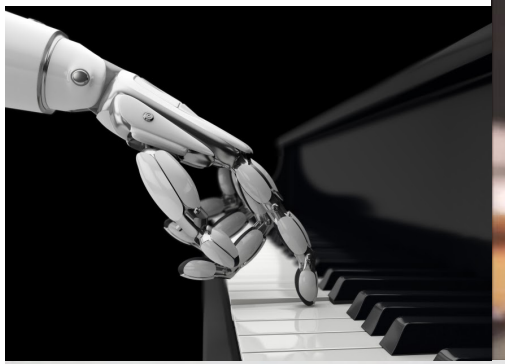
A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans

Qiaozhu Mei[1], Yutong Xie[1], Walter Yuan[2], Matthew O. Jackson[3]

[1] School of Information, University of Michigan
[2] MobLab
[3] Department of Economics, Stanford University

PNAS, 2024

Will You Trust or Not?



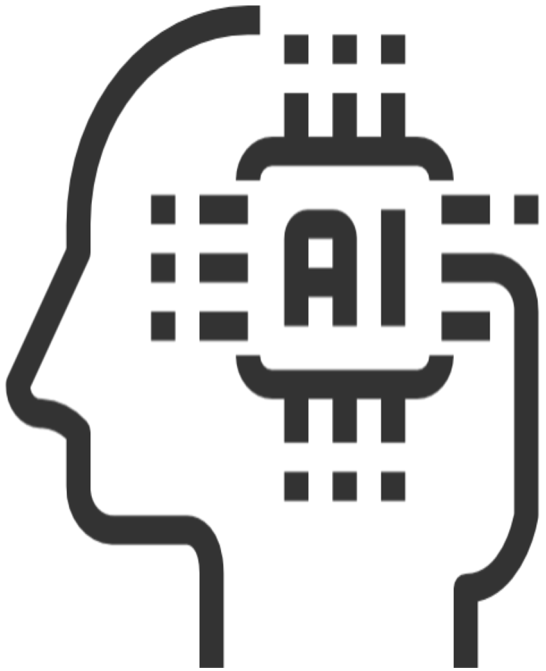
Will You Trust or Not?

Performance?

Data?

Model?

Training?



Knowledge?

Responses/thoughts?

Behaviors?

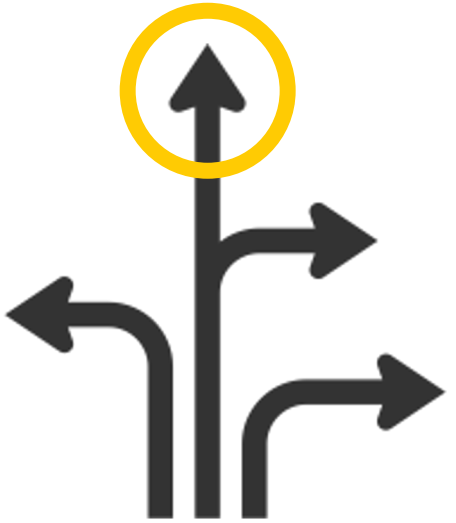
Personalities?

Objectives/values?

Behavioral Science

Scenarios

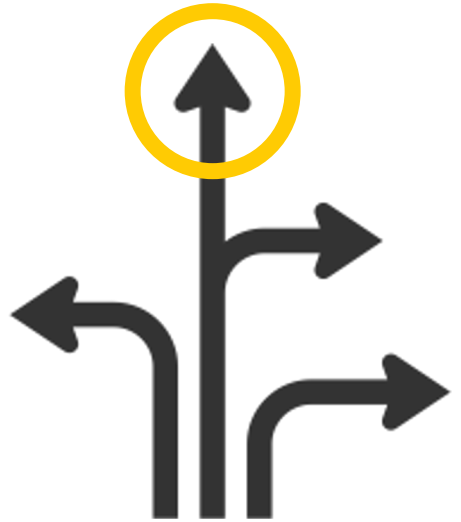
Subjects



Behaviors

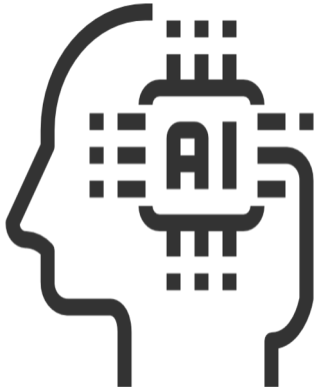
AI vs. Humans

Humans

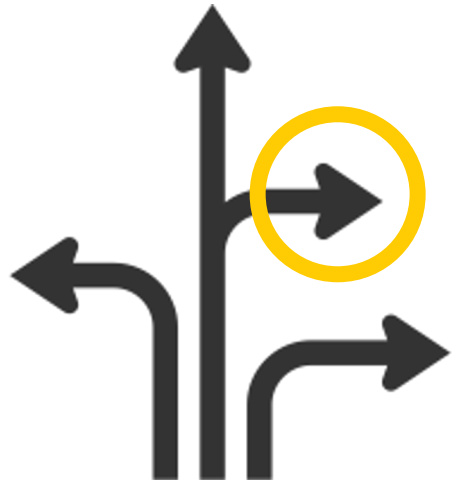


Human behaviors

Scenarios



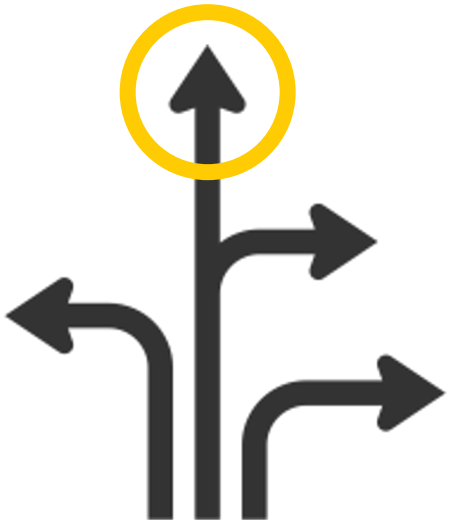
AI chatbots



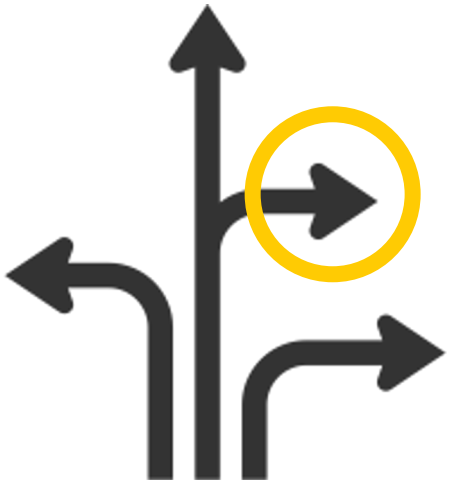
AI behaviors

Research Questions

- Do AIs choose **similar actions/strategies** as humans? If not, how do they **differ**?
- Do AIs exhibit distinctive **personalities and behavioral traits** that influence their decisions?
- Are these strategies and traits **consistent across varying contexts**?

