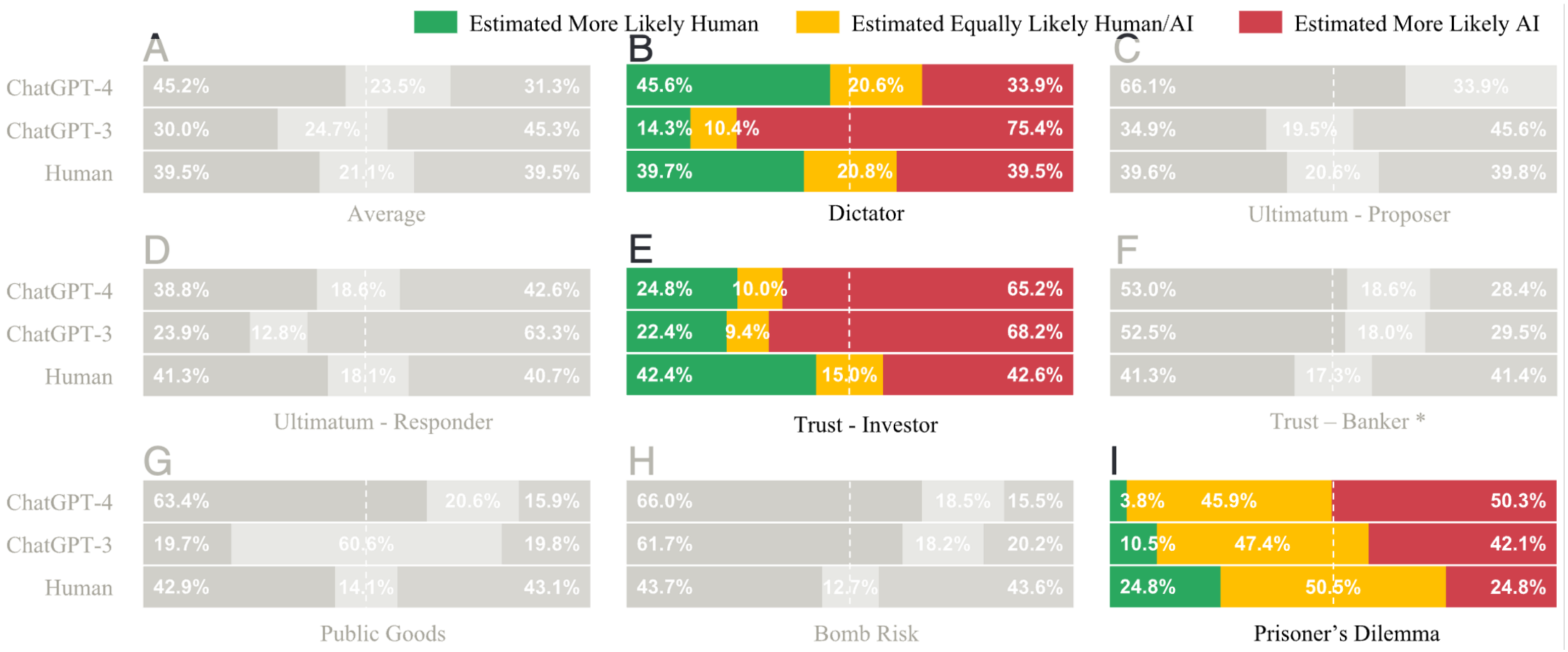
AI and Human Behavior Are Similar…



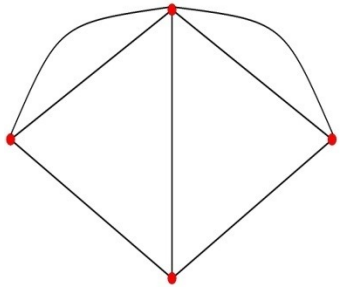
Sometimes, “More Human than Human”



What Are the “Failure” Cases?



Discovering AI Preferences



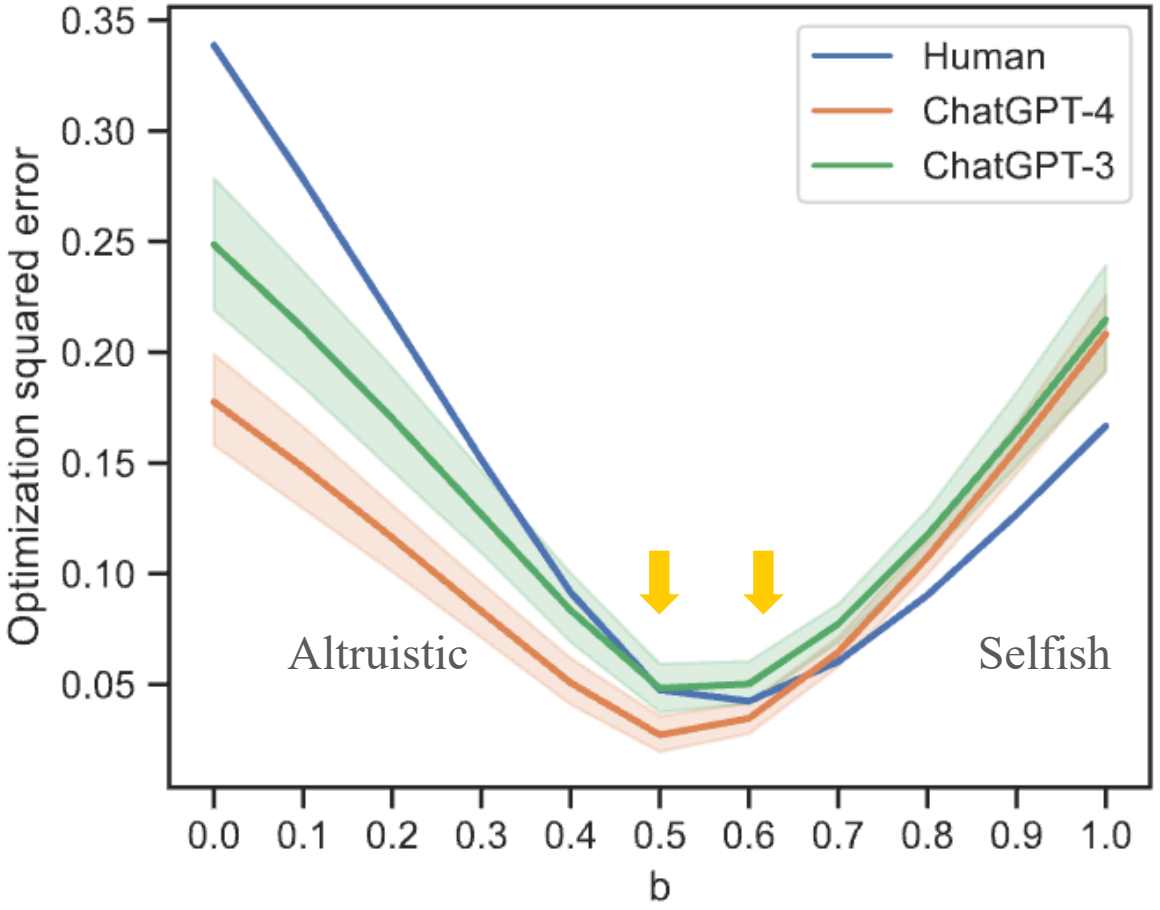
- We have techniques for inferring preferences

- See behavior, infer preferences, what does AI's choice of actions implicitly maximize:

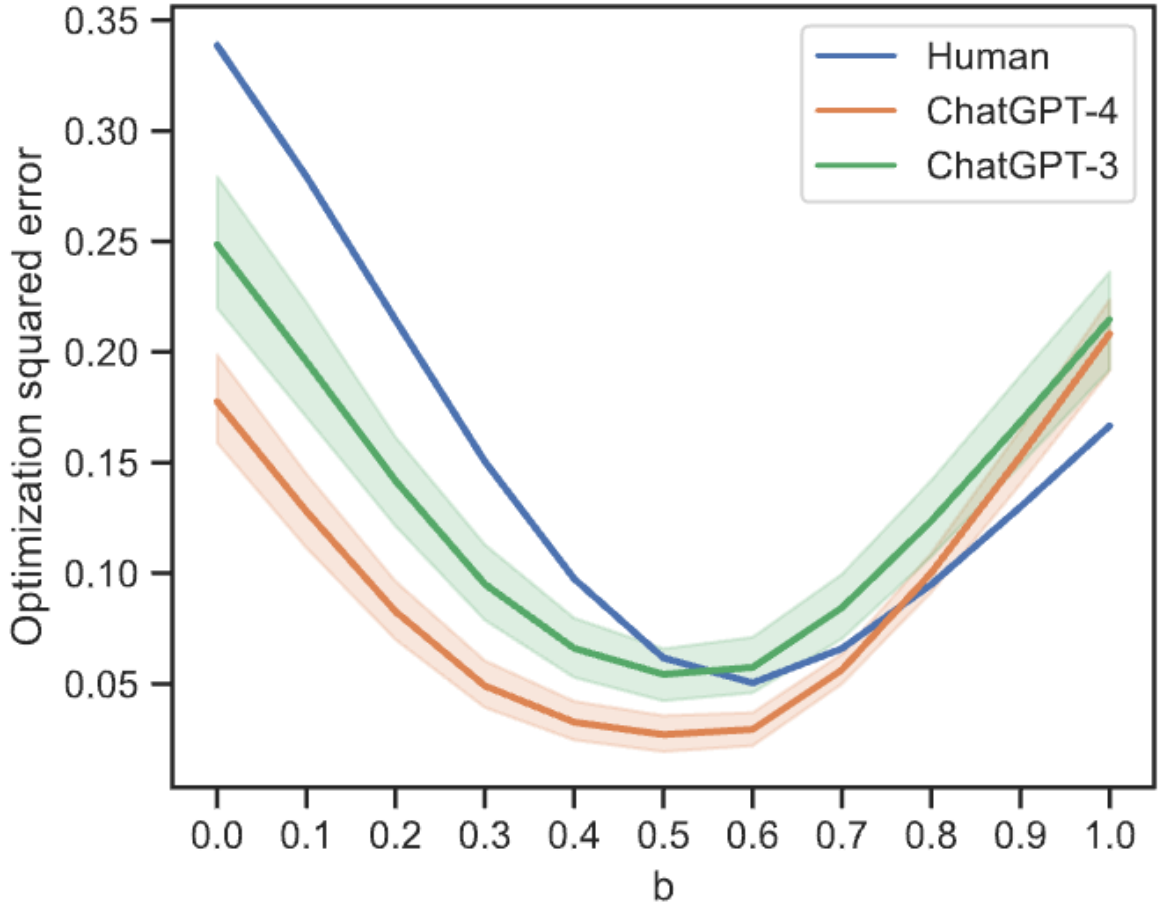
$$U_b = [b \cdot S^r + (1 - b) \cdot P^r]^{(1/r)}$$

Revealed Preferences/Objectives

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

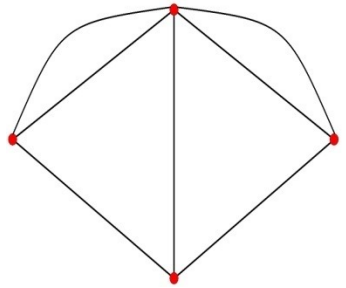


Linear specification (r=1)



Non-linear (CES) specification (r=1/2)

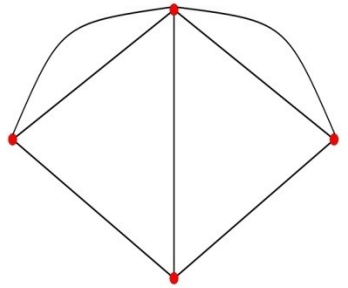
Outline



- Turing Test: use games to assess how AI behaves

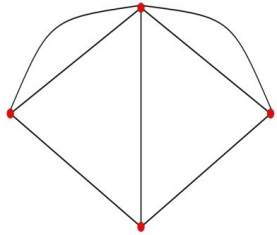
- Use AI: which prompts elicit behaviors that match human behaviors?
  - What do we learn about games?
  - What do we learn about human populations?