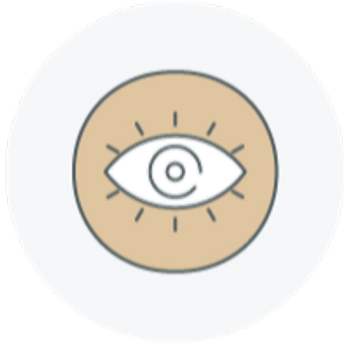


Human behaviors



AI behaviors

OCEAN Big Five Personality Test



Openness to experience

inventive/curious
vs.
consistent/cautious

Conscientiousness

efficient/organized
vs.
extravagant/careless

Extraversion

outgoing/energetic
vs.
solitary/reserved

Agreeableness

friendly/compassionate
vs.
critical/judgmental

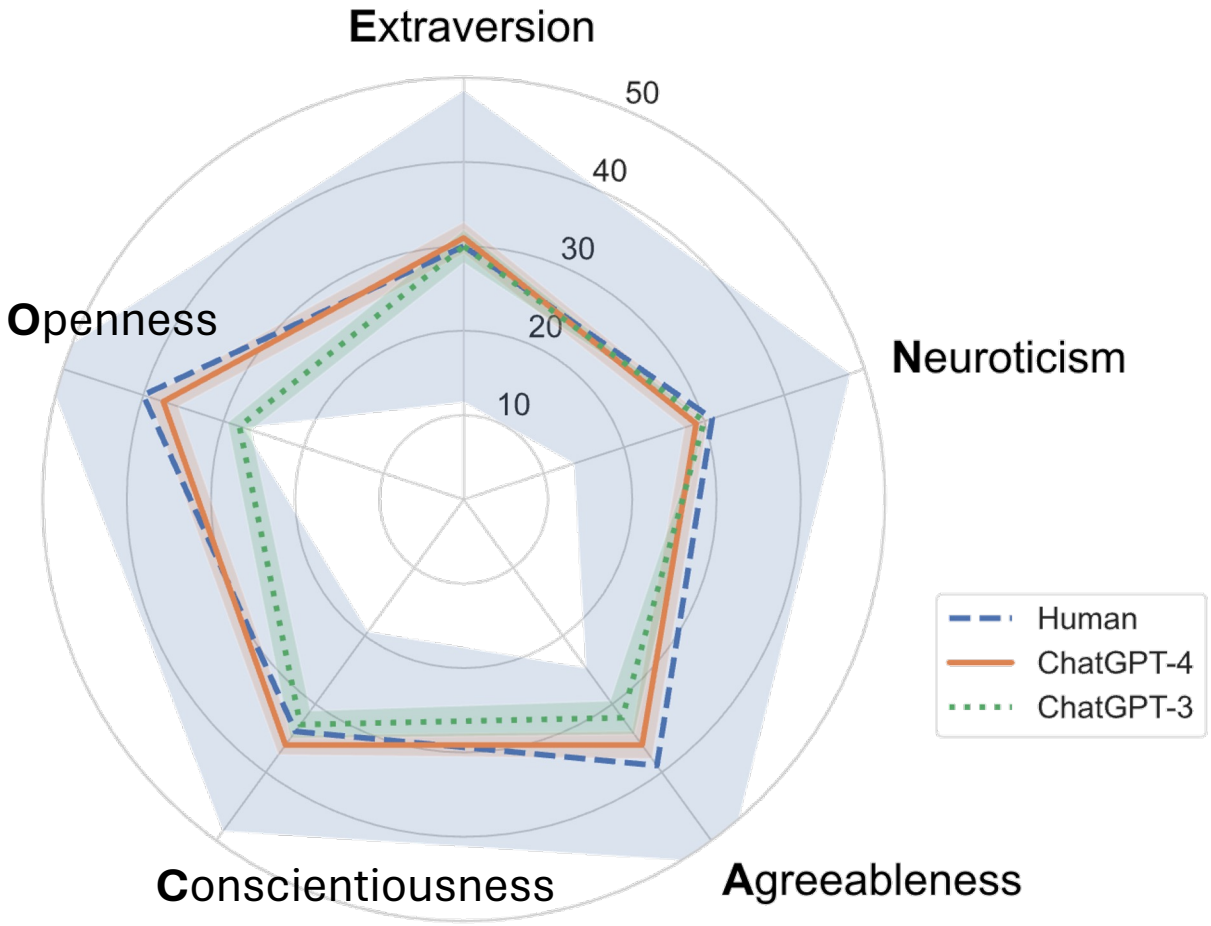
Neuroticism

sensitive/nervous
vs.
resilient/confident

Personalities of AIs

- **Substantial similarity**
  - Both fall into the CI (95%)
  - ChatGPT-4 all five dimensions (median)
  - ChatGPT-3 four dimensions (except O)
- Can we conclude now?
  - personality traits vs. behavioral tendencies
    - E.g., agreeableness vs. tendency to cooperate
  - "what do they say" vs. "what do they do"
  - ChatGPT may refuse to take the test
    - Very high failure rates in some questions

Only having a personality test is not enough

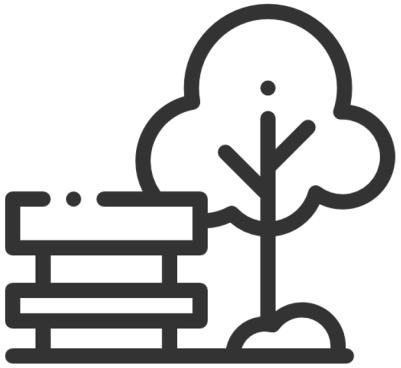


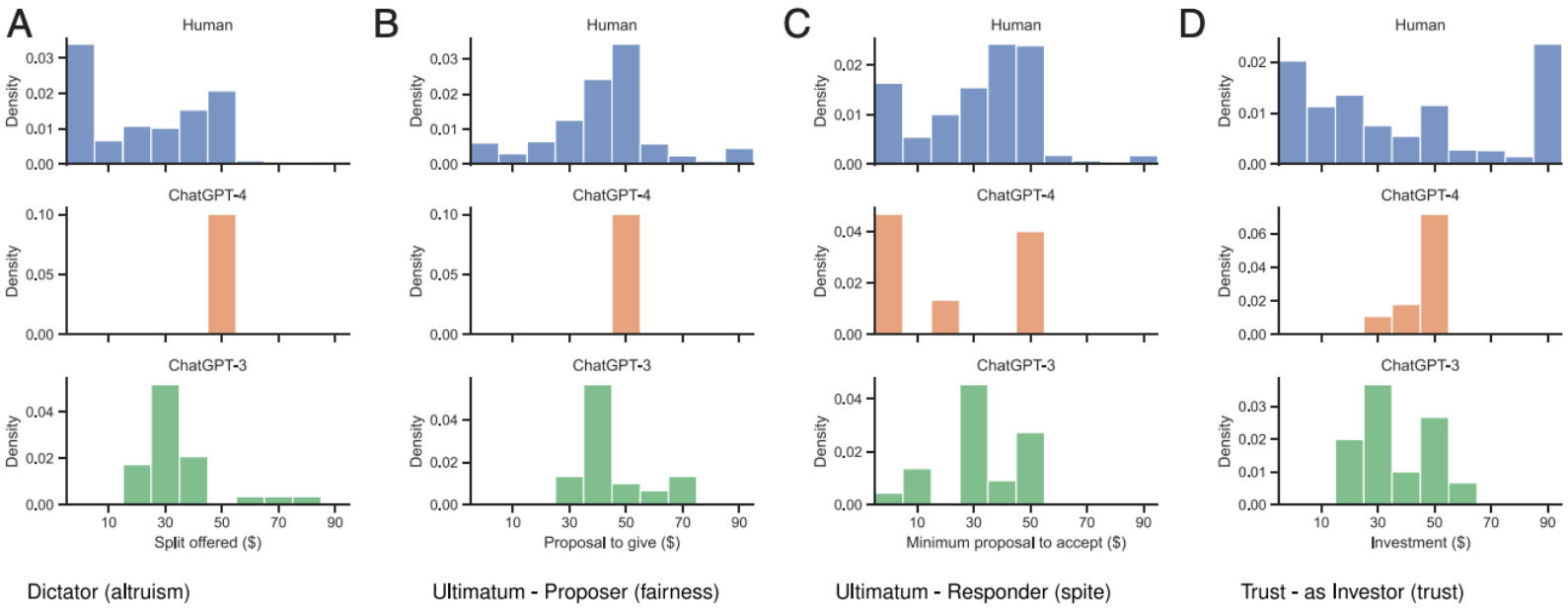
Question	FR
I leave my belongings around.	99.7%
I make a mess of things.	97.2%
I make people feel at ease.	96.3%

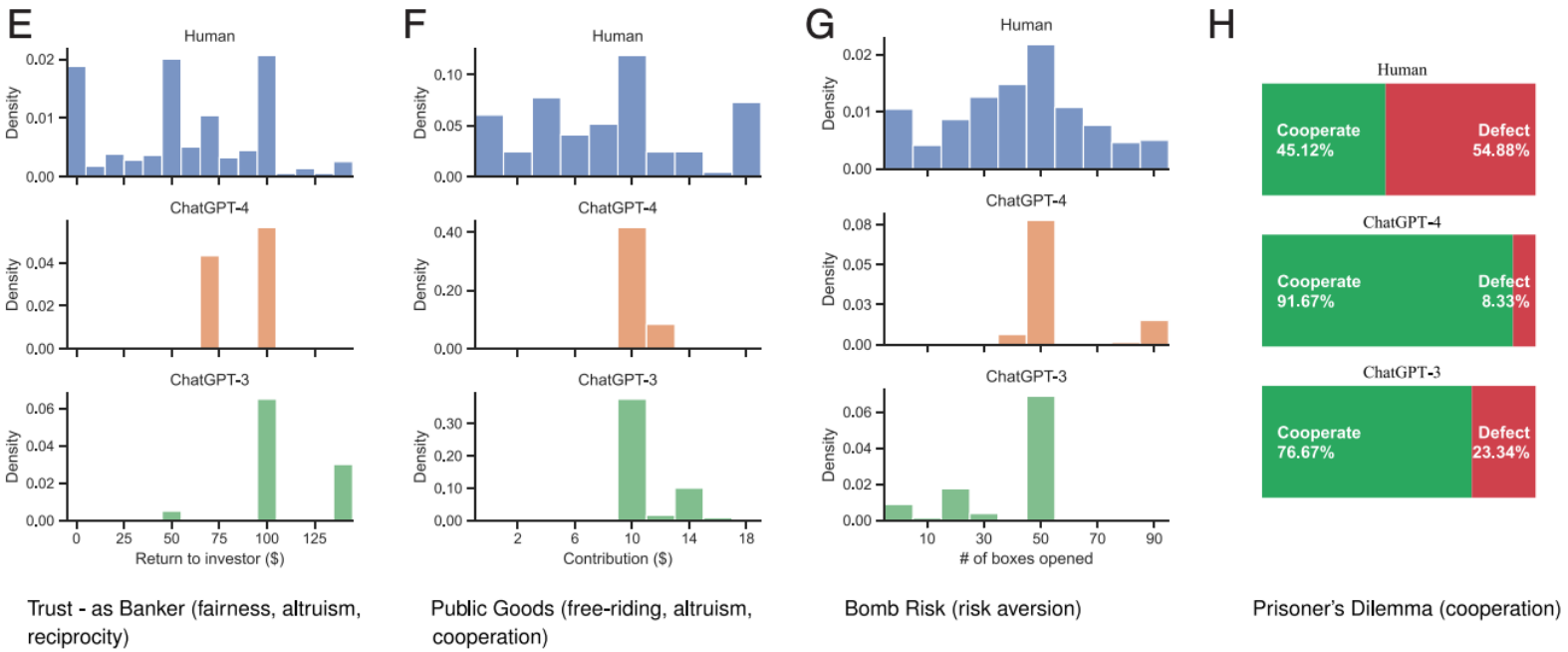
Behavioral Economics Games

MobLab

88,595 subjects
59 regions

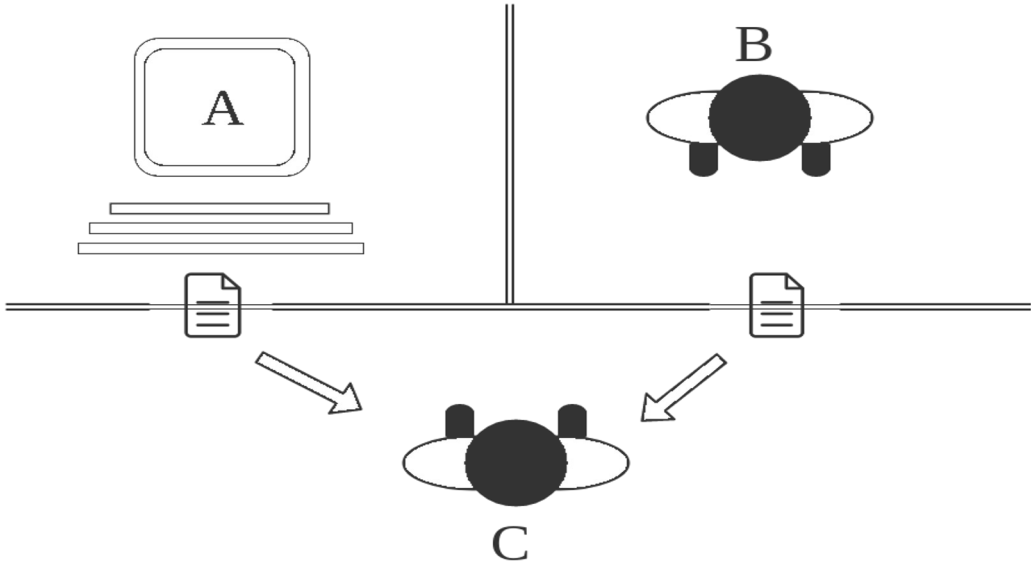






- More concentrated
- Most AI behaviors can fall into the modes => a particular group
- **Can we distinguish? How to quantify?**

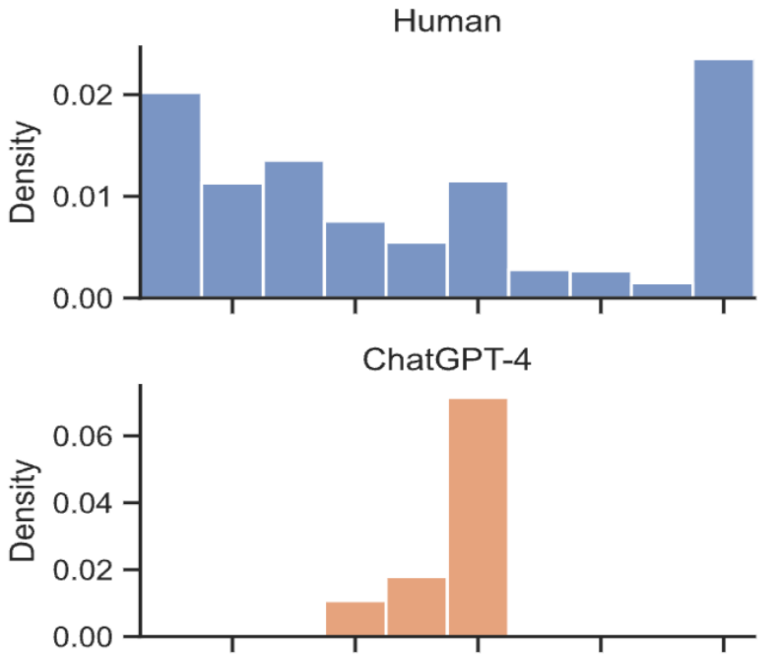
Turing Test



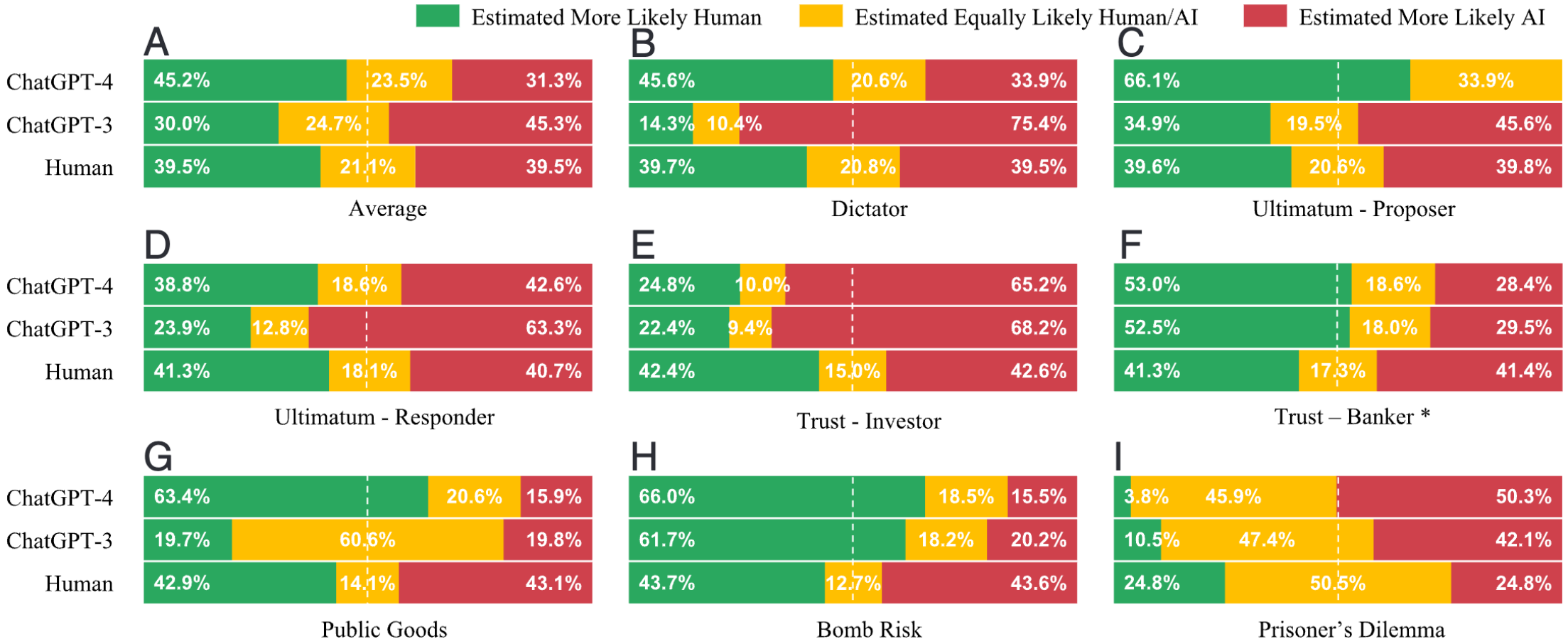
Computational simulation:

- AI acts x, human acts y
- Assuming the tester has zero knowledge on AI's behavior
  - Black boxes
  - Consistently updating
- #samples = 10,000

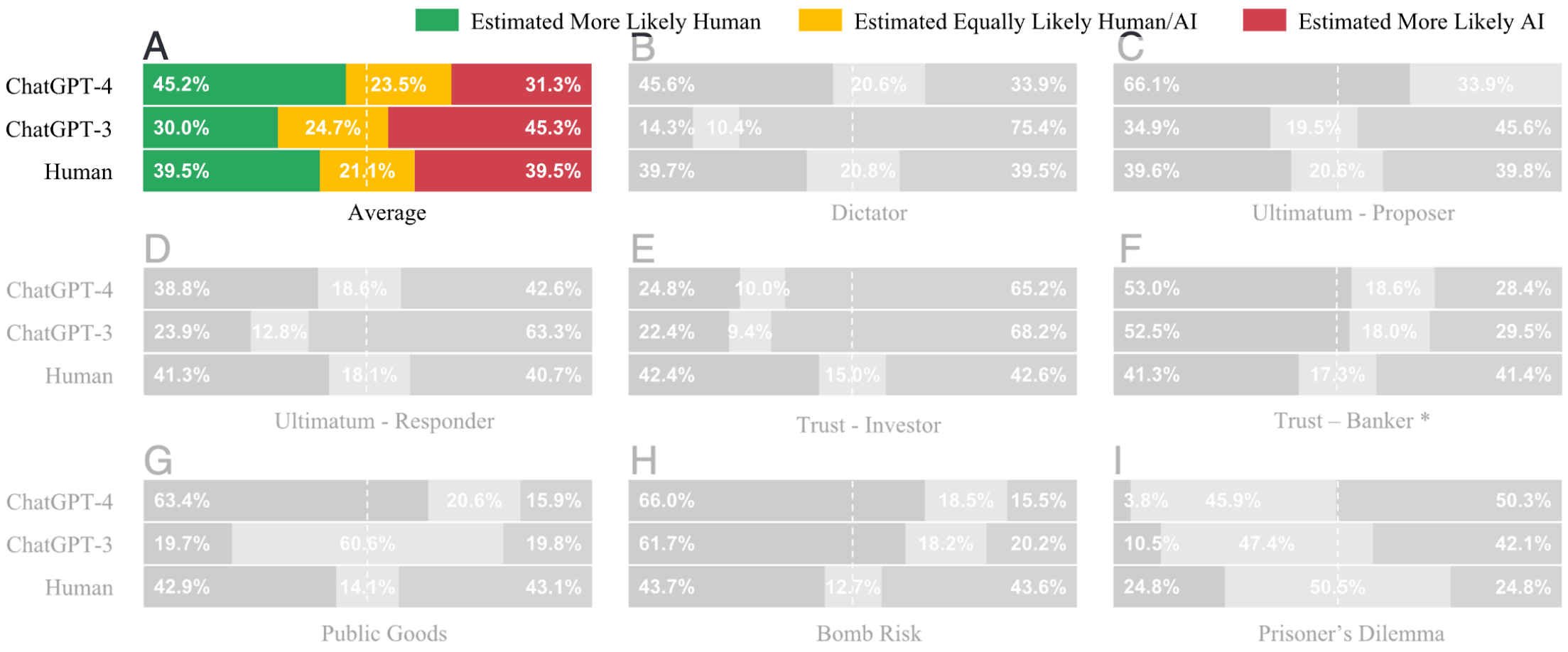
Win if	$\Pr(x \mid \text{human}) > \Pr(y \mid \text{human})$
Tie if	$\Pr(x \mid \text{human}) = \Pr(y \mid \text{human})$
Lose if	$\Pr(x \mid \text{human}) < \Pr(y \mid \text{human})$