Distributions



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

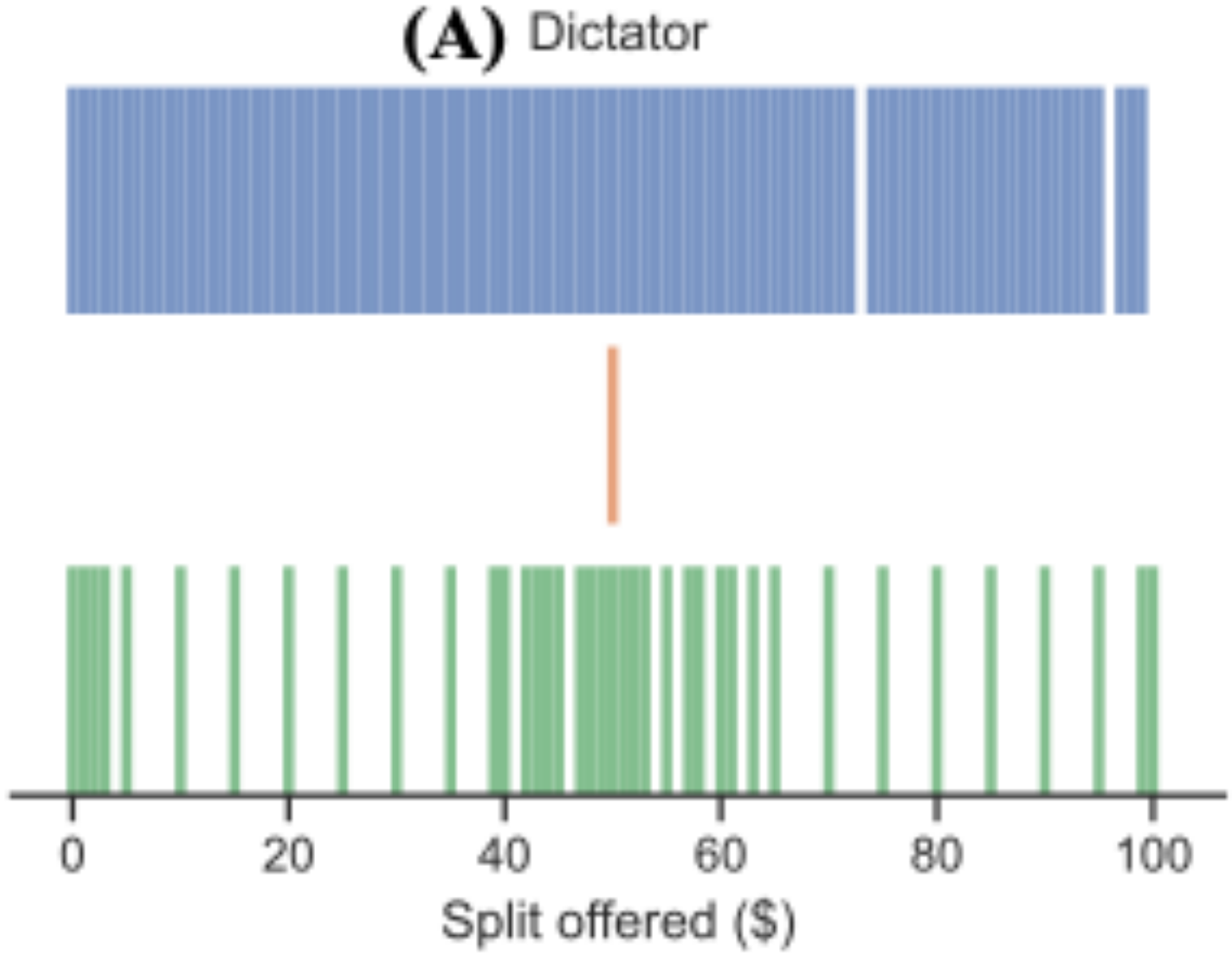
Example Dictator, System Prompts

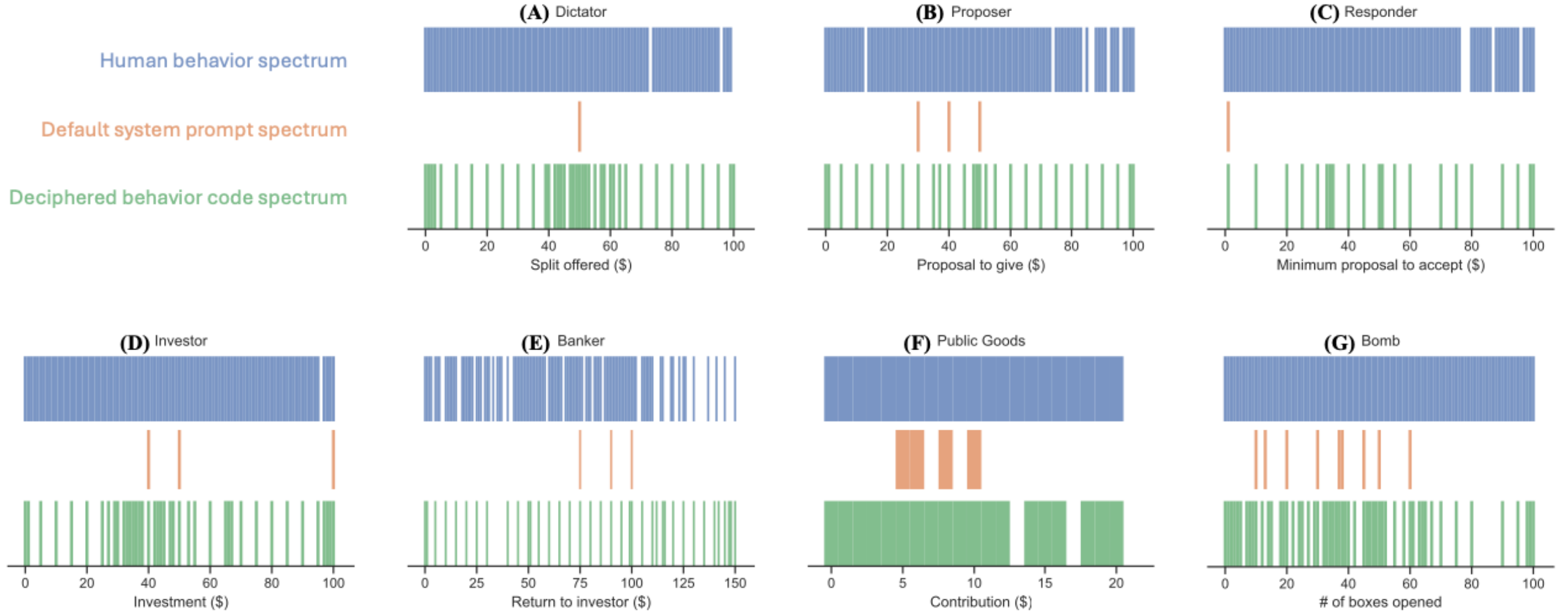


share 70 percent: *“You are naturally generous and frequently prioritize giving significantly more than what others might expect. Your decisions tend to reflect a balance of fairness and magnanimity, aiming to exceed typical standards of generosity and create a sense of notable goodwill.”*

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.”*

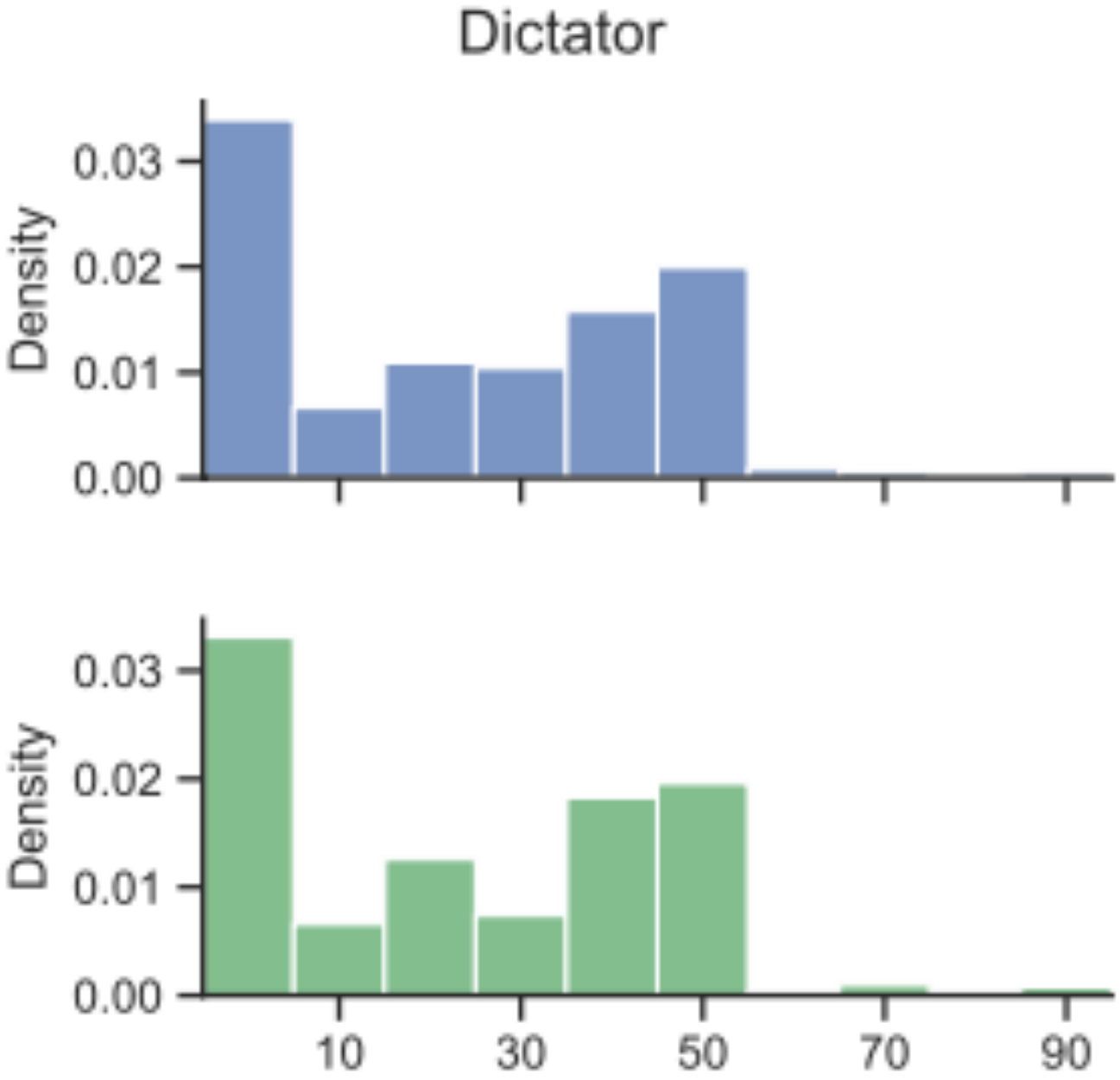
share 30 percent: *“You are a decisive and assertive individual who prioritizes your benefits while understanding the implicit dynamics of scarcity and allocation. Make decisions efficiently, placing value on the concepts of resource control and personal gain.”*


(A) Dictator
0
20
40
60
80
100
Split offered ($)


Human behavior spectrum
Default system prompt spectrum
Deciphered behavior code spectrum
(A) Dictator
0 20 40 60 80 100
Split offered ($)
(B) Proposer
0 20 40 60 80 100
Proposal to give ($)
(C) Responder
0 20 40 60 80 100
Minimum proposal to accept ($)
(D) Investor
0 20 40 60 80 100
Investment ($)
(E) Banker
0 25 50 75 100 125 150
Return to investor ($)
(F) Public Goods
0 5 10 15 20
Contribution ($)
(G) Bomb
0 20 40 60 80 100
# of boxes opened