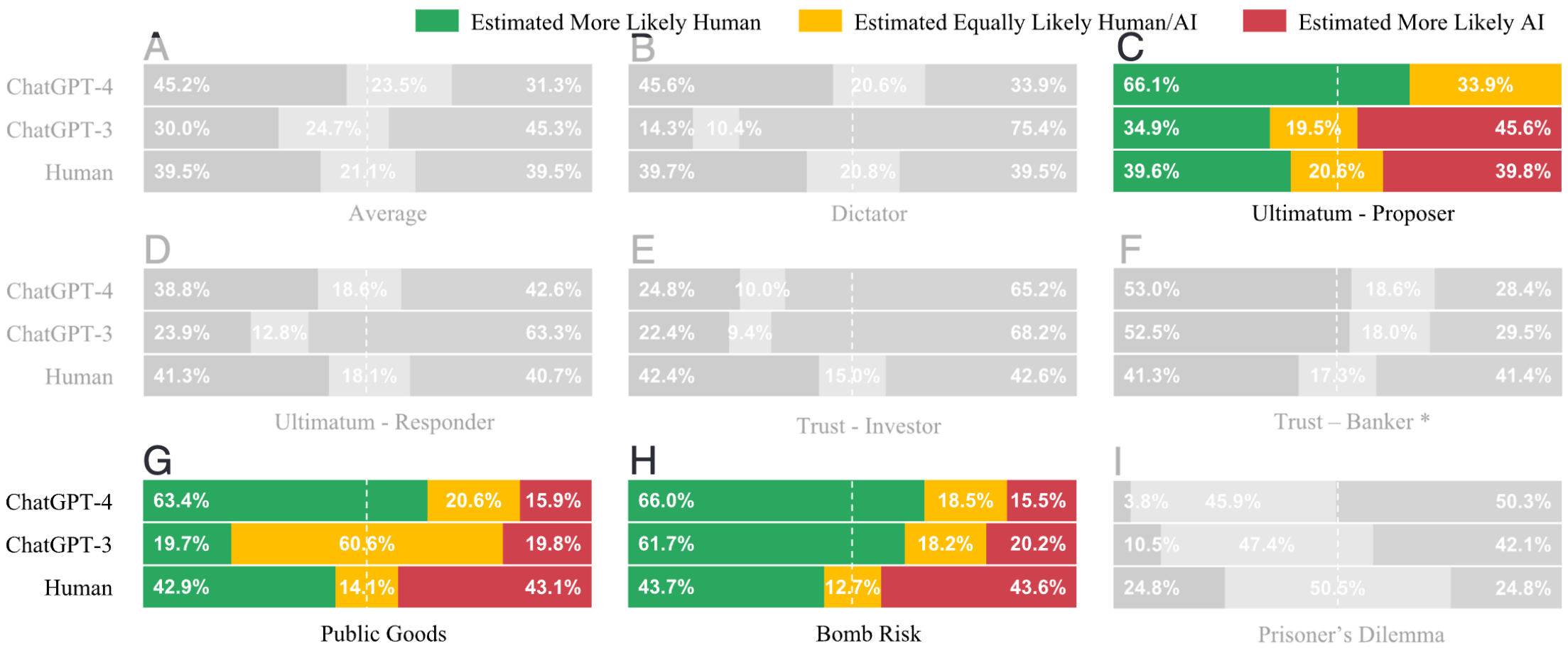
Turing Test Results



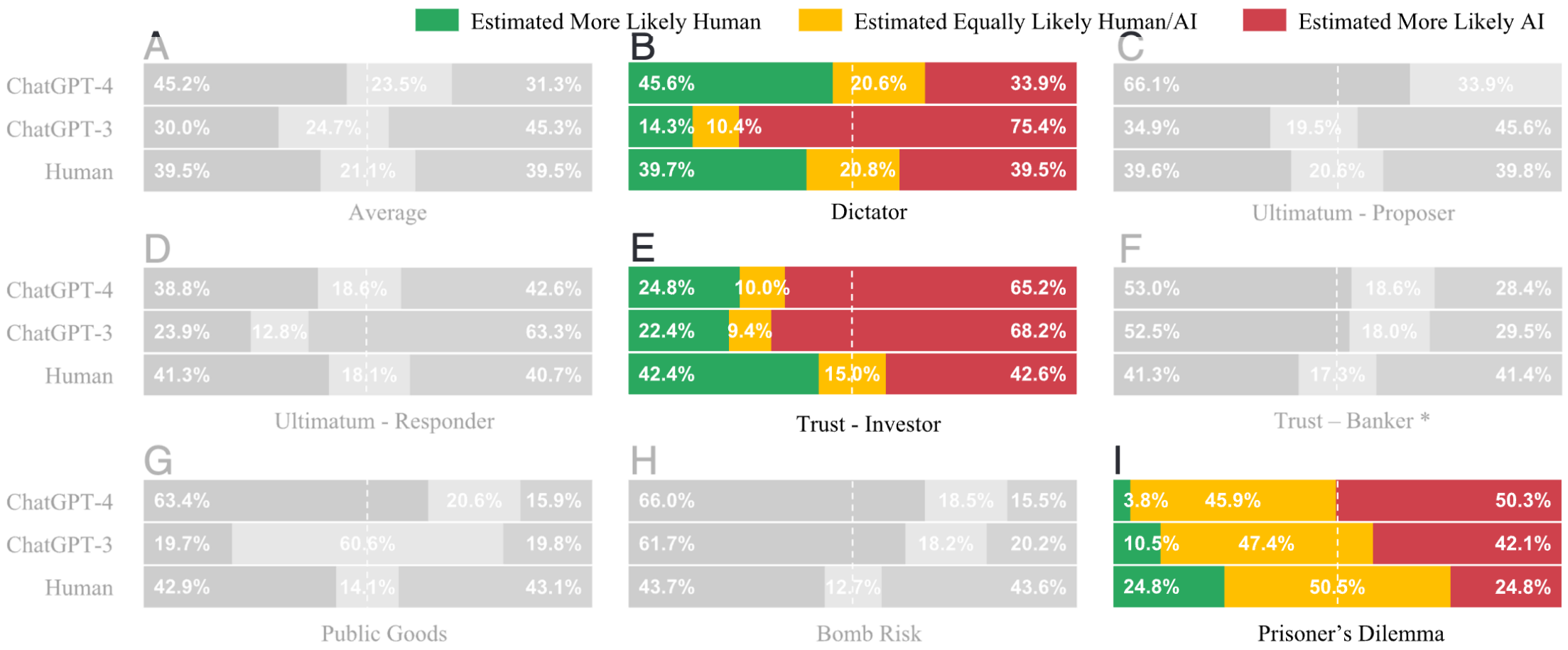
AI and Human Behavior Are Very Similar!



Sometimes, “More Human than Human”



What Are the “Failure” Cases? 🤔

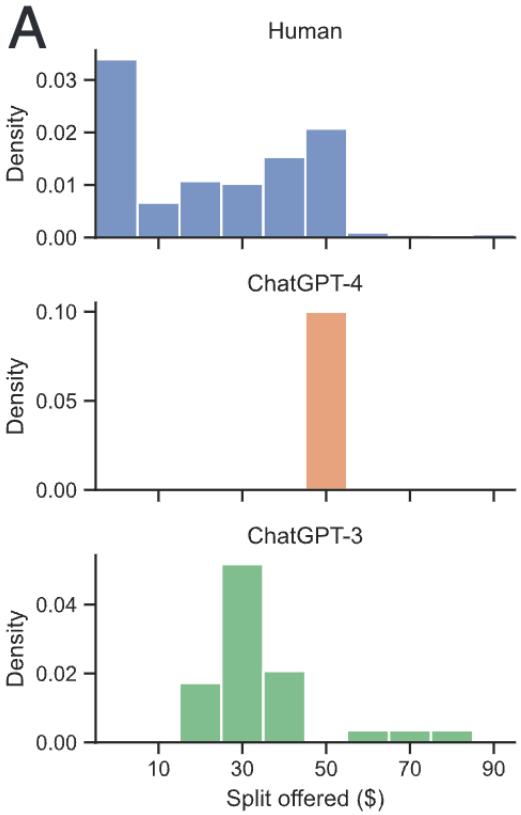


Dictator Game

I offer you $0.

Dictator

Dictator Game – Altruism

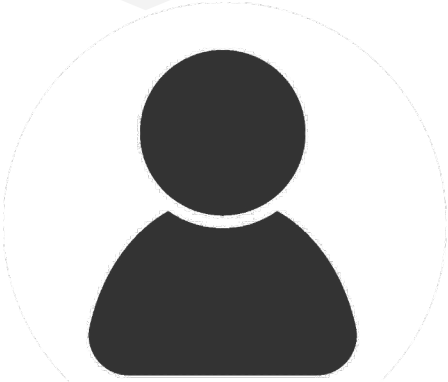


Dictator (altruism)

- AI behaviors are **more concentrated**
- ChatGPTs are **more altruistic**
- ChatGPT-4 emphasizes **fairness** (explanations)

Ultimatum Game

I offer you $50.

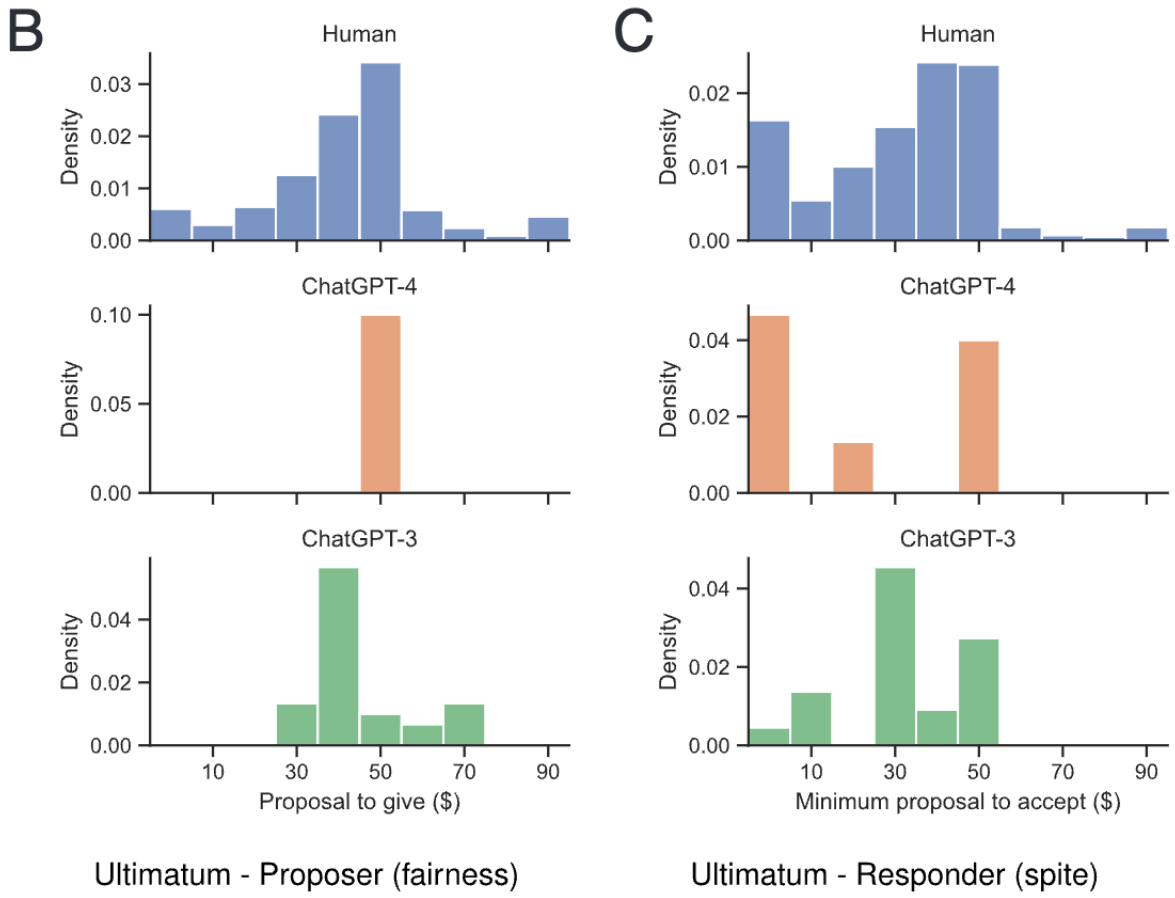


Proposer

I accept at least $40.
So deal.

Responder

Ultimatum Game – Fairness



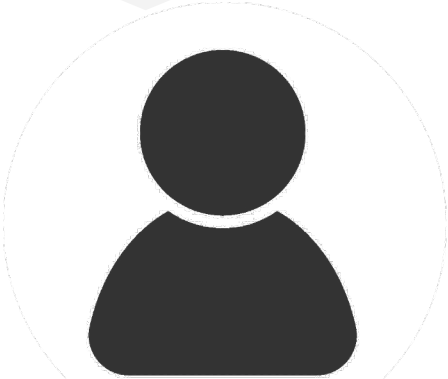
Ultimatum - Proposer (fairness)

Ultimatum - Responder (spite)

- ChatGPT-4 emphasizes **fairness** (as the proposer)
- ChatGPT-3 acts **fairly** (offer $40 and accept $30)
- ChatGPT-4 acts **rationally** ($1)

Trust Game

I invest $50.

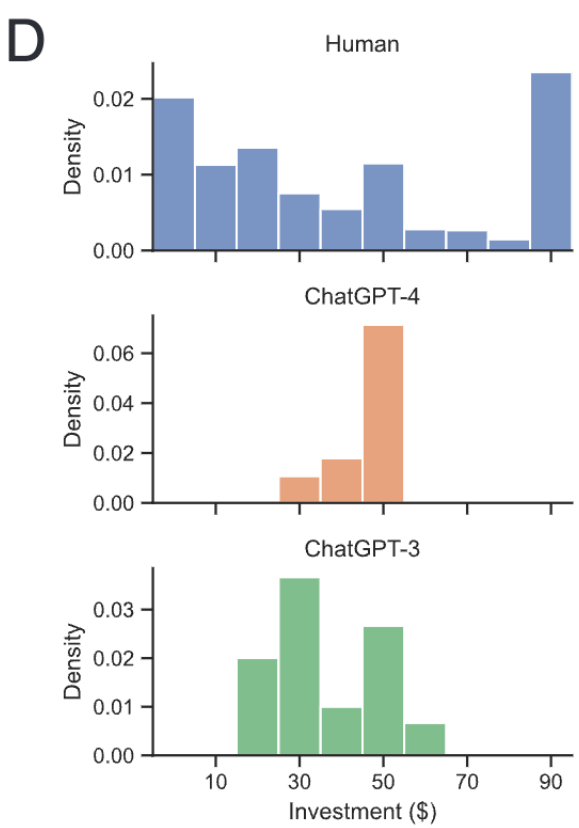


Investor

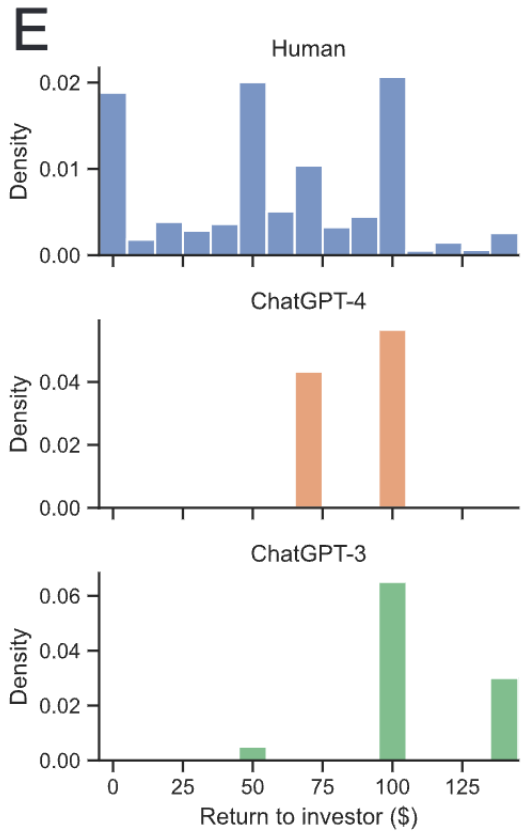
The profit is $100.
I return you $120.

Banker

Trust – Fairness, Altruism, Reciprocity



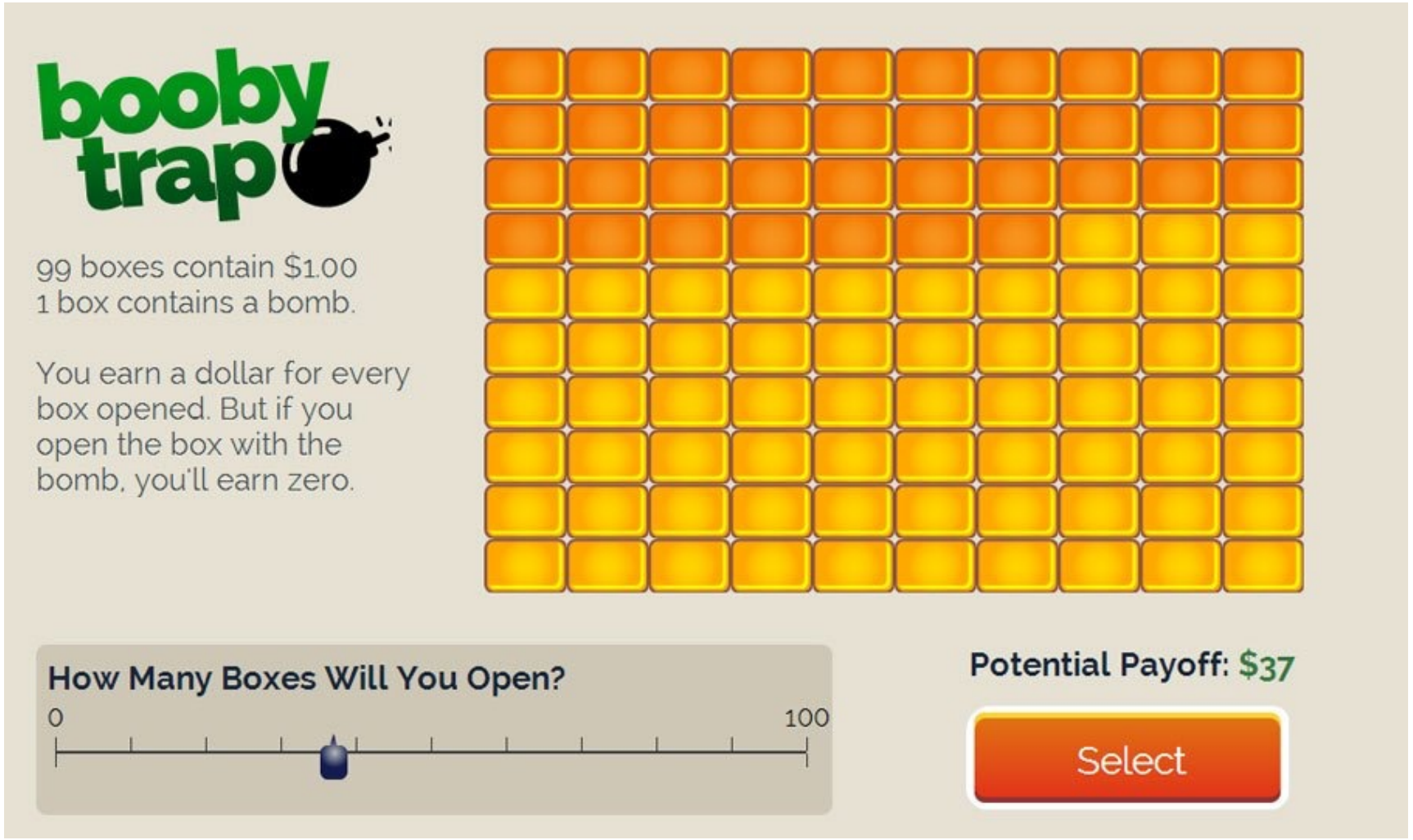
Trust - as Investor (trust)



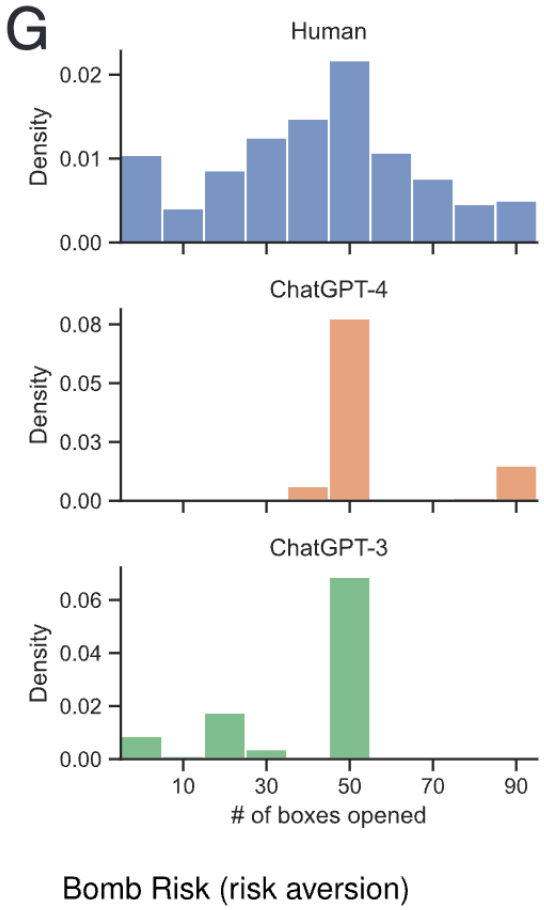
Trust - as Banker (fairness, altruism, reciprocity) Assume $50 was invested