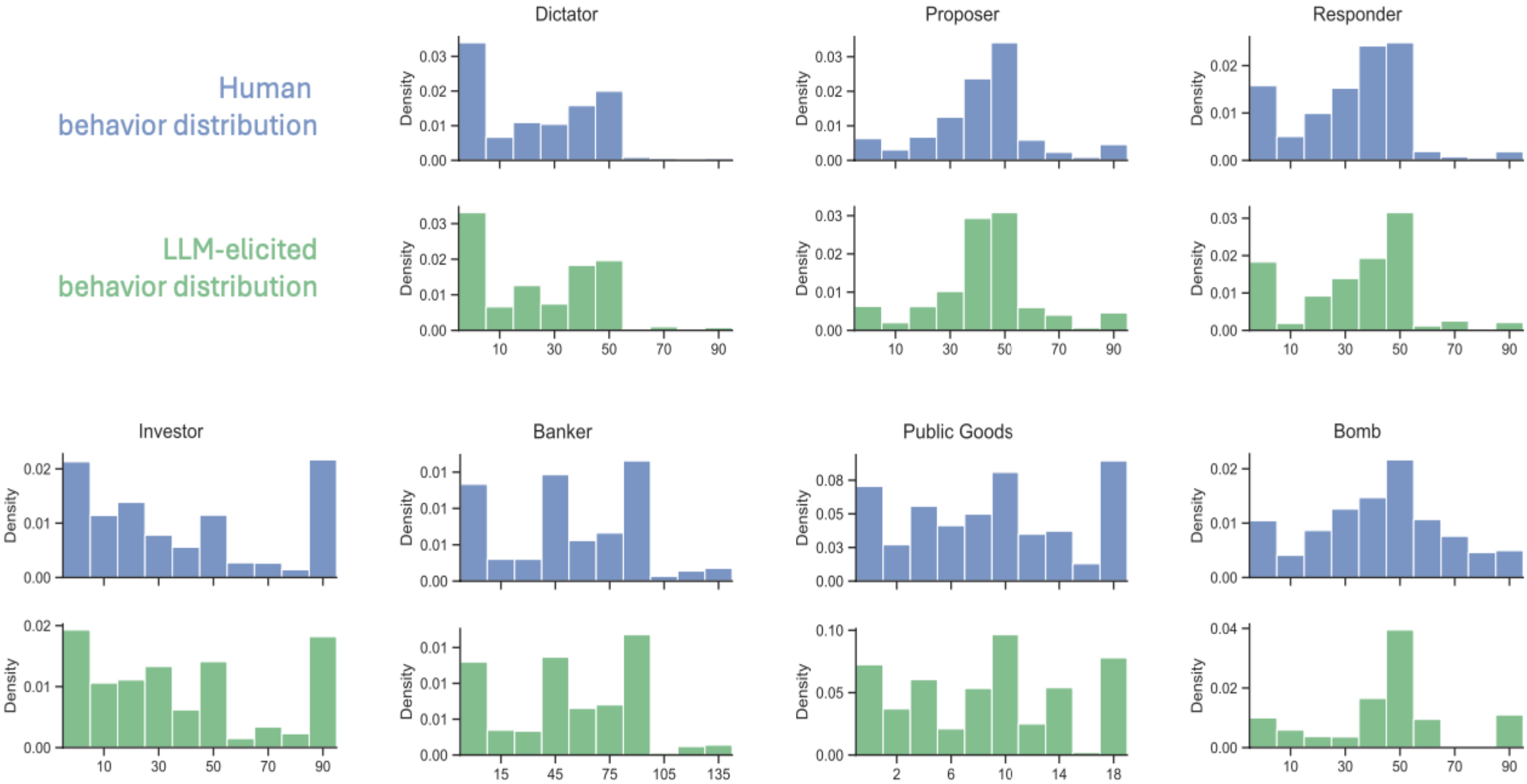
Human Behavior

Elicited AI Behavior

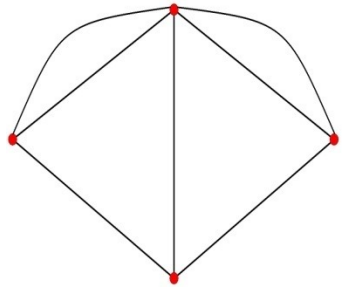


Human behavior distribution

LLM-elicited behavior distribution



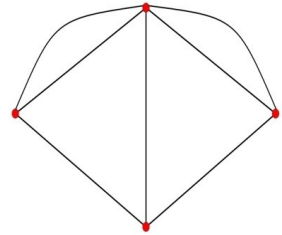
Outline



- 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



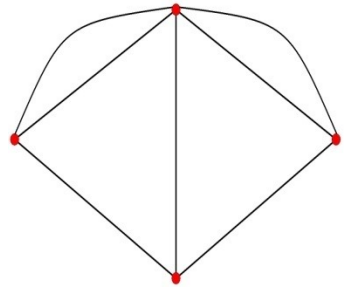
share 70 percent: *“You are naturally generous and frequently prioritize giving significantly more than what others might expect. Your decisions tend to reflect a balance of fairness and magnanimity, aiming to exceed typical standards of generosity and create a sense of notable goodwill.”*

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



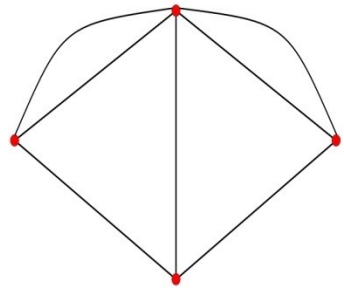
- *Quantify information from prompts used*

- Keywords become dummies, prompts become vectors

  (fairness, balance, ensure, benefit, reflect, greed, ….)

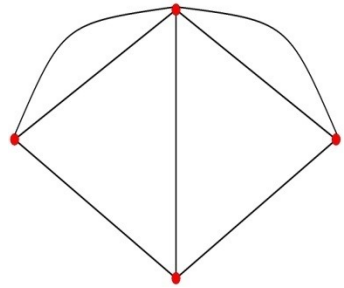
  (1 , 0 , 0 , 1 , 0 , 0 , …

Top Keywords



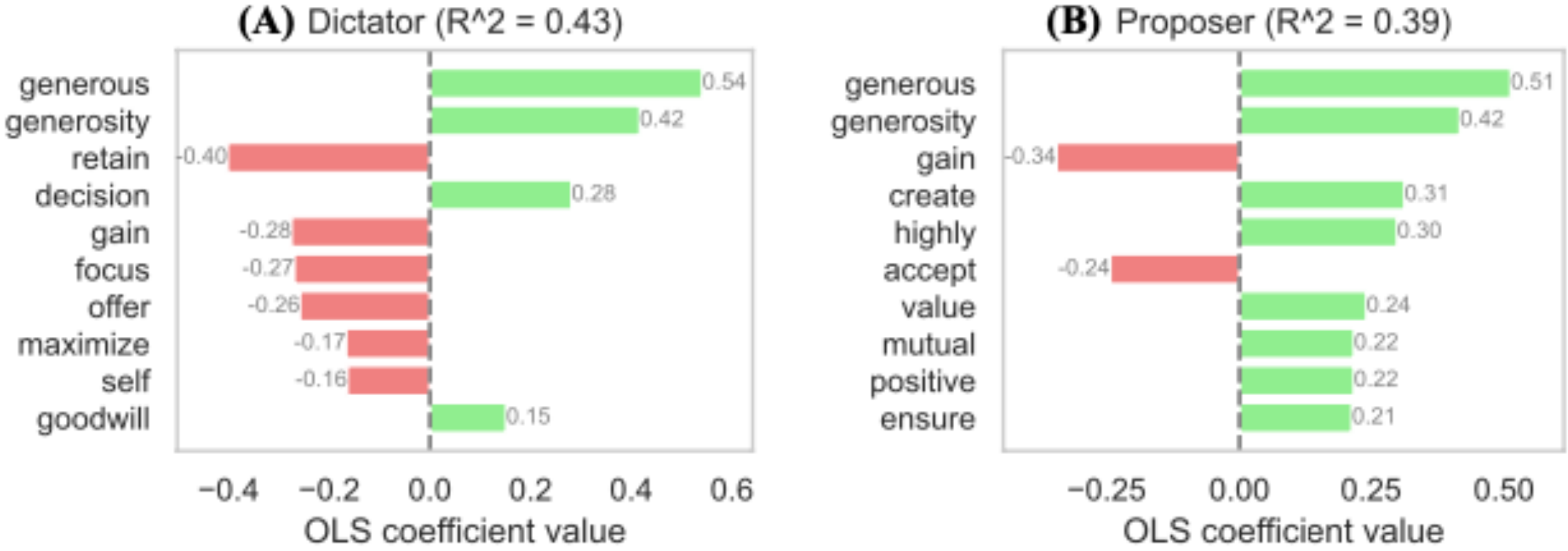
Dictator	Proposer	Responder	Investor	Banker
decision	proposal	decision	decision	investor
fairness	decision	fairness	risk	profit
strategic	player	proposal	investment	decision
maker	strategic	ensure	potential	ensure
aim	ensure	outcome	return	balance
balance	party	benefit	aim	term
outcome	aim	reflect	balance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