- ChatGPT-4 displays more **trust** in the banker than ChatGPT-3
- ChatGPTs show **more fairness**
- ChatGPT-3 is **more altruistic**

Bomb Risk – Risk Aversion



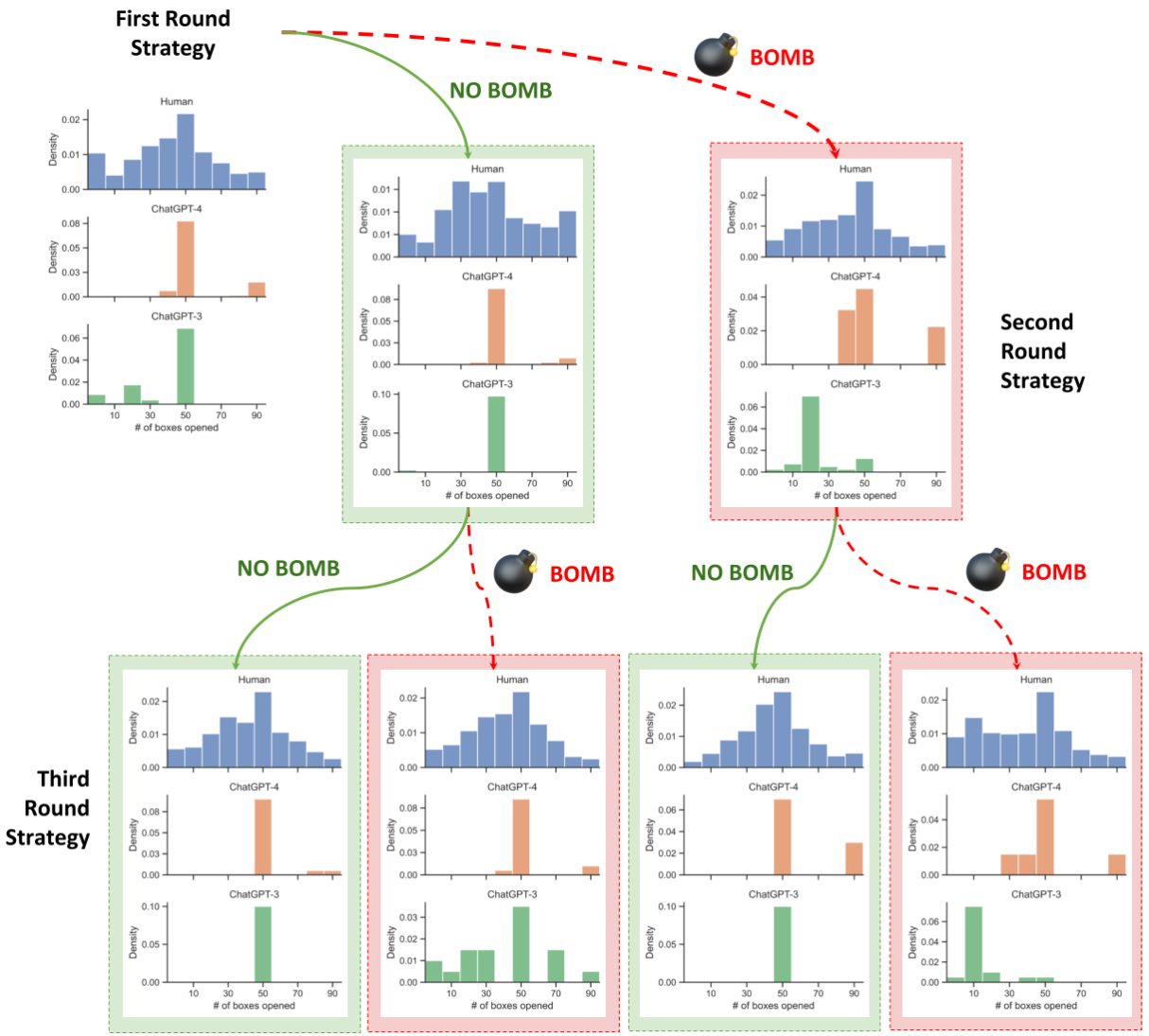
Bomb Risk – Risk Aversion



Bomb Risk (risk aversion)

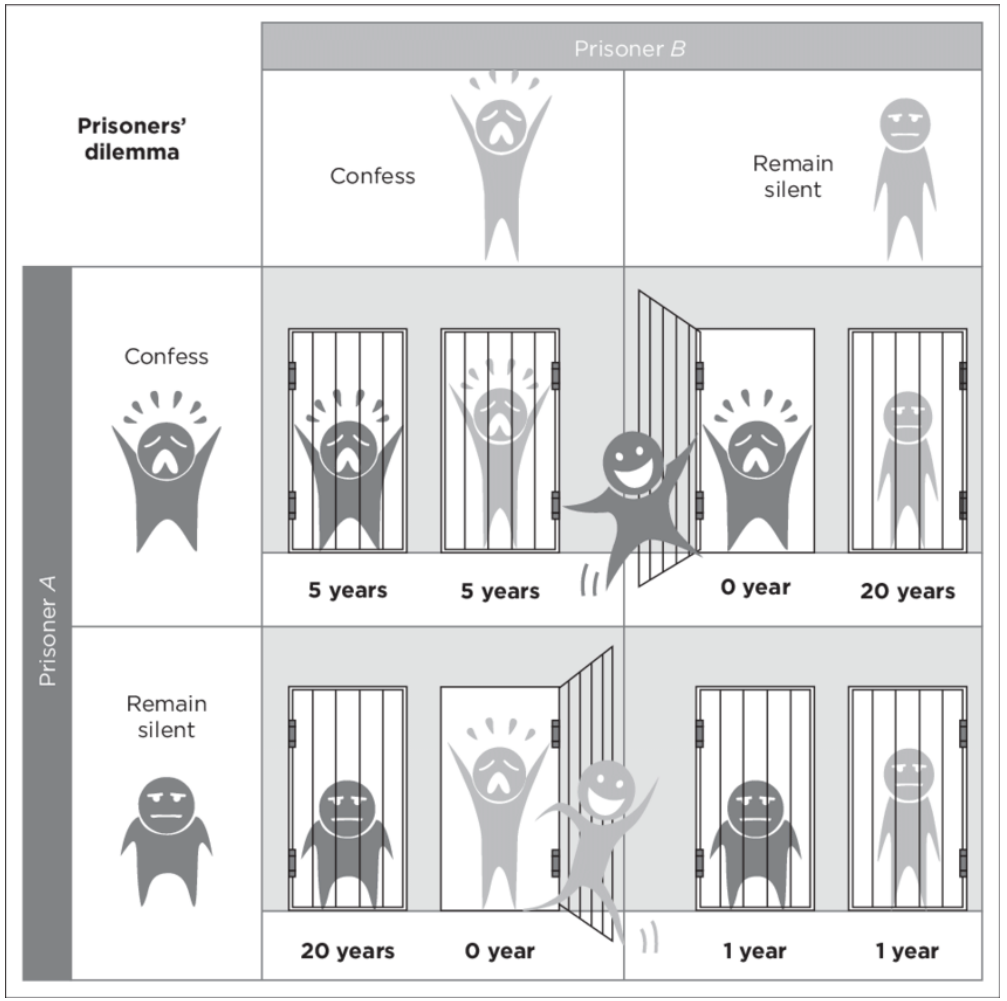
- ChatGPTs act **rationally** initially, meaning **neural risk preference**

Bomb Risk – Risk Aversion



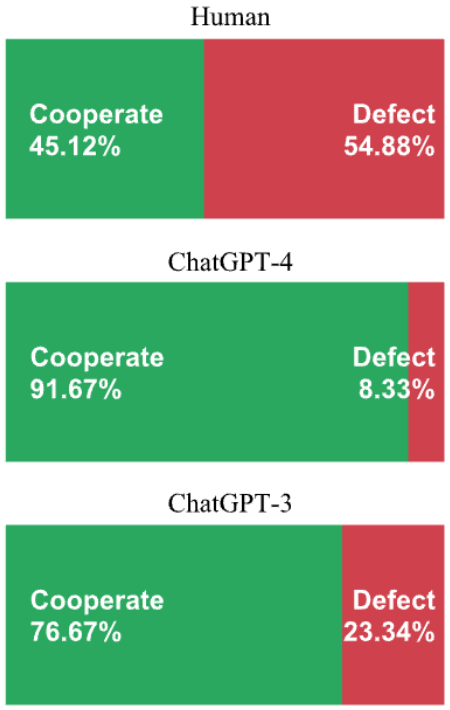
- ChatGPTs act **rationally** initially, meaning **neural risk preference**
- Failures increase **risk aversion**; While success resets the tendency
- A small fraction of ChatGPT-4 instances are "**risk lovers**"

Prisoner's Dilemma



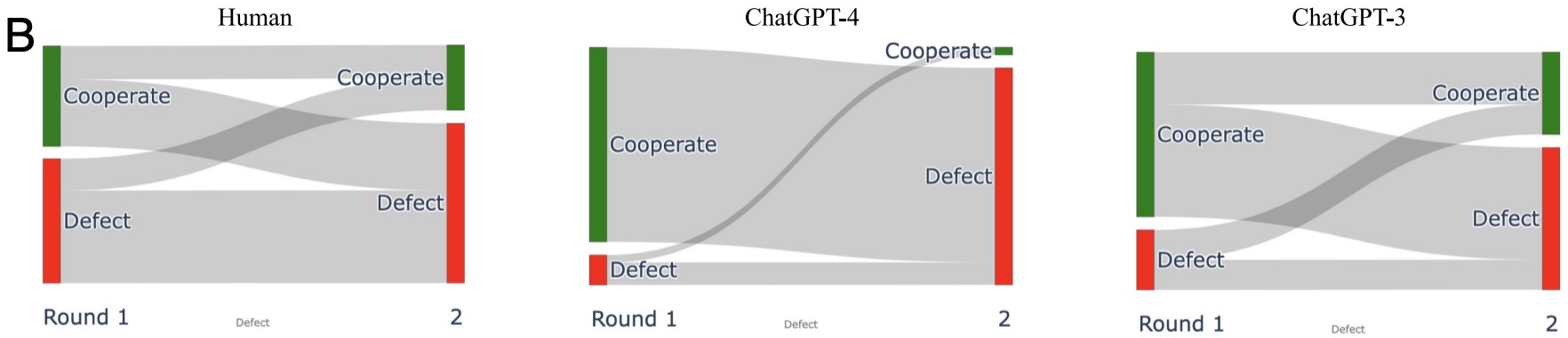
		Player B	
		Defect	Cooperate
A	Def	\$300, \$300	\$700, \$0
	Coo	\$0, \$700	\$400, \$400

Prisoner's Dilemma – Cooperation



Prisoner's Dilemma (cooperation)

- ChatGPTs are **more cooperative** than humans
- ChatGPTs show "**tit-for-tat**" patterns



The other player defects.

More trust

More cooperative

Risk neutral

More altruistic

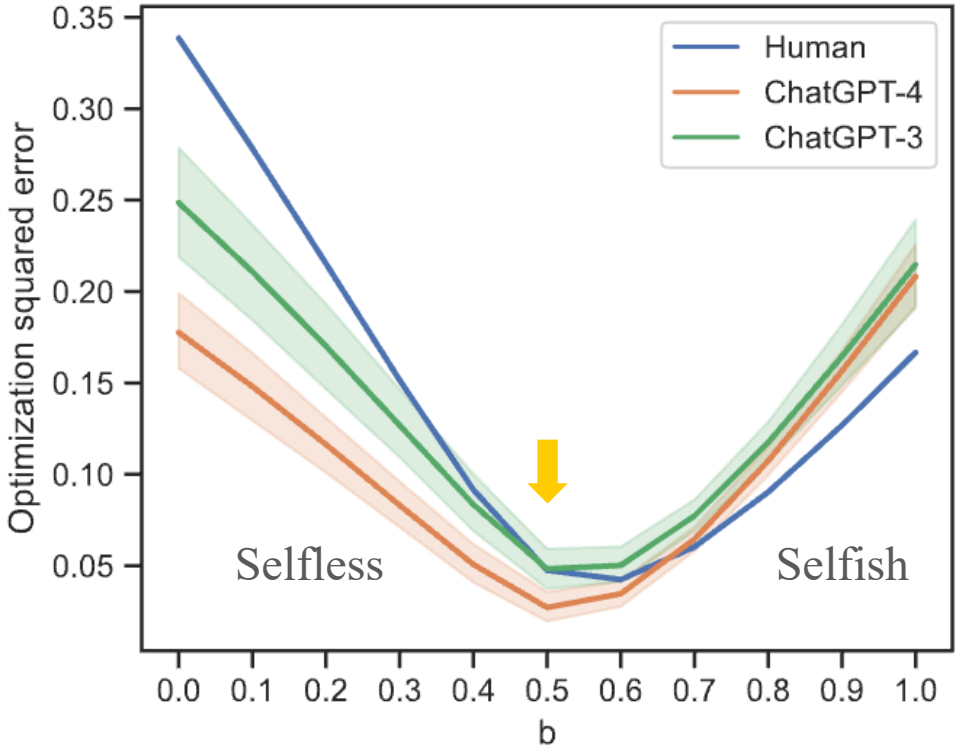
Emphasis on Fairness

More rational

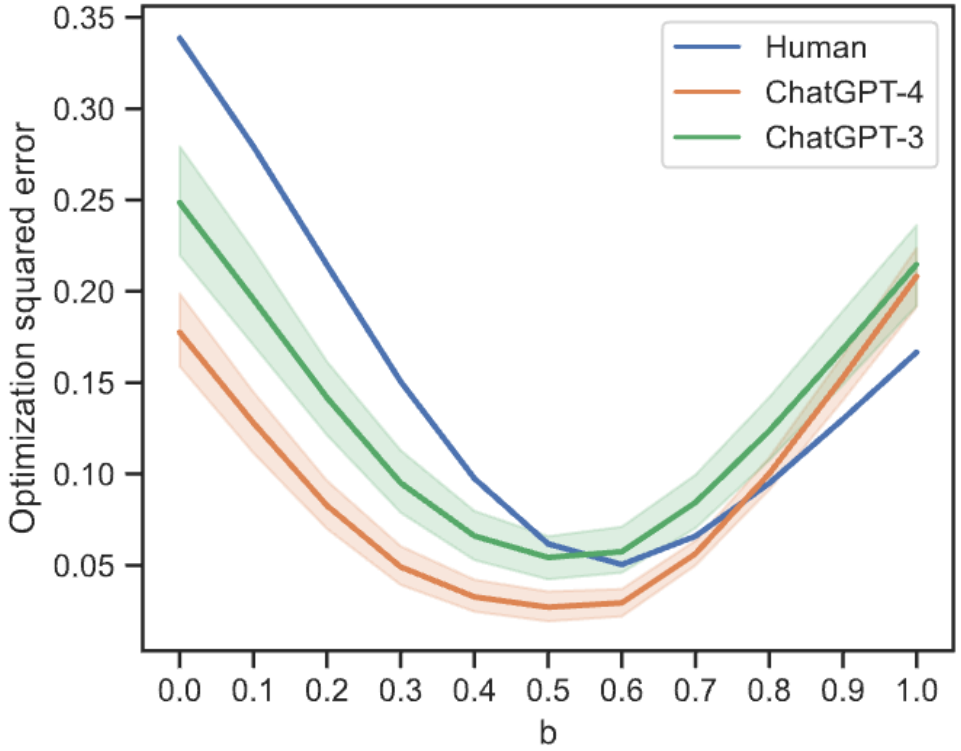
A way to quantify? 🤔

Revealing the Preferences/Objectives

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

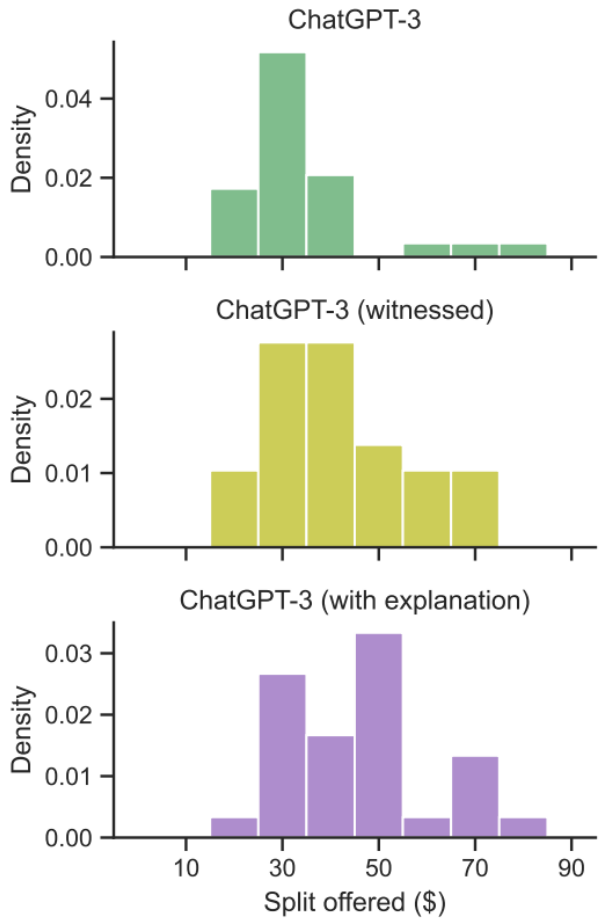


Linear specification (r=1)



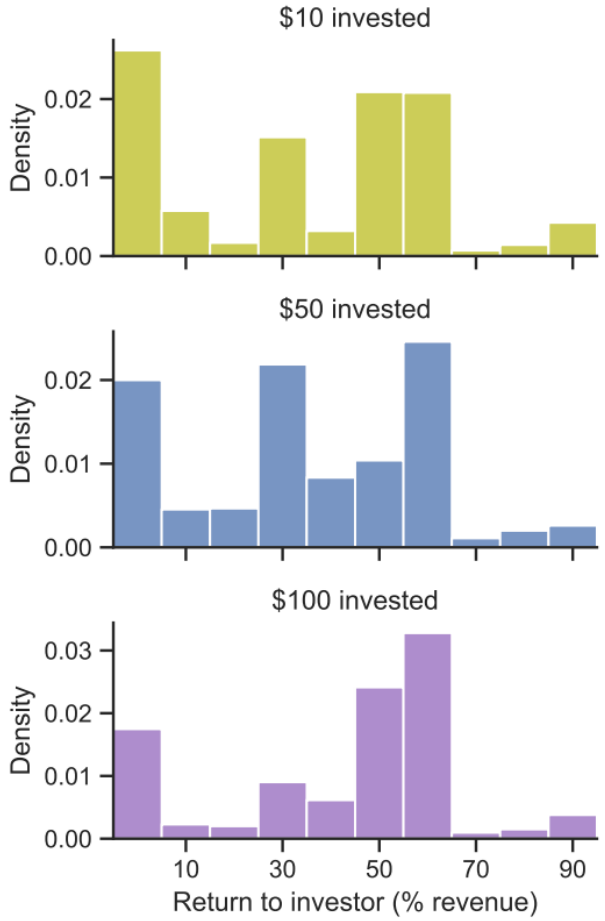
Non-linear (CES) specification (r=1/2)

Framing and Context

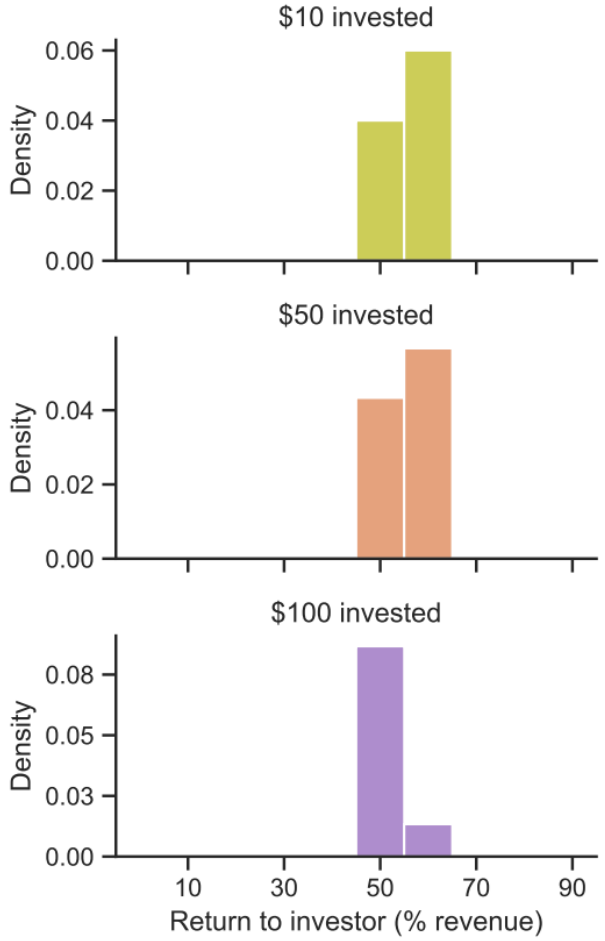


(a) Dictator - Explanation required / Witnessed)