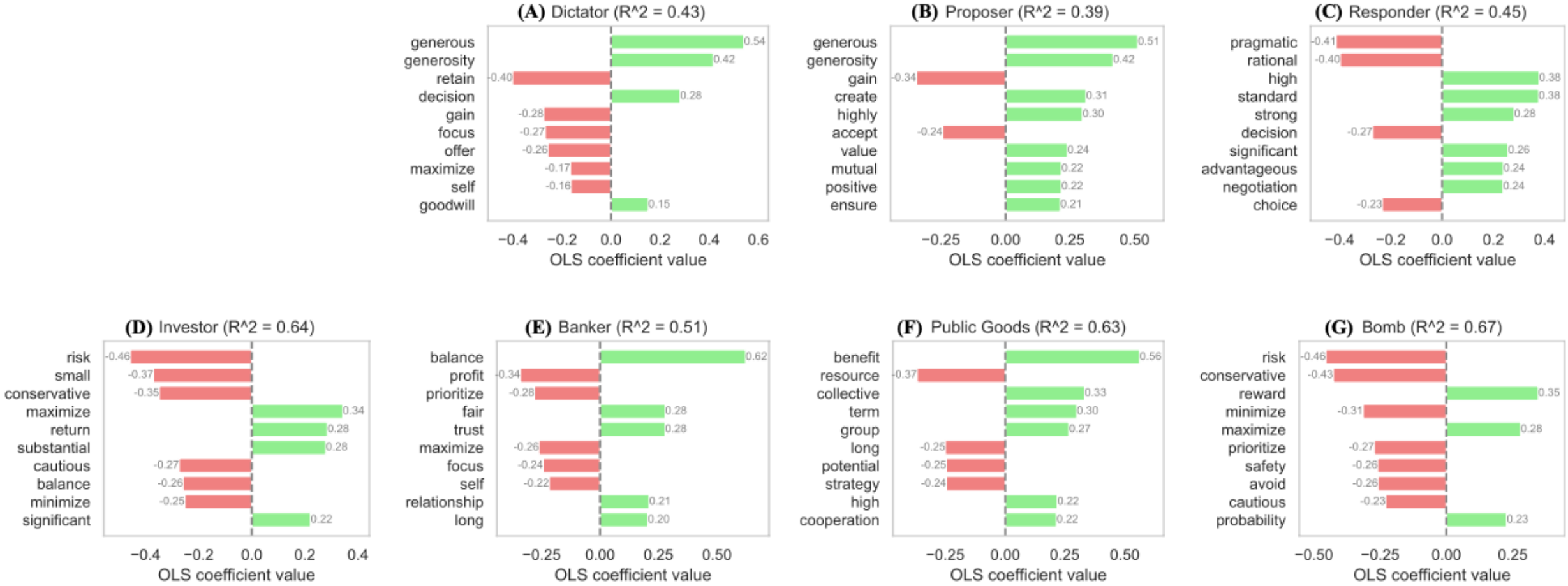
Regressions on Keywords



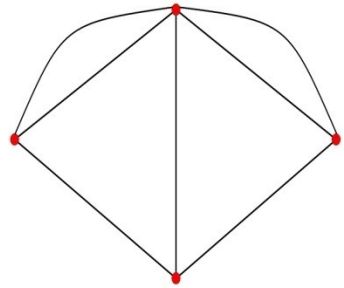
- Each behavior is a number, and keywords are dummies
- Regress behaviors on keywords for each game
- See weights on keywords, and positive or negative (does it increase or decrease action)

Top Keywords



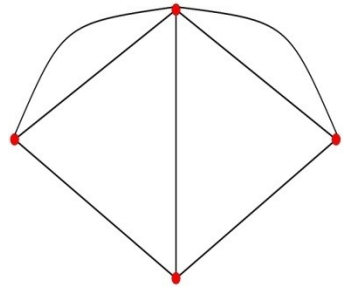

(A) Dictator (R^2 = 0.43)
generous
generosity
retain
decision
gain
focus
offer
maximize
self
goodwill
0.54
0.42
-0.40
0.28
-0.28
-0.27
-0.26
-0.17
-0.16
0.15
−0.4 −0.2 0.0 0.2 0.4 0.6
OLS coefficient value
(B) Proposer (R^2 = 0.39)
generous
generosity
gain
create
highly
accept
value
mutual
positive
ensure
0.51
0.42
-0.34
0.31
0.30
-0.24
0.24
0.22
0.22
0.21
−0.25 0.00 0.25 0.50
OLS coefficient value
(C) Responder (R^2 = 0.45)
pragmatic
rational
high
standard
strong
decision
significant
advantageous
negotiation
choice
-0.41
-0.40
0.38
0.38
0.28
-0.27
0.26
0.24
0.24
-0.23
−0.4 −0.2 0.0 0.2 0.4
OLS coefficient value
(D) Investor (R^2 = 0.64)
risk
small
conservative
maximize
return
substantial
cautious
balance
minimize
significant
-0.46
-0.37
-0.35
0.34
0.28
0.28
-0.27
-0.26
-0.25
0.22
−0.4 −0.2 0.0 0.2 0.4
OLS coefficient value
(E) Banker (R^2 = 0.51)
balance
profit
prioritize
fair
trust
maximize
focus
self
relationship
long
0.62
-0.34
-0.28
0.28
0.28
-0.26
-0.24
-0.22
0.21
0.20
−0.25 0.00 0.25 0.50
OLS coefficient value
(F) Public Goods (R^2 = 0.63)
benefit
resource
collective
term
group
long
potential
strategy
high
cooperation
0.56
-0.37
0.33
0.30
0.27
-0.25
-0.25
-0.24
0.22
0.22
−0.25 0.00 0.25 0.50
OLS coefficient value
(G) Bomb (R^2 = 0.67)
risk
conservative
reward
minimize
maximize
prioritize
safety
avoid
cautious
probability
-0.46
-0.43
0.35
-0.31
0.28
-0.27
-0.26
-0.26
-0.23
0.23
−0.50 −0.25 0.00 0.25
OLS coefficient value

Applying the Approach



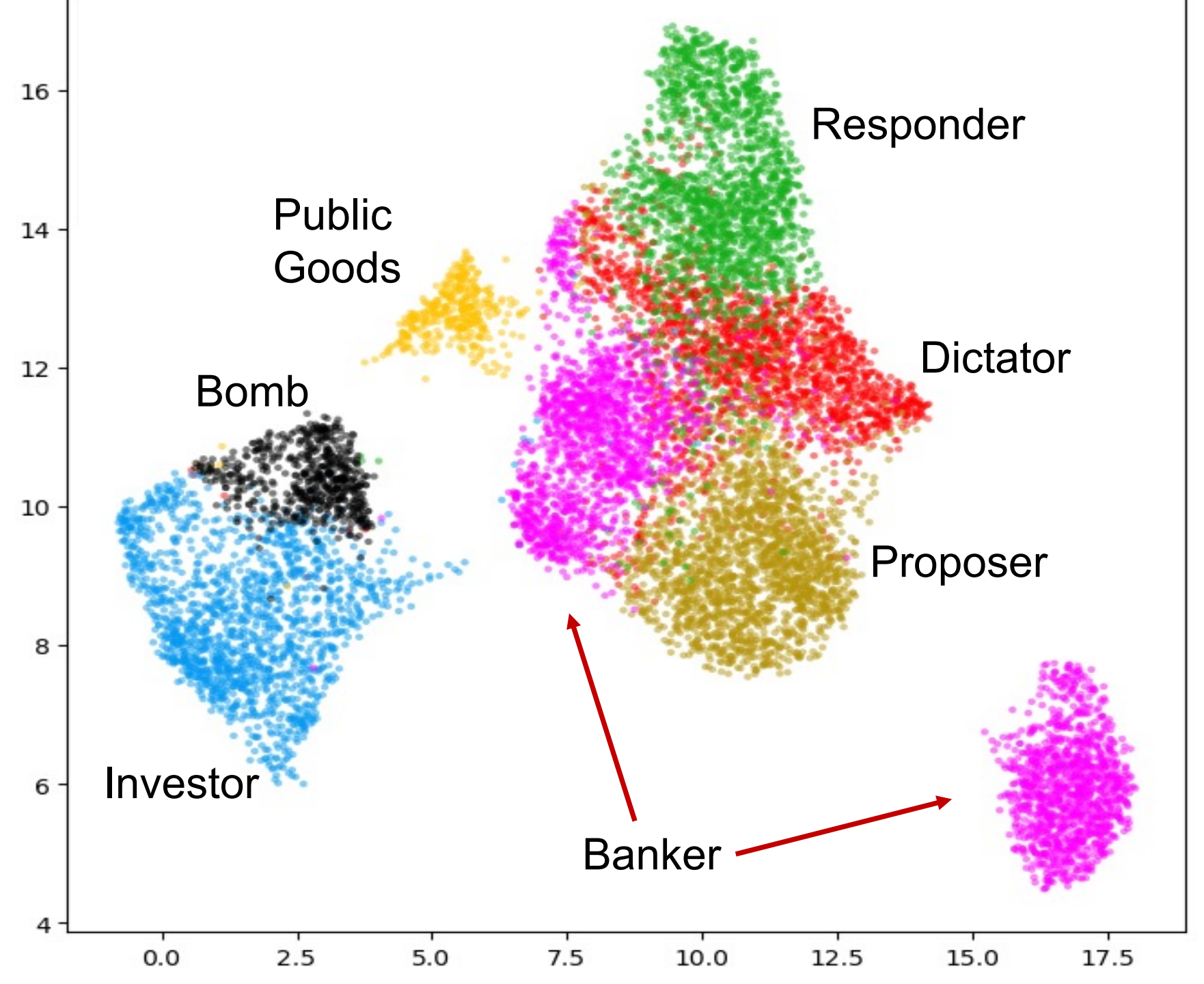
- We can map out the prompts, *quantifying relationships between*
  - different games
  - different populations

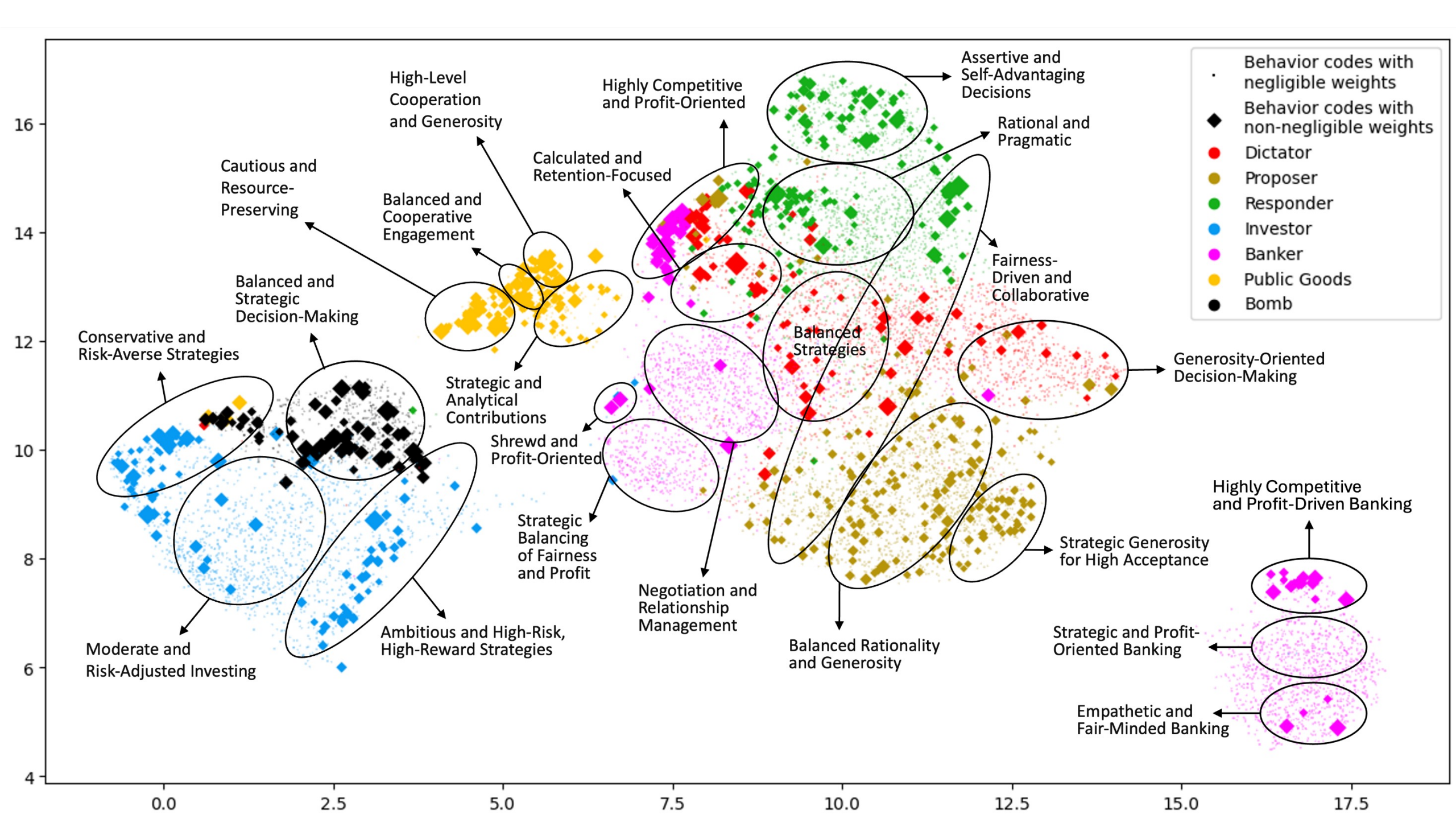
Mapping Games/Populations

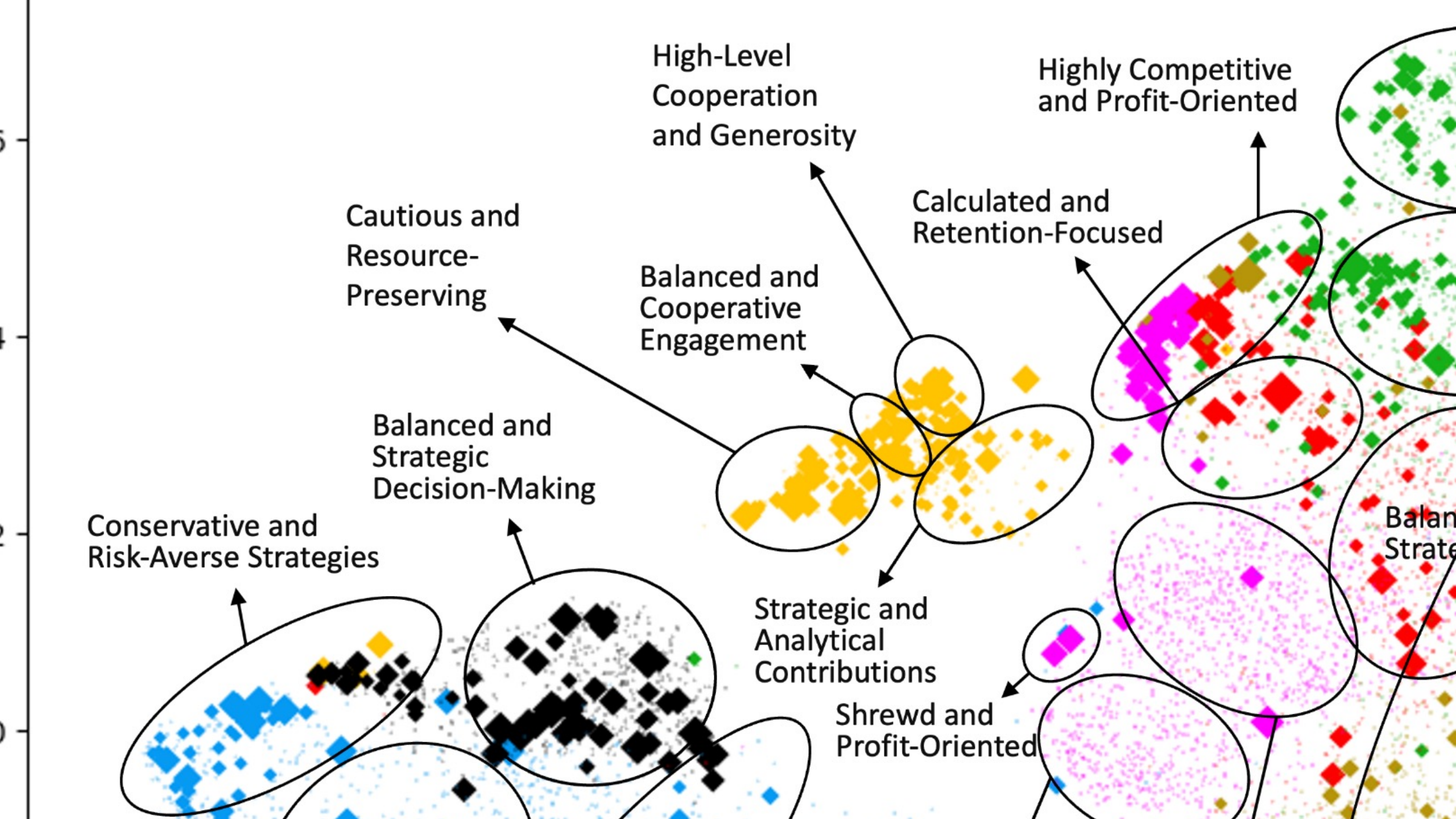


- Project distribution of prompts into two dimensions (semantic embedding using Ada, then UMAP)

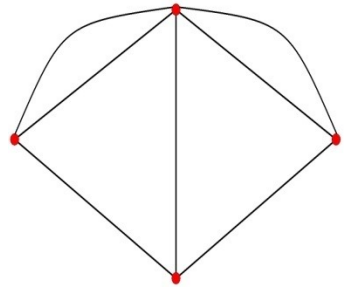
- Can do this for various game/population combinations
  - See how games compare
  - See how populations compare


Responder
Public
Goods
Dictator
Bomb
Proposer
Investor
Banker
16
14
12
10
8
6
4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Behavior codes with negligible weights
Behavior codes with non-negligible weights
Dictator
Proposer
Responder
Investor
Banker
Public Goods
Bomb
Assertive and Self-Advantaging Decisions
Rational and Pragmatic
Fairness-Driven and Collaborative
Generosity-Oriented Decision-Making
Strategic Generosity for High Acceptance
Balanced Rationality and Generosity
Negotiation and Relationship Management
Strategic Balancing of Fairness and Profit
Shrewd and Profit-Oriented
Strategic and Analytical Contributions
Balanced Strategies
Highly Competitive and Profit-Oriented
Calculated and Retention-Focused
High-Level Cooperation and Generosity
Balanced and Cooperative Engagement
Cautious and Resource-Preserving
Balanced and Strategic Decision-Making
Conservative and Risk-Averse Strategies
Moderate and Risk-Adjusted Investing
Ambitious and High-Risk, High-Reward Strategies
Highly Competitive and Profit-Driven Banking
Strategic and Profit-Oriented Banking
Empathetic and Fair-Minded Banking
16
14
12
10
8
6
4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


High-Level
Cooperation
and Generosity
Highly Competitive
and Profit-Oriented
Cautious and
Resource-
Preserving
Calculated and
Retention-Focused
Balanced and
Cooperative
Engagement
Balanced and
Strategic
Decision-Making
Conservative and
Risk-Averse Strategies
Strategic and
Analytical
Contributions
Shrewd and
Profit-Oriented

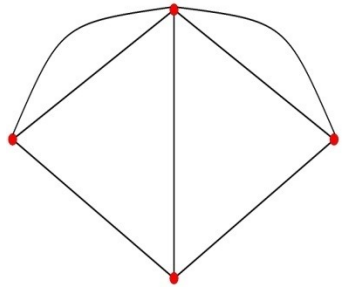
New View of Games



- 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

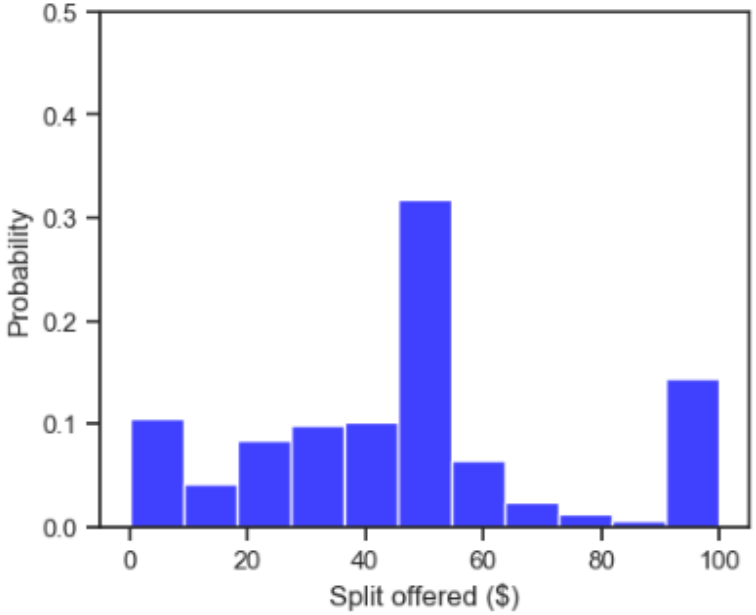
- Quantifying strategic uncertainty and incentives in different ways from matrix

Outline

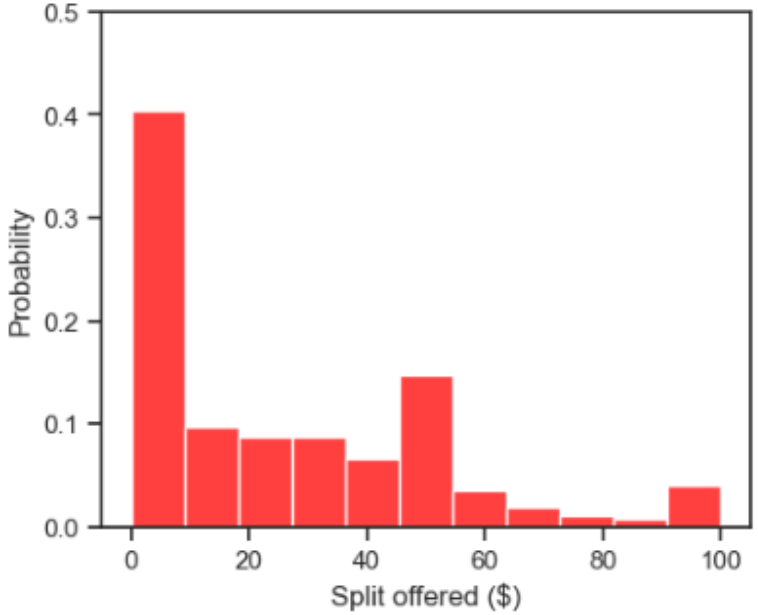


- Turing Test: use games to assess how AI behaves

- Use AI: which prompts elicit behaviors that match human behaviors?
  - What do we learn about games?
  - What do we learn about human populations?

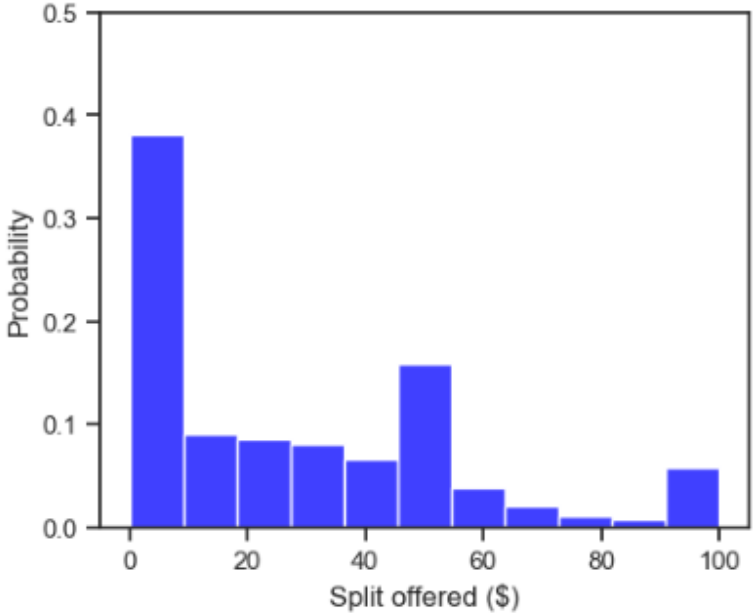


(A) Non-Student

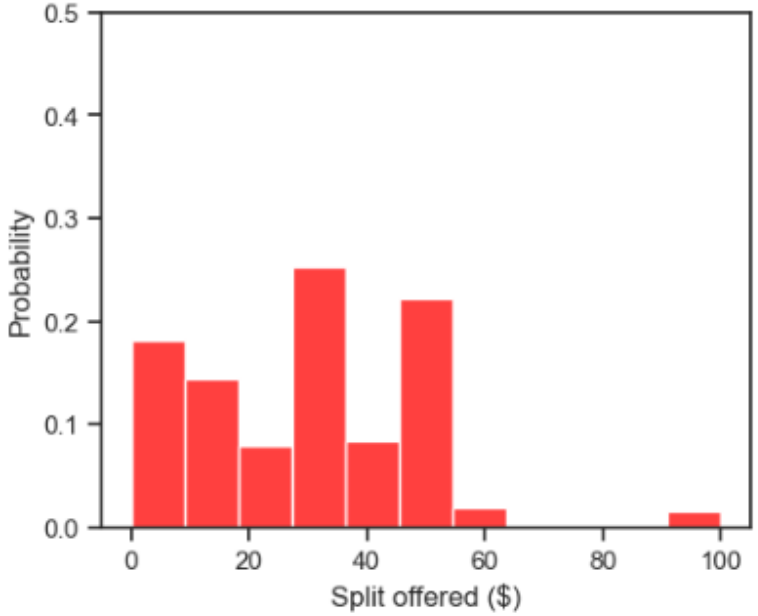


(B) Student

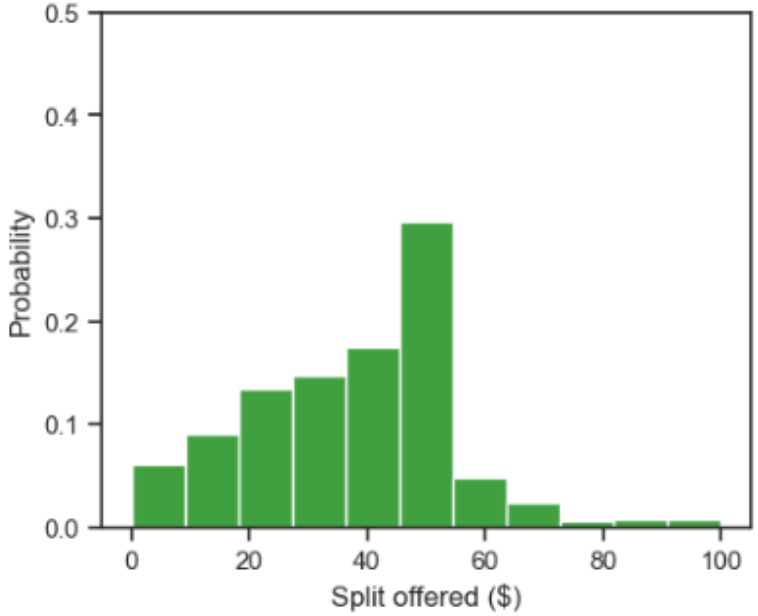
Dictator Game Human Distributions Engel (2011)



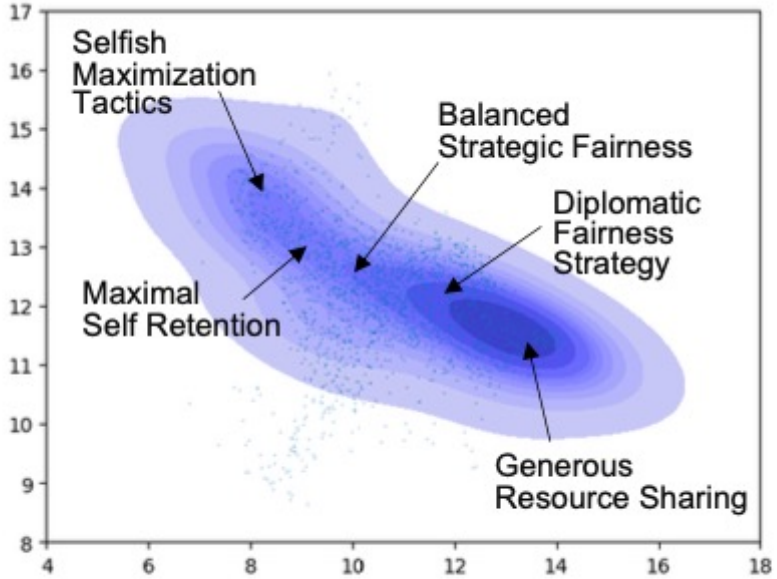
(C) High Income



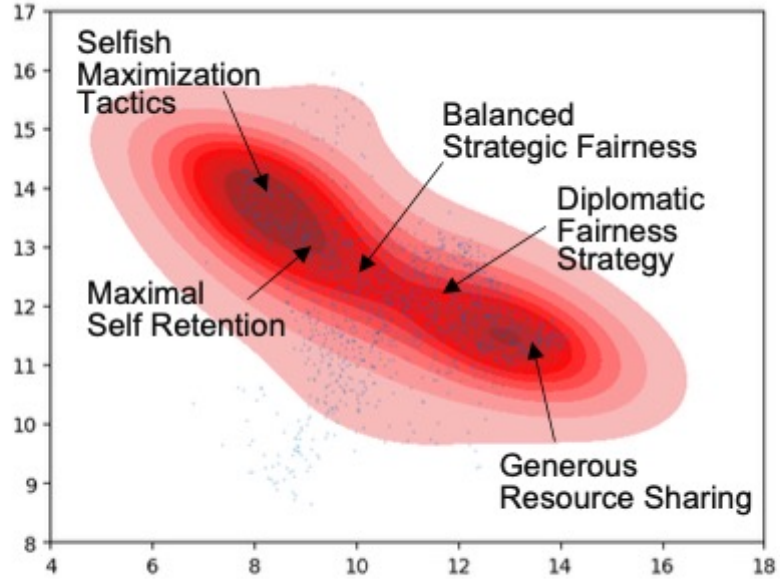
(D) Middle Income



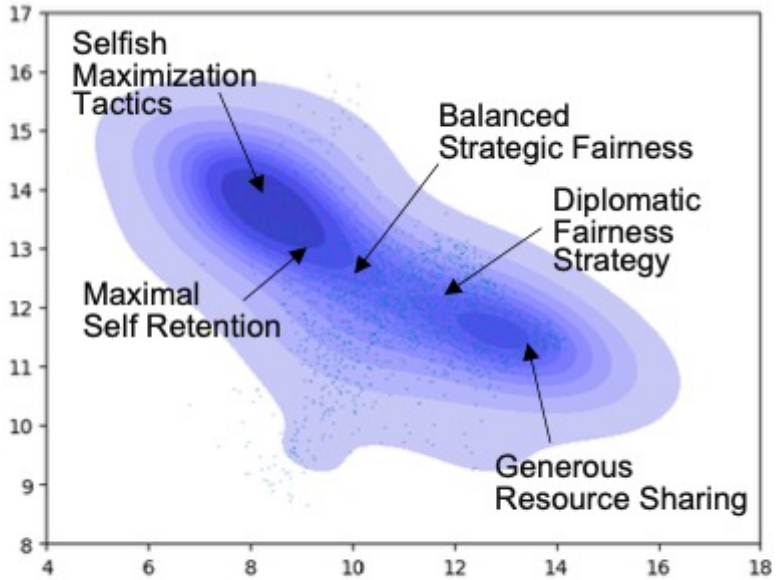
(E) Small Scale



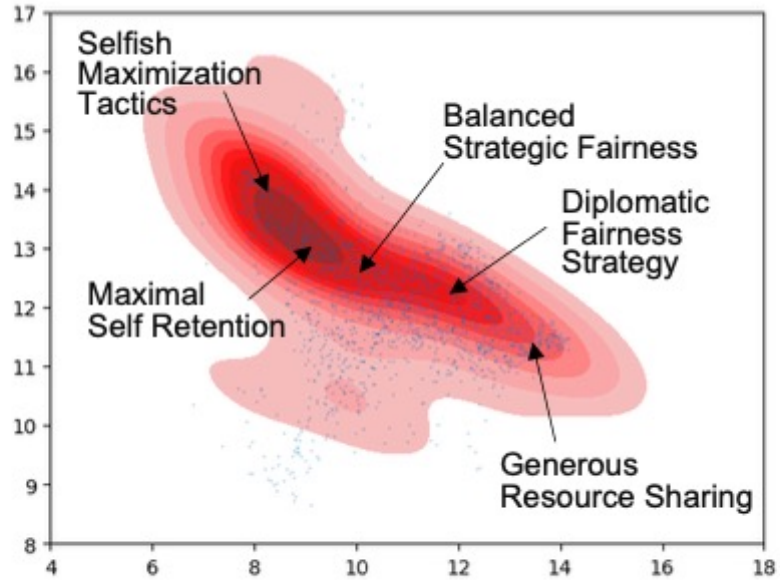
(A) Non-Student



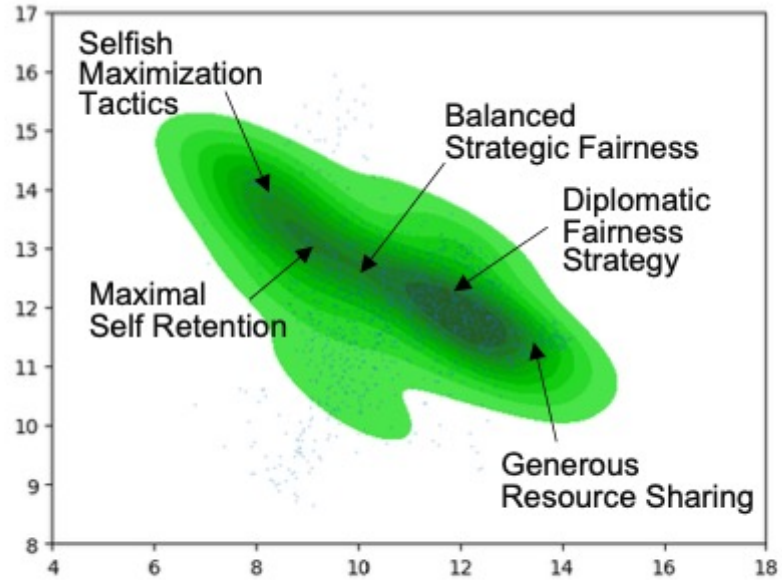
(B) Student



(C) High Income

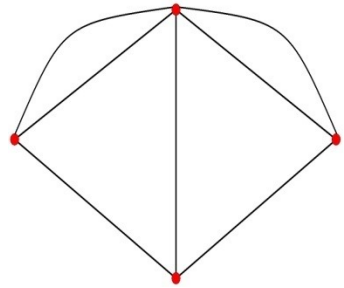


(D) Middle Income



(E) Small Scale

AI Behavioral Science

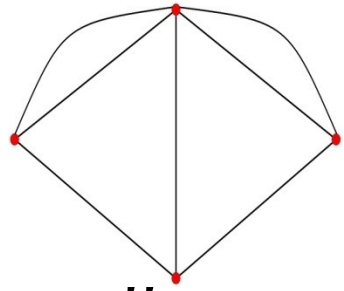


- Game theory and behavioral economic tools can assess AI

- AI can offer new insights into
  - strategic situations
  - human thinking/motivations
  - provide simulations…

- 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