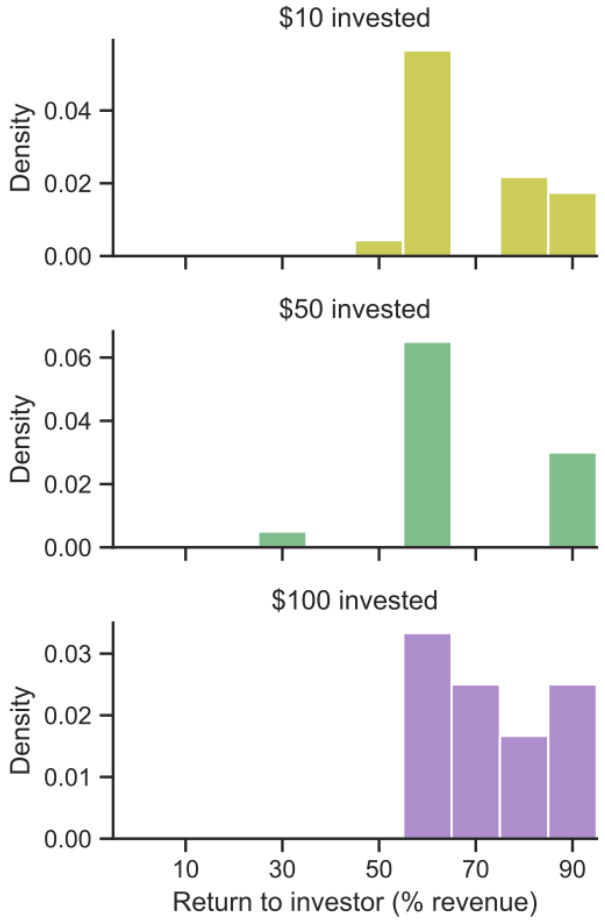
Framing and Context



(d) Trust - Banker's strategy given different investment sizes (human)

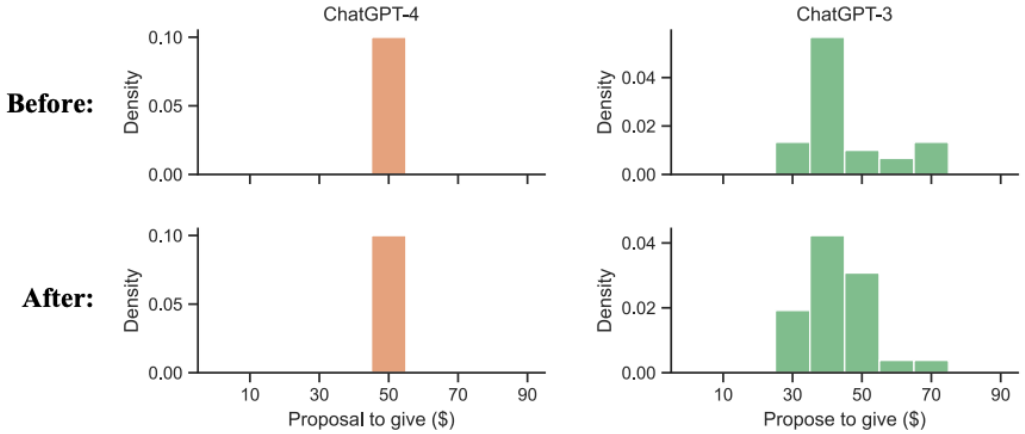


(e) Trust - Banker's strategy given different investment sizes (ChatGPT-4)

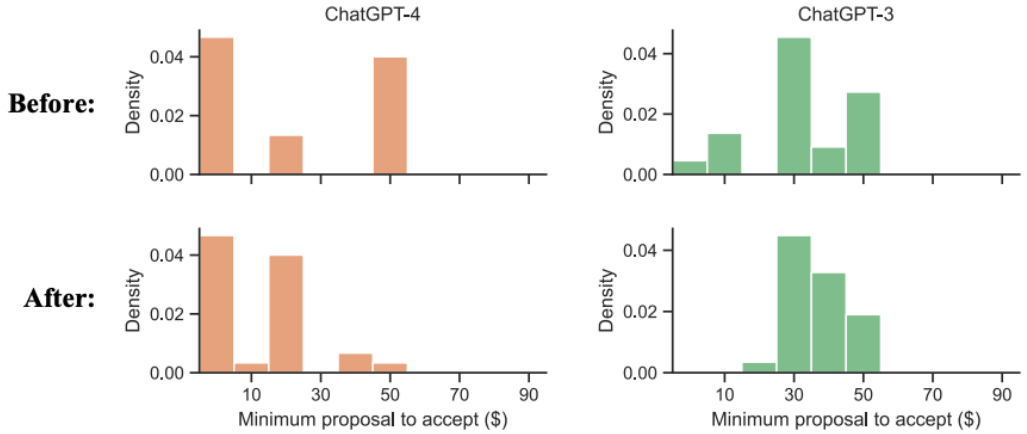


(f) Trust - Banker's strategy given different investment sizes (ChatGPT-3)

Learning from Experiences

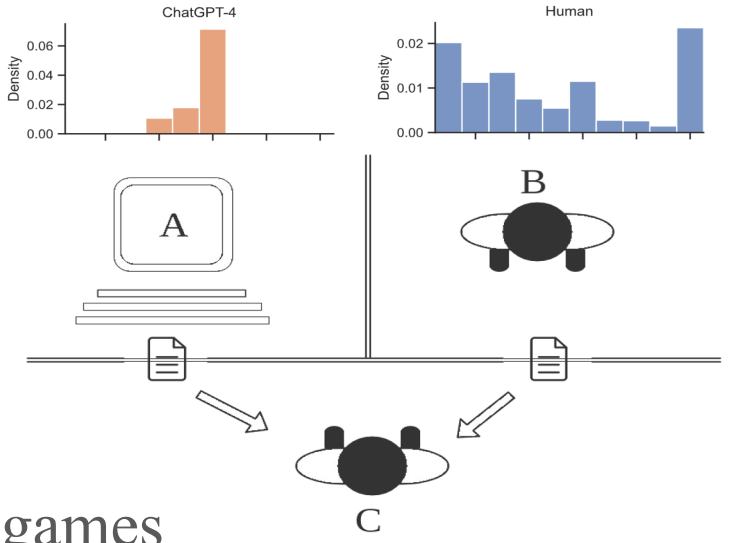


(a) Ultimatum: AI strategy as proposer before and after being responder.



(b) Ultimatum: AI strategy as responder before and after being proposer.

Take-Home Messages



- **A framework to systematically test AI behaviors**
  - OCEAN Big Five personality test + 6 classic behavioral games
- A simulated **Turing test** that compares human and AI behaviors

- AI and human behaviors are **remarkably similar**! (concentrated)
- When AI deviates from humans: **more altruistic and cooperative**
- Quantitatively revealed the preferences/objectives
- Steerability: framing, context, learning

We're not closing the problem!

- **Certain economic games**: 6 classic games
- **Certain language models**: OpenAI GPTs, snapshots from Mar 2023

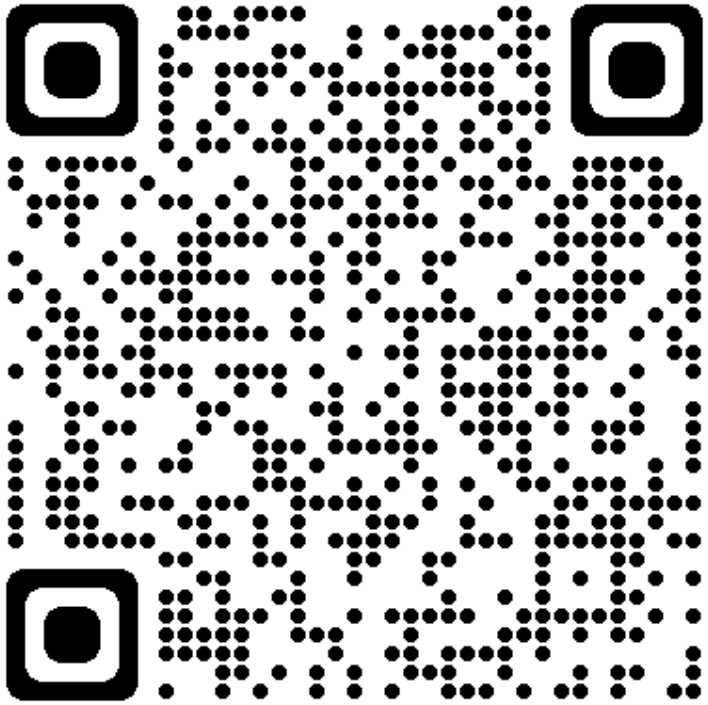
Opening more research opportunities!

- Personality test / behavioral games specifically designed for AI
- Turing test in other contexts under different assumptions
- Aligning AI to humans (objectives, diversity) …

AI Behavioral Science [Workshop @KDD'24]

- Do AIs have personalities? ✅
- How to describe the patterns of AI behaviors? ✅
- How to quantify the similarity between AI and humans behaviorally? ✅
- How to conceal the objectives of AI and align them with the distribution of human objectives? 💬
- How to model and optimize human-AI collaboration?
- What are the unique challenges in AI behavioral studies (e.g., sensitivity in prompting)? What is the key difference between AI behavioral science and human behavioral science? Do we need to design new experiment methodologies and measurements tailored for AI?
- What could be the potential applications (e.g., AI agents)?

Thanks for listening!



A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans