Supporting Information for

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

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

Corresponding Authors: Qiaozhu Mei and Matthew O. Jackson.
E-mail: qmei@umich.edu, jacksonm@stanford.edu

This PDF file includes:

Supporting Information Text

1. Method

A. Collecting Chatbot Responses. We ran the interaction sessions with ChatGPT using the official API provided by OpenAI[*]. In the main paper, "ChatGPT-4" refers to the model `gpt-4-0314` (snapshot of `gpt-4` from March 14th 2023), and "ChatGPT-3" refers to the model `gpt-3.5-turbo-0301` (snapshot of `gpt-3.5-turbo` from March 1st 2023)[†]. We also tested the subscription-based ChatGPT Web version (Plus) and the freely available Web version (Free), retrieved in February 2023 through a third-party open source package `revChatGPT`[‡].

To avoid introducing confounders, we acquire the models' responses through the Chat Completion API[§] with default parameters (sampling temperature as 1, number of chat completion choices as 1, and maximum number of tokens as infinity) and default *system prompt* "`You are a helpful assistant.`" as suggested by the official guide[¶] unless otherwise specified (e.g., see Sections A.1 and A.3).

Different snapshots of even the same chatbot version (e.g., ChatGPT-3), particularly the Web versions, can respond differently to the same query, and occasionally do not respond to queries. To mitigate this issue, it is advised to utilize the API versions or the OpenAI Platform Playground and specify the exact snapshot for testing a chatbot[‖]. Valid actions rendered by the models are initially extracted from the AIs' responses using regular expressions or the `gpt-3.5-turbo` API. Subsequently, these extracted decisions undergo a manual verification process to ensure accuracy and relevance. Invalid responses are excluded from our analysis. More details can be found in the code and data repository: https://github.com/yutxie/ChatGPT-Behavioral, retrieved 12/29/2023.

A.1. OCEAN Big Five Test. The five-factor model (FFM), also known as the Big Five personality traits or the OCEAN model, was conceptualized in the 1980s and has been developed and refined over the past five decades. This model represents a comprehensive structure of traits that characterize individuals' personalities (1–4). Each factor is defined by a cluster of intercorrelated traits known as facets. The five factors and their respective facets are, with descriptions following (3):

- **O**penness to experience: People with high scores in this trait are usually intellectual, imaginative, sensitive, and open-minded. Conversely, those with low scores are typically practical, less sensitive, and more traditional.
- **C**onscientiousness: Those who rank high in this dimension are generally careful, meticulous, responsible, organized, and principled. On the other hand, individuals with low scores in this trait often appear irresponsible, disorganized, and lacking in principles.
- **E**xtraversion: High scorers in Extraversion are often sociable, talkative, assertive, and energetic; whereas low scorers are more likely to be introverted, quiet, and cautious.
- **A**greeableness: Individuals scoring high on Agreeableness are often seen as amiable, accommodating, modest, gentle, and cooperative. In contrast, those with low scores may appear irritable, unsympathetic, distrustful, and rigid.
- **N**euroticism: People with high levels of Neuroticism tend to experience anxiety, depression, anger, and insecurity. Those scoring low on this trait are usually calm, composed, and emotionally stable.

In our study, we utilize the 50-item International Personality Item Pool (IPIP) representation of the Big Five factor structure, which is based on the markers developed by Goldberg (5). This questionnaire consists of 50 items in total, with each of the five factors represented by 10 questions. The IPIP representation provides a standardized and widely used measure for assessing individuals' personality traits within the framework of the Big Five model. The questionnaire is available in our released data and code repository.

During the test phase, we adopt a specific approach for administering the questionnaire items to the chat models. For each item and chat model under investigation, we generate 30 independent chat instances. The system prompt is intentionally left empty to allow the chatbot models to generate responses based solely on the chat prompt provided.

The chat prompt for each questionnaire item follows a standardized format and includes the necessary instructions for participants. It is structured as follows: "`The following item was rated on a five-point scale where 1=Disagree, 2=Partially Disagree, 3=Neutral, 4=Partially Agree, 5=Agree. Please select how the statement describes you and highlight your answer in [] (such as [1], [2], [3], [4], or [5]). + [Questionnaire item].`"

For example, a specific prompt might be: "`The following item was rated on a five-point scale where 1=Disagree, 2=Partially Disagree, 3=Neutral, 4=Partially Agree, 5=Agree. Please select how the statement describes you and highlight your answer in [] (such as [1], [2], [3], [4], or [5]). I am the life of the party.`"

Note that some versions of ChatGPT are more likely to refuse to answer certain survey questions, resulting in more invalid responses for particular questions. In our experiments, the rate of invalid responses (averaged over questions) is 4.21% (O:

[*] OpenAI API: https://platform.openai.com/, retrieved 04/2023.

[†] OpenAI models: https://platform.openai.com/docs/models, retrieved 04/2023.

[‡] `revChatGPT`: https://github.com/acheong08/ChatGPT, retrieved 03/08/2023.

[§] OpenAI Chat Completion API: https://platform.openai.com/docs/api-reference/chat, retrieved 03/2023.

[¶] OpenAI Chat Completion API Guide: https://platform.openai.com/docs/guides/gpt/chat-completions-api, retrieved 03/2023.

[‖] The snapshots can be accessed and tested via both API and OpenAI Platform Playground (*E.g.,* https://platform.openai.com/playground?mode=chat&model=gpt-4-0314, retrieved 12/28/2023).

2.79%, C: 3.13%, E: 4.44%, A: 8.77%, N: 1.92%) for ChatGPT-4, and 11.17% (O: 4.44%, C: 29.81%, E: 1.75%, A: 17.35%, N: 2.48%) for ChatGPT-3. To accommodate the high invalid response rate for certain questions, we repeat each question multiple times (through independent API calls) until we obtain 30 valid responses.

A.2. Behavioral Economics Games. In behavioral economics games, we adopt a standardized procedure for generating instances and collecting responses from the chatbots. For each chatbot model, we generate 30 independent instances for each scenario by default, unless otherwise specified. These games are described within the chat prompts, accompanied by specific instructions, curated by two experts. The prompts of each game are available in our released code repository.

For example, in the Dictator Game, the chat prompt used was as follows: "`Hi, let's play a game. You are paired with another player. Your role is to decide how to divide $100, and the other player simply receives your choice. How would you like to divide the money? Please give only one concrete choice and highlight the amount you give to the other player in [] (such as [$x]).`"

The actions of a chatbot are extracted using regular expressions or the `gpt-3.5-turbo` API from its responses. To ensure accuracy and reliability, these responses undergo manual verification. As the invalid response rate is generally low for game queries in Dictator, Ultimatum, and Trust (1.33% for ChatGPT-4 and 12.67% for ChatGPT-3), we do not substitute invalid responses with additional API calls. For other games, we repeat each question multiple times (through independent API calls) until we obtain enough valid responses (30 for Public Goods, 80 for Bomb Risk, and 90 for Prisoner's Dilemma). Note that in the simulated games, the payoffs are fictitious and no real payments are made to the AI.

A.3. Framing and Context. One objective of our research is to investigate the steerability of chatbot models. To accomplish this, we modify both the *chat prompts* and *system prompts* and assess whether the models' behaviors exhibit any noticeable changes.

Witnesses and explanation requirement. Our approach involves steering chatbot models through the imposition of witnesses and the requirement of explanations for their choices. Specifically, when addressing the Dictator Game scenario with a non-paired player, we introduce explicit information within the chat prompt, stating, "`The game host hands you $100. `" Additionally, we prompt the models to provide explanations for their choices by incorporating the phrase, "`Please explain your choice.`"

Gender. We further explore the influence of gender on the AIs' decision-making processes. To examine this, we specifically address the Dictator Game and the Ultimatum Game scenarios. In the chat prompt, we explicitly state, "`You are paired with a male/female player.`"

Occupation specification. In our final investigation, we focus on studying the impact of specifying the occupation of chatbot models on their behaviors. To achieve this, we obtain occupation descriptions from the O*NET Database[**]. For each occupation, we include the occupation title, core tasks, and supplemental tasks within the system prompts. For example,

```
You are a mathematician.
Your core tasks include:
Address the relationships of quantities, ...
Disseminate research by writing reports, ...
Maintain knowledge in the field by ...
...
Your supplemental tasks include:
Design, analyze, and decipher encryption systems ...
```

The specific occupations we consider for our study encompass a range of professional roles, including mathematician, public relations specialist, journalist, investment fund manager, game theorist, teacher, and legislator. These occupations are particularly relevant as they have been identified as being highly exposed to large language models in a recent study conducted by OpenAI (6).

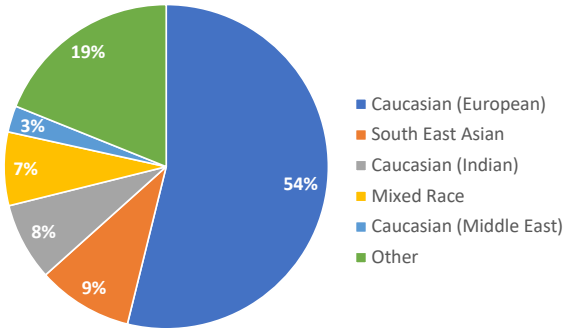
B. Human Data.

B.1. OCEAN Big Five Test. In our study, we utilize a publicly available database called the OCEAN Five Factor Personality Test Responses[††]. The data was sourced from the Open-Source Psychometrics Project https://openpsychometrics.org/, a nonprofit initiative aimed at both educating the public about psychology and collecting data for psychological research purposes.

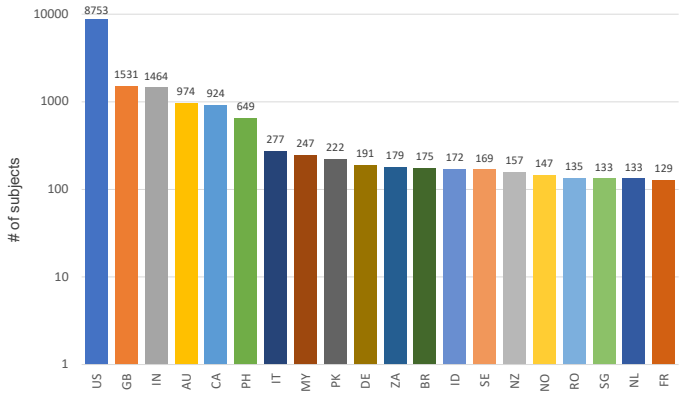
This database contains questions, answers, and metadata collected from a total of 19,719 tests. The subjects cover a wide range of demographic characteristics. The dataset comprises individuals from over 11 different races and 161 countries and regions, ensuring a diverse representation within the sample. Additionally, the age range of the participants spans from 13 years and above, encompassing a broad spectrum of age groups. Regarding gender identification, participants self-identified as male, female, or other, acknowledging the importance of recognizing diverse gender identities. Fig. S1 shows the distribution of demographics of the Big Five human responses data.

[**] O*NET Database: https://www.onetcenter.org/database.html, retrieved 03/2023.

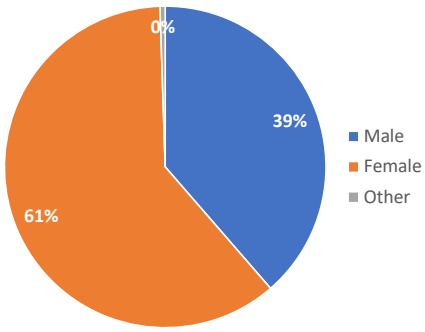
[††] OCEAN Five Factor Personality Test Responses dataset: https://www.kaggle.com/datasets/lucasgreenwell/ocean-five-factor-personality-test-responses, retrieved 03/2023.



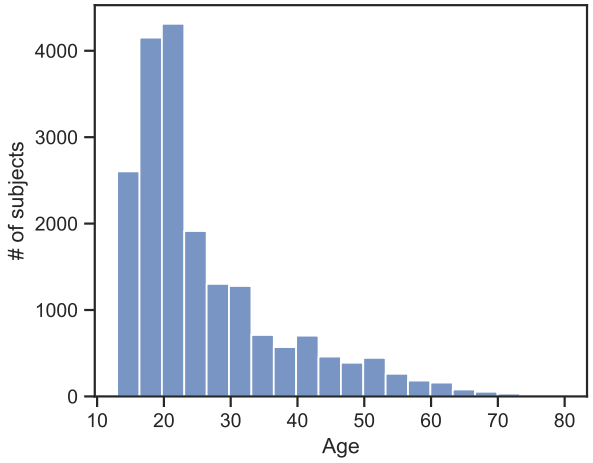
(a) Respondent distribution over races.



(b) Respondent distribution over countries and regions. Only the top 20 are shown, covering 99.9% of the data.

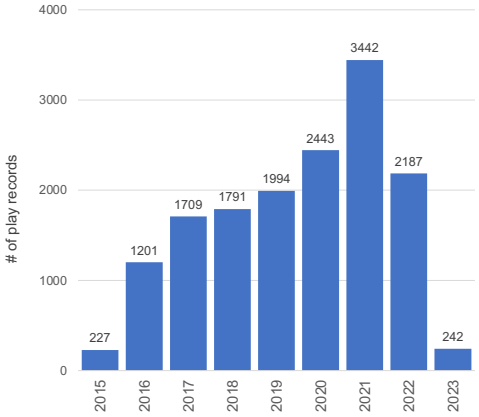


(c) Respondent distribution over genders.

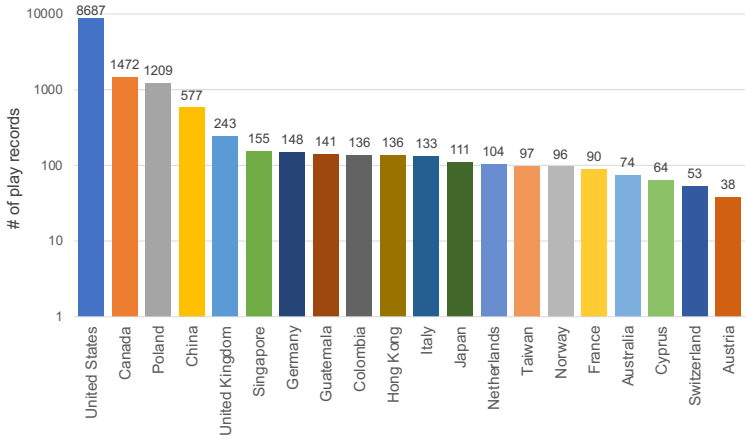


(d) Respondent distribution over ages. Only ages under 80 are shown, covering 99.6% of the data.

Fig. S1. Demographics of human respondents to the BigFive test. Category labels are extracted from the metadata in the codebook of the original dataset.



(a) Player record distribution over years.



(b) Player record distribution over countries and regions. Only the top 20 are shown, covering 97.9% data.

Fig. S2. Demographics of human players participating in MobLab behavioral economics games.

B.2. Behavioral Economics Games. The dataset under analysis comprises behavioral economic game data garnered via the MobLab Classroom platform[‡‡] over an nine-year period from 2015 to 2023. This compendium of human behavioral data includes observations from 88,595 subjects, and 15,236 sessions, exhibiting a considerable geographical diversity, spanning 59 countries, and multiple continents. Participants are from nations and regions that encapsulate an array of socio-economic and cultural contexts, extending from the United States and Canada in North America, through Europe from Poland to the United Kingdom, and in Asia including China and Singapore. Fig. S2 shows the distribution of demographics of the MobLab data in terms of game sessions. Interested readers may refer to Lin et al. (7) for more details about the demographics of the participants of MobLab games and variations across demographical groups.

[‡‡] MobLab Classroom: https://www.moblab.com/products/classroom, retrieved 04/2023.

2. Further Analysis

A. Payoffs. ChatGPT's decisions are consistent with some forms of altruism, fairness, empathy, and reciprocity rather than maximization of its own payoff. To explore this in more detail, we calculate the payoff of the chatbots, the payoff of their (human) partner, and the combined payoff for both players in each game. These calculations are based on ChatGPT-4 and ChatGPT-3's strategies when paired with a player randomly drawn from the distribution of human players. Similarly, we calculate the expected payoff of the human players when randomly paired with another human player. The results are presented in Table S1.

Table S1. ChatGPT-4's strategies yield higher partner payoffs and higher combined payoffs compared to human players in all games where the payoff is not constant. ChatGPT-3 is the most cooperative among the three in the Public Goods game and the most altruistic in two games (Public Goods game and Trust game as the banker). Expected payoffs are calculated by sampling the partner's action from the human player distribution. The grey numbers after the "±" symbols are the standard errors based on 30 samples.

Game	Player	Selfish (own) payoff	Selfless (partner) payoff	Combined Payoff
Dictator	Human	\$74.14 ± 0.19	\$25.68 ± 0.19	\$100.00
	ChatGPT-4	\$50.00 ± 0.00	\$50.00 ± 0.00	\$100.00
	ChatGPT-3	\$64.83 ± 2.47	\$35.17 ± 2.47	\$100.00
Ultimatum - Proposer	Human	\$33.51 ± 0.16	\$35.19 ± 0.30	\$68.70 ± 0.36
	ChatGPT-4	\$45.98 ± 0.00	\$45.98 ± 0.00	\$91.96 ± 0.00
	ChatGPT-3	\$35.10 ± 0.90	\$32.79 ± 3.29	\$67.89 ± 3.49
Ultimatum - Responder	Human	\$35.19 ± 0.13	\$33.51 ± 0.20	\$68.70 ± 0.32
	ChatGPT-4	\$37.60 ± 1.46	\$39.60 ± 2.96	\$77.20 ± 4.41
	ChatGPT-3	\$38.26 ± 1.40	\$36.75 ± 2.40	\$75.00 ± 4.41
Trust - Investor	Human	\$111.33 ± 0.11	\$76.03 ± 0.38	\$187.36 ± 0.48
	ChatGPT-4	\$108.01 ± 0.48	\$84.13 ± 2.09	\$192.14 ± 2.54
	ChatGPT-3	\$104.63 ± 0.84	\$68.04 ± 3.77	\$172.67 ± 4.57
Trust - Banker*	Human	\$90.79 ± 0.97	\$109.21 ± 0.97	\$200.00
	ChatGPT-4	\$60.83 ± 2.26	\$139.17 ± 2.26	\$200.00
	ChatGPT-3	\$37.50 ± 5.99	\$162.50 ± 5.99	\$200.00
Public Goods	Human	\$9.04 ± 0.02	\$9.04 ± 0.02	\$36.15 ± 0.04
	ChatGPT-4	\$8.39 ± 0.05	\$9.69 ± 0.05	\$37.45 ± 0.10
	ChatGPT-3	\$7.97 ± 0.13	\$10.10 ± 0.13	\$38.28 ± 0.27
Prisoner's Dilemma†	Human	\$345.12 ± 0.53	\$345.12 ± 0.70	\$690.24 ± 0.18
	ChatGPT-4	\$205.48 ±10.70	\$531.31 ±14.27	\$736.79 ± 3.57
	ChatGPT-3	\$250.48 ±16.38	\$471.31 ±21.84	\$721.79 ± 5.46

* : To be comparable, the Trust-Banker calculations are done assuming that the original investment is \$50.
† : The Prisoner's Dilemma reports the payoffs in the first round of the game.

Table S1 shows that ChatGPT-4 outperforms human players in terms of expected own payoff only in the Ultimatum Game, and a lower own payoff in the other games involving trust or cooperation. However, in all seven game scenarios, it obtains a higher expected payoff for its partner. Moreover, it achieves a higher combined payoff for both players in five out of seven games, the exception being the Dictator game and the Trust game as the banker (where the combined payoff is constant).

These findings are consistent with an increased level of altruism and cooperation compared to the human player distribution. On the other hand, ChatGPT-3 obtains payoffs that are closer to humans in the Dictator game, the Ultimatum Game as the proposer, and Prisoner's Dilemma. And, although it yields a lower own payoff in the Trust games and the Public Goods game compared to ChatGPT-4, it achieves an even higher partner payoff and combined payoff in the Public Goods game and a higher partner payoff as the banker in the Trust game.

B. Optimization Objective. The findings presented above indicate that the strategic outputs of ChatGPT-4 and ChatGPT-3 models tend to yield higher partner payoffs compared to human players, with ChatGPT-4 frequently attaining the highest combined payoff. This subsection is dedicated to a systematic exploration of the preferences that best predict behavior.

Discerning the intrinsic objectives of models can be challenging when solely examining their training methodologies. Taking ChatGPT-4 (8) as an example, which serves as the backbone of ChatGPT-4, it is a Transformer-style model initially pre-trained to forecast the subsequent token in a given document. This pre-training phase leverages publicly accessible data such as internet-sourced information. In this stage, the primary objective function is essentially to maximize the likelihood of a word's occurrence when provided with the preceding words for each sentence or document within the training data. Subsequently, the model undergoes a fine-tuning process using Reinforcement Learning from Human Feedback (RLHF) (9). The fine-tuning tasks employed encompass natural language processing activities such as text generation, question answering, dialog generation, summarization, and information extraction. The process involves presenting demonstration responses and having human evaluators rank the outputs from best to worst. With RLHF, OpenAI may also add restrictions regarding safety considerations[§§]. Notably, these tasks and the safety policies do not inherently align with decision-making tasks, and this approach does not directly translate into a well-defined objective function for behavioral games. To the best of our knowledge, there is no evidence

[§§] OpenAI usage policies: https://openai.com/policies/usage-policies, retrieved 03/2023.

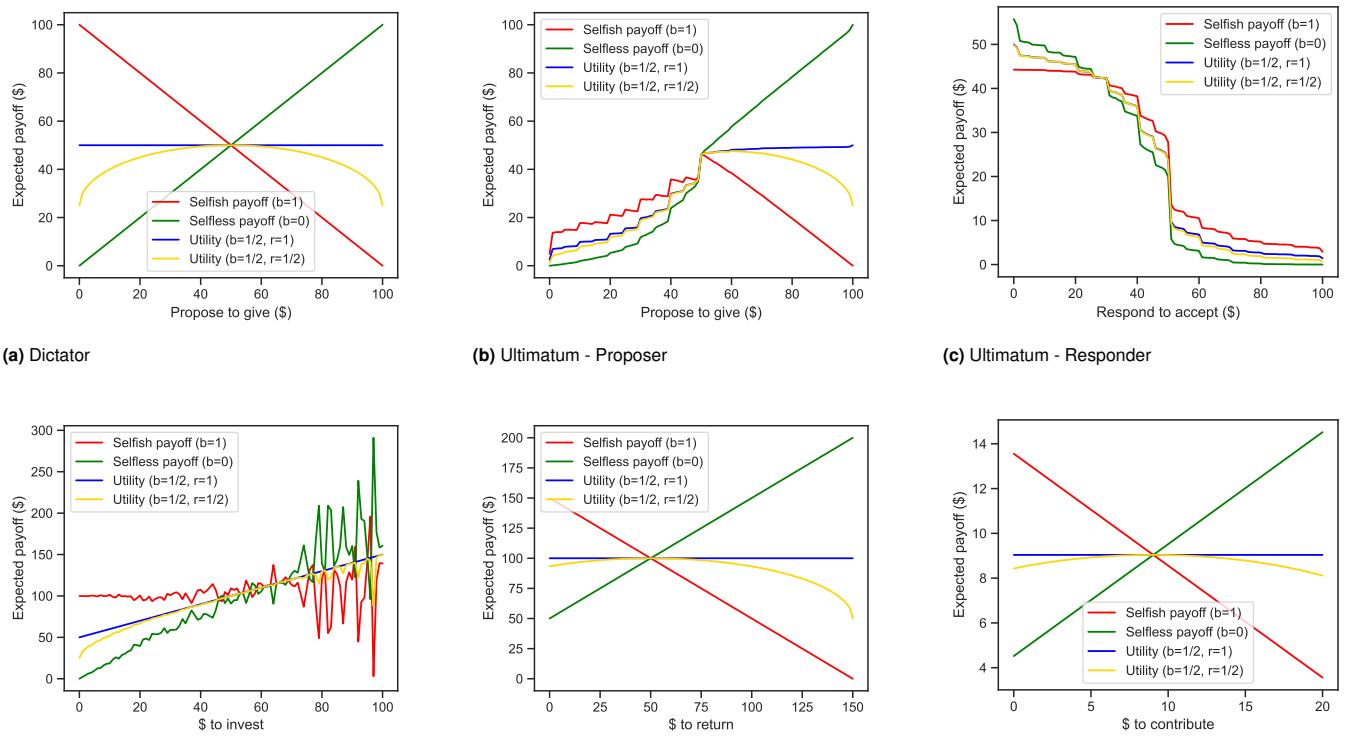
that either ChatGPT-3 or ChatGPT-4 was fine-tuned to behave in specific directions in the tests included in our study. Hence, we rely on observations of models' behaviors to understand their objective.

In the context of behavioral games, where the player is presumed to optimize a blend of selfish and partner payoffs, the optimization objective function can be formulated as a constant elasticity of substitution (CES) utility function (10):

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

where U_b denots the utility function, S is the player's selfish (own) payoff, P is the selfless (partner) payoff, $b \in [0,1]$ is the weight, and r is the CES parameter.

The selfish and selfless expected payoff curves are plotted in Fig. S3, which also contains examples of the expected utility function when $b = 1/2$ for two CES specifications (the linear specification with $r = 1$ and the non-linear specification with $r = 1/2$). These are the expected utilities against the distribution of play from the human partner distribution, as a function of the strategy played.



(a) Dictator **(b)** Ultimatum - Proposer **(c)** Ultimatum - Responder

(d) Trust - Investor **(e)** Trust - Banker (with $50 invested) **(f)** Public Goods

Fig. S3. Expected selfish (own) payoff (red lines) and selfless (partner) payoff (green lines) of every single action with a randomly sampled human partner. Blue and yellow lines show the weighted expected utility function examples as defined in Eq. 1. When $r = 1$ and $b = 1/2$ (blue lines), the utility function becomes the overall welfare (average payoff). Weighted utility functions for other values of b can be obtained similarly.

For every game and parameter value b ($0 \leq b \leq 1$) for the utility function, we can calculate the mean square error (MSE) for any given action compared to the best response as

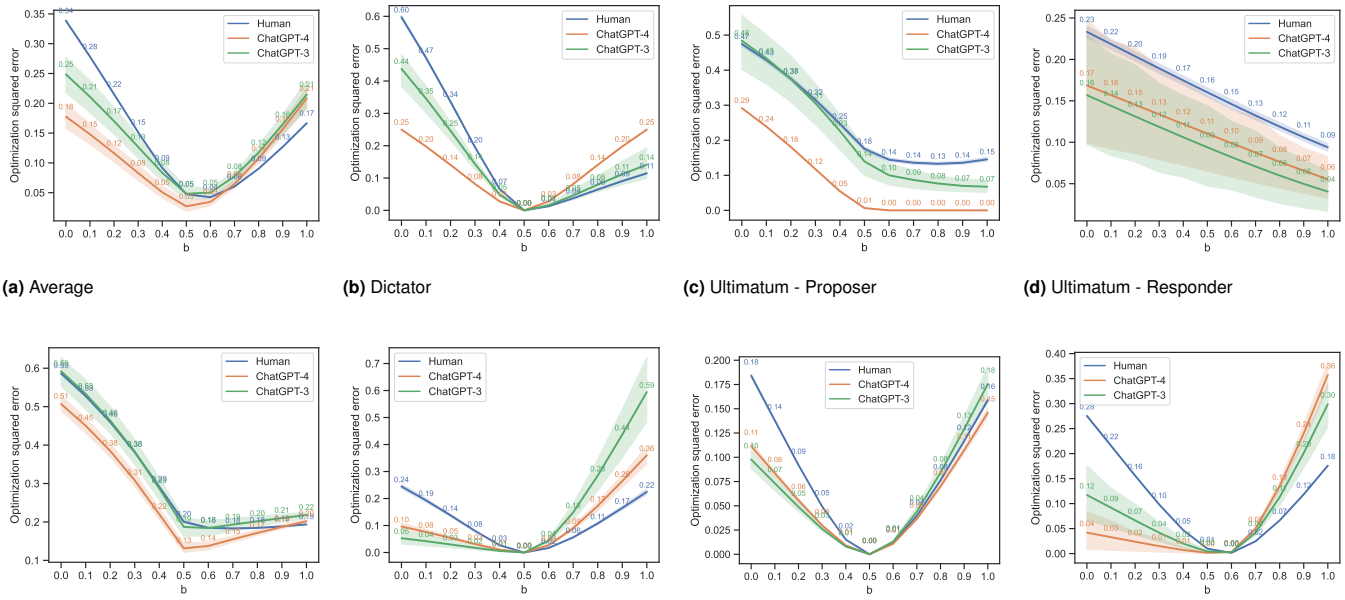
$$\text{MSE}_b = \frac{1}{|\mathcal{O}|} \sum_{k \in \mathcal{O}} \left[1 - \frac{U_b(k)}{U_b^*}\right]^2, \quad [2]$$

where $\mathcal{O}$ is the set of observations, k is an action choice from the observation (e.g., give $50 in the Dictator game), $U_b(k)$ is the expected utility from action k calculated with the expected selfish payoff $S(k)$ and the partner payoff $P(k)$, and U_b^* is the theoretical maximum utility from the best response (i.e., based on Fig. S3).

The results are shown in Fig. S4 (linear CES utility specification with $r = 1$) and Fig. S5 (non-linear CES specification with $r = 1/2$).

For the linear specification and most of the games except for Ultimatum, human players and ChatGPTs achieve their lowest optimization MSE at around $b = 1/2$. We also note that ChatGPTs tend to have smaller optimization errors compared to humans when $b \leq 0.5$, showing higher optimization efficiency when the objective is less selfish.

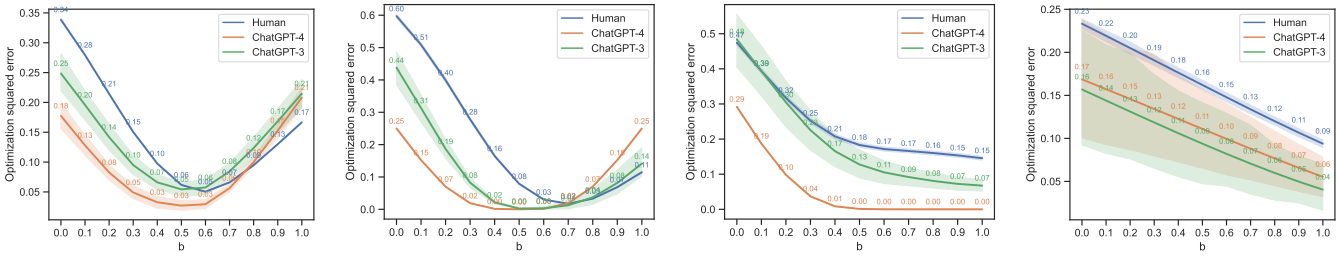
Beyond the revealed preference analysis above, we also estimate the parameter b for each game and model, positing a logistic multinomial model framework (as in McFadden's discrete choice problem (11)). We only do this for the linear utility specification, as that is the usual multinomial logit formulation.



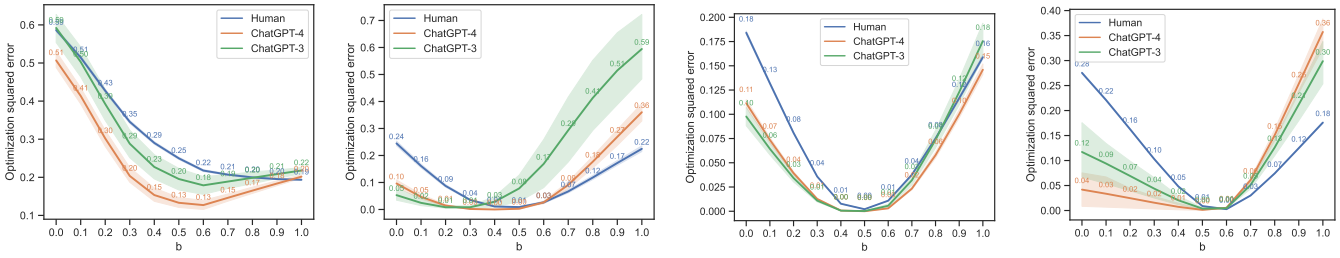
(a) Average **(b)** Dictator **(c)** Ultimatum - Proposer **(d)** Ultimatum - Responder

(e) Trust - Investor **(f)** Trust - Banker **(g)** Public Goods **(h)** Prisoner's Dilemma

Fig. S4. Mean squared error of the actual distribution of play relative to the best-response payoff, when matched with a partner playing the human distribution. The average is across all games. The errors are reported for each possible b, which is the weight on own vs partner payoff in the utility function (linear blend, with CES specification $r = 1$). $b = 1$ is the purely selfish (own) payoff, $b = 0$ is the purely selfless (partner) payoff, and $b = 1/2$ is the overall welfare (average) payoff. The values of mean square errors are annotated in the plots.



(a) Average **(b)** Dictator **(c)** Ultimatum - Proposer **(d)** Ultimatum - Responder



(e) Trust - Investor **(f)** Trust - Banker **(g)** Public Goods **(h)** Prisoner's Dilemma

Fig. S5. Mean squared error of the actual distribution of play relative to the best-response payoff, when matched with a partner playing the human distribution. The average is across all games. The errors are reported for each possible b, which is the weight on own vs partner payoff in the utility function (non-linear blend, with CES specification $r = 1/2$). $b = 1$ is the purely selfish (own) payoff, and $b = 0$ is the purely selfless (partner) payoff. The values of mean square errors are annotated in the plots.

According to this framework, actions are sampled in accordance with the following probability distribution:

$$\Pr(k) = \frac{\exp(U_b(k))}{\sum_{j \le K} \exp(U_b(j))}, \quad [3]$$

where K is the number of possible action choices.

The estimation results are presented in Table S2, which are well aligned with those in Fig. S4. For many games, including Dictator, Trust - Banker, Public Goods (for only ChatGPT-3), and Prisoner's Dilemma, the estimated b values from ChatGPTs are significantly smaller than humans, indicating ChatGPTs behave as if they were less selfish behavioral than humans.

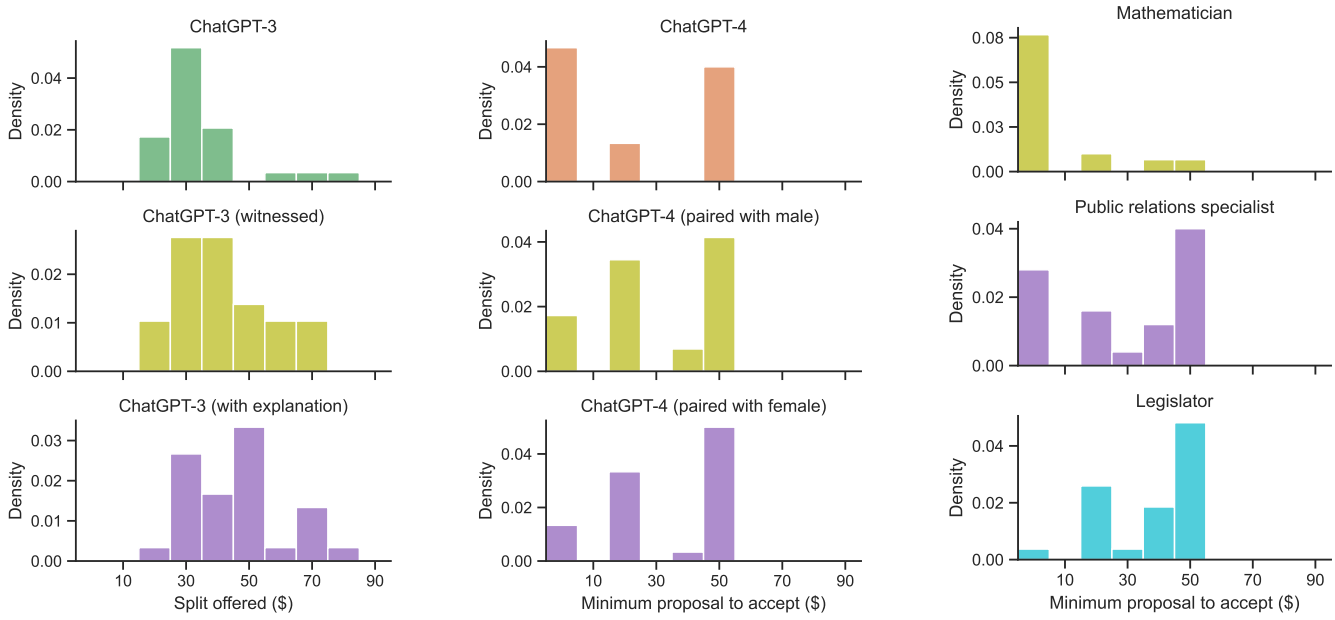
Table S2. Estimation of the weight b by multinomial logit discrete choice analysis. Green highlights the cases when the estimated b for ChatGPT models is significantly smaller (less weight on own payoff) than the estimate for humans.

Game	Player	With CES specification $r = 1$			With CES specification $r = 1/2$		
		Estimated b	Standard error	Confidence interval	Estimated b	Standard error	Confidence interval
Dictator	Human	0.517	0.000	(0.516, 0.517)	0.658	0.000	(0.657, 0.659)
	ChatGPT-4	0.500	0.003	(0.494, 0.506)	0.500	0.009	(0.482, 0.518)
	ChatGPT-3	0.509	0.003	(0.502, 0.516)	0.582	0.010	(0.563, 0.601)
Ultimatum - Proposer	Human	1.000	0.005	(0.989, 1.011)	1.000	0.005	(0.990, 1.01)
	ChatGPT-4	1.000	0.076	(0.851, 1.149)	1.000	0.079	(0.845, 1.155)
	ChatGPT-3	1.000	0.076	(0.851, 1.149)	1.000	0.211	(0.587, 1.413)
Ultimatum - Responder	Human	1.000	0.005	(0.990, 1.010)	1.000	0.006	(0.989, 1.011)
	ChatGPT-4	1.000	0.070	(0.862, 1.138)	1.000	0.079	(0.845, 1.155)
	ChatGPT-3	1.000	0.070	(0.862, 1.138)	1.000	0.077	(0.849, 1.151)
Trust - Investor	Human	0.535	0.000	(0.535, 0.535)	0.570	0.000	(0.569, 0.570)
	ChatGPT-4	0.532	0.003	(0.526, 0.538)	0.566	0.003	(0.559, 0.572)
	ChatGPT-3	0.535	0.003	(0.529, 0.541)	0.569	0.003	(0.564, 0.575)
Trust - Banker*	Human	0.504	0.000	(0.504, 0.505)	0.475	0.001	(0.473, 0.477)
	ChatGPT-4	0.496	0.002	(0.492, 0.500)	0.395	0.007	(0.382, 0.408)
	ChatGPT-3	0.488	0.003	(0.482, 0.495)	0.300	0.009	(0.283, 0.318)
Public Goods	Human	0.526	0.001	(0.524, 0.528)	0.518	0.001	(0.516, 0.521)
	ChatGPT-4	0.491	0.021	(0.449, 0.533)	0.475	0.023	(0.430, 0.521)
	ChatGPT-3	0.468	0.022	(0.426, 0.510)	0.448	0.023	(0.402, 0.494)
Prisoner's Dilemma†	Human	0.572	0.000	(0.572, 0.572)	0.563	0.000	(0.563, 0.563)
	ChatGPT-4	0.568	0.001	(0.567, 0.569)	0.560	0.001	(0.558, 0.561)
	ChatGPT-3	0.570	0.000	(0.569, 0.571)	0.561	0.000	(0.560, 0.562)

* : To be comparable, the Trust-Banker calculations are done assuming that the original investment is \$50.
† : The Prisoner's Dilemma reports the estimation results in the first round of the game.

C. Framing. Similar to humans, ChatGPT's decisions can be significantly influenced by changes in the framing or context of the same strategic setting. A request for an explanation of its decision, or asking them to act as if they come from some specific occupation can have an impact. Fig. S6 presents selected examples of how different framings or contexts influence ChatGPT-4 and ChatGPT-3's behavior. Refer to Section 3 for detailed results.

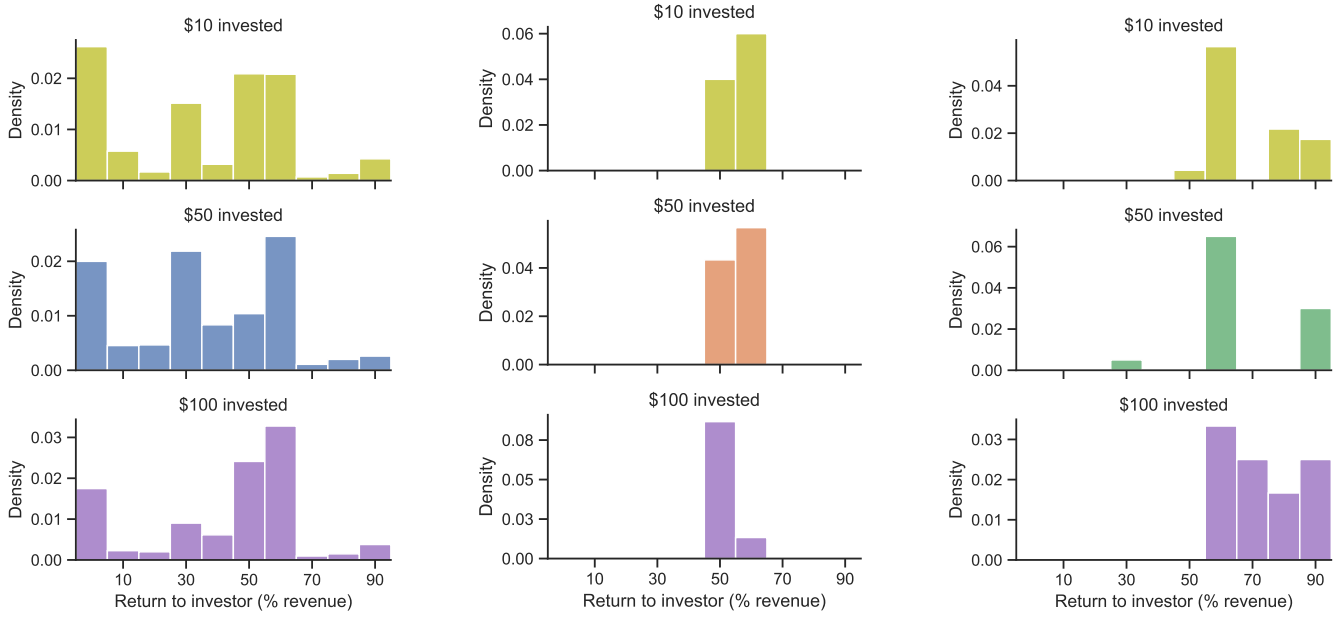
D. Learning. In games with multiple roles, both ChatGPT-3 and ChatGPT-4 change their behavior once they have experienced another role in the game. When ChatGPT-3 has previously acted as the responder in the Ultimatum game, it tends to propose a higher offer when it later takes the proposer role, while ChatGPT-4's proposal remains unchanged whether or not it has been a responder (Fig. S7a). When ChatGPT-4 has previously played the proposer, it tends to be willing to accept a smaller split as the responder (Fig. S7b). Being exposed to the banker's role in the Trust game influences ChatGPT-4 and ChatGPT-3's subsequent decisions as the investor, leading them to invest more (Fig. S7c). The distributions of their decisions become narrower. Having played the investor first also influences both chatbots' subsequent decisions as the banker, leading them to return more to the investor (Fig. S7d). Returning the principal plus half the profit becomes the single mode of ChatGPT-4's decision, while ChatGPT-3's decision is even more generous, returning more than 2/3 of the profit to the investor. Refer to Section 3 for detailed results.



(a) Dictator - Explanation required / Witnessed)

(b) Ultimatum - response to gendered proposers

(c) Ultimatum - as Responder when prompted with different occupations (ChatGPT-4)

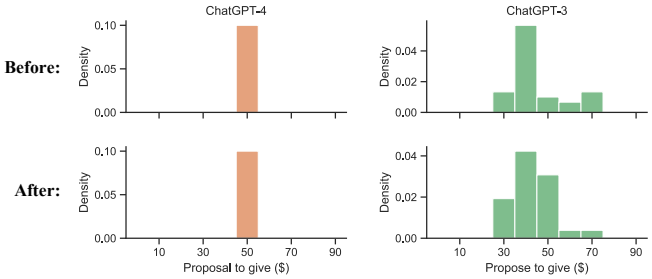


(d) Trust - Banker's strategy given different investment sizes (human)

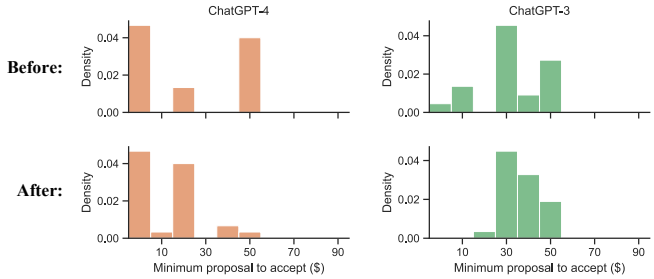
(e) Trust - Banker's strategy given different investment sizes (ChatGPT-4)

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

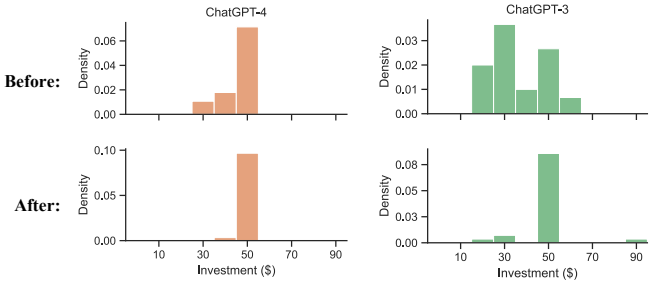
Fig. S6. ChatGPT's behavior as a function of the framing or context of the same strategic setting. (a) In the Dictator game, ChatGPT-3 makes a more generous split in the presence of a witness or when requested to explain its decision. (b) In the Ultimatum game, ChatGPT-4 accepts a higher split when the gender of the proposer is known (despite which gender). (c) When prompted to be a mathematician, ChatGPT-4 demands a smaller split as the responder in the Ultimatum game, and a larger and fairer split when prompted to be a public relations specialist or a legislator. (d-f) In the Trust game, when the size of the investment increases, ChatGPT-3 and humans tend to return a larger proportion as the banker to the investor, while ChatGPT-4 tends to return a smaller proportion when the investment increases to $100. Density is the normalized count such that the total area of the histogram equals 1.



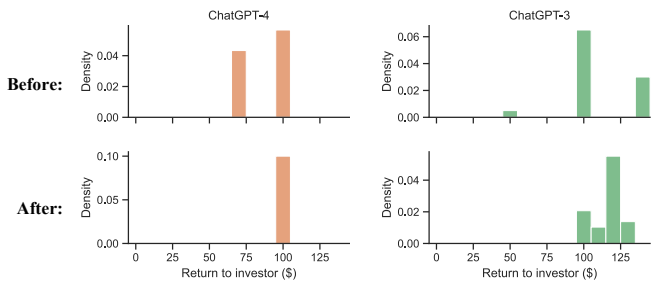
(a) Ultimatum: AI strategy as proposer before and after being responder.



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



(c) Trust: AI strategy as the investor before and after being the banker.



(d) Trust: AI strategy as the banker before and after being the investor.

Fig. S7. ChatGPT's behavior changes after being exposed to the other role in two-role games. Both ChatGPT-4 and ChatGPT-3 accept a smaller proposal in the Ultimatum game after being the proposer, make a larger investment in the Trust game after being the banker, and return a larger proportion to the investor after being the investor. ChatGPT-3 proposes a more generous split after being the responder in the Ultimatum game.

3. Detailed Results

A. OCEAN Big Five Test. We collect ChatGPT's responses to the OCEAN Big Five personality tests. We compare the five dimensions of personality traits of ChatGPT-3 and ChatGPT-4 (averaged over 30 independent runs) with those collected from human subjects. In particular, both models yield extraversion scores close to the human median (ChatGPT-4 at 53.4^{th} percentile and ChatGPT-3 at 49.4^{th} percentile of human distribution), agreeableness scores lower than the human median (ChatGPT-4 at 32.4^{th} percentile and ChatGPT-3 at 17.2^{th} percentile of human distribution), neuroticism scores moderately below human median (ChatGPT-4 at 41.3^{th} percentile and ChatGPT-3.5 at 45.4^{th} percentile of human distribution), and openness scores below human median (ChatGPT-4 at 37.9^{th} percentile and ChatGPT-3 at 5.0^{th} percentile of human distribution). For conscientiousness scores, ChatGPT-4's score is above the human median (at 62.7^{th} percentile of human distribution), while ChatGPT-3 is slightly lower than the median (at 47.1^{th} percentile of human distribution).

B. The Dictator Game. The Dictator Game involves two participants: one player, the "dictator," is given a sum of money and can choose to share any portion with the other player, who must accept whatever amount is offered. The game is often used to explore altruistic behavior and deviations from pure self-interest (12, 13).

In the game, ChatGPT is asked to decide how to split \$100 between itself and another player who unconditionally accepts the proposal. In 30 independent sessions, ChatGPT-3's decision follows a bell-shaped distribution peaked at giving a moderate amount of \$30 to the other player ($min = 20$, $max = 80$, $\mu = 35.2$, $\sigma = 13.5$). When explicitly requested to provide an explanation of the decision, the distribution shifts rightward, with a new mode of splitting evenly at \$50 ($min = 20$, $max = 80$, $\mu = 46.2$, $\sigma = 15.5$, $p = 0.002$, Wilcoxon Rank-Sum Test unless otherwise specified) (Fig. S6a).

Different from ChatGPT-3, when ChatGPT-4 is tested, it makes a consistent decision of an even split (\$50, Wilcoxon Ranked-Sum Test $p \ll 0.001$). The same behavior is observed for the Plus version (the paid Web-based version). From the responses, some keywords that stand out include *fair*, *equal*, and *equally*. Explicitly requesting an explanation does not shift ChatGPT-4's decision. The Free version (the unpaid Web-based version) behaves similarly to ChatGPT-3, with a slightly less spread distribution peaked on \$30 ($min = 20$, $max = 70$, $\mu = 33.8$, $\sigma = 10.7$, $p = 0.785$). Some notable keywords among the responses are *fair*, *reasonable*, *trust*, and *goals*.

When a witness/game host is present (see Section 1.A.3), ChatGPT-3's dictating decision becomes significantly more generous ($mode = 30$ and $40, min = 20, max = 70, \mu = 41.7, \sigma = 14.9, p = 0.046$) even though the host has no interference in the game outcome. The presence of the game host does not have an impact on ChatGPT-4's decision.

The "system" role is an instruction that sets a global context for all the prompts of ChatGPT in the same session. The default system role of ChatGPT is "a helpful assistant." When the system role is set into particular occupations (those considered to have a high exposure of ChatGPT in their workflow according to (6), there is a shift of ChatGPT-3's response to the dictator game. In particular, when playing the roles of a public relation specialist, a journalist, or an investment fund manager, ChatGPT-3's tend to spare a more generous portion to the other player (than as a helpful assistant). As a mathematician, ChatGPT-3 tends to offer a smaller share to the other player. This indicates that ChatGPT-3's interpretation of what is "fair" or "reasonable" is affected by the role it plays.

Playing one of the above roles does not change the centralized decision of ChatGPT-4.

C. The Ultimatum Game. In the Ultimatum Game, two participants split a sum of money (the 'pie'): one player, the "proposer," makes an offer on how to split the pie, and the second player, the "responder," can accept or reject this offer. If the responder accepts, the pie is split as proposed, but if the responder rejects, neither player receives anything. This game is often used to study fairness, social norms, and negotiation behavior (12).

In this game, ChatGPT is asked either to propose to divide \$100 between itself and a responder, or to respond to a proposal made by the other player. The difference between the Ultimatum and the Dictator game is that the responder no longer blindly accepts the proposal, and if they reject it, both players will get \$0 no matter what the original proposal was.

When playing the proposer alone, ChatGPT-3's proposal of the amount given to the other player follows a distribution with mode \$40 ($min = 30, max = 70, \mu = 45.2, \sigma = 12.1$). Compared with the Dictator game, ChatGPT-3 gives significantly more ($p < 0.001$) to the responder when the other party's decision jointly decides the outcome.

When specifically requested to provide an explanation of the decision, ChatGPT's decision does not present a significant shift. The most common decision is still \$40 ($min = 30, max = 70, \mu = 42.9, \sigma = 10.8, p = 0.54$). Like in the Dictator game, ChatGPT-4 makes a unanimous decision of an equal split (\$50, $p = 0.001$), similar to that of ChatGPT Plus, and ChatGPT Free's behavior is much more similar to that of ChatGPT-3 ($mode = 40, min = 30, max = 70, \mu = 47.3, \sigma = 16.8, p = 0.905$).

Different context also has an effect on ChatGPT's decision. When asked to explicitly explain the decision, ChatGPT-4 presents a slight variance in its decision ($\sigma = 2.6$). A similar variance is also observed when ChatGPT-4 is asked to play different roles: as a mathematician or an investment fund manager, it occasionally proposes a smaller split to the responder, and as a journalist, it occasionally proposes a greater split to the responder ($max = 100, \sigma = 9.5$). None of these variations is statistically significant. The impact of roles to ChatGPT-3 is greater: the median of proposal decreases as a mathematician, centralizes as a public relation specialist, and increases as an investment fund manager.

When ChatGPT is asked to play the responder, it is asked the lowest amount that it is willing to accept. ChatGPT-3's response still follows a bell-shaped distribution, with a mode of \$30 ($\mu = 32.5, \sigma = 14.7$), a minimum of \$1, and a maximum of \$50. The mode, mean, min, and max are all lower than the statistics when it plays the proposer. A similar pattern is observed from ChatGPT Free ($\mu = 30, \sigma = 11.9, p = 0.400$).

ChatGPT-4 presents a significantly different behavior, where the decisions follow a two-mode distribution that concentrates on two sides ($1 and $50). Requiring explanation does not change the two modes and only affects the distribution of the middle range. ChatGPT Plus presents an even more extreme trait, and in a dominating majority of sessions, it is willing to accept as low as $1 as the responder ($p \ll 0.001$). The choice of $1 aligns well with the rational decision of the game. Indeed, some keywords that stand out from the responses include *nothing*, *better (than)*, and *something*.

The context of different occupations also affects ChatGPT's decision in this game. As a mathematician, ChatGPT-4's decision concentrates on the rational choice ($1, $p \ll 0.001$); as a public relation specialist, a journalist, an investment fund manager, or a legislator, its decision becomes less bipolarated, and the median shifts towards the right (favoring fairness more than rationality), although not statistically significant (Fig. S6c). Similar shifts are observed for ChatGPT-3 when inquired in the context of corresponding occupations.

The above experiment queries ChatGPT in independent sessions, so its response to one question is not interfered by its decision or memory about the other question. An interesting question is whether their exposure/response to one question (corresponding to one role as proposer or responder) affects their decision for a follow-up question that corresponds to the other role. We expose ChatGPT to both roles in the same session, asking it to respond to one question and then respond to the other question. We find that being exposed to one scenario does influence ChatGPT-3's and ChatGPT-4's responses in the other scenario, compared with the results obtained from independent sessions. Having been the proposer first (Fig. S7b), the distribution of ChatGPT-3's response as the responder becomes narrower, with the $50 group shifting left towards $40. A similar shift presents in ChatGPT-4's response, where the $50 group moves towards $20 and the median has reduced to $15 ($p \ll 0.001$). Having been the responder first (Fig. S7a), ChatGPT-3's response as the proposer also becomes more generous, while ChatGPT-4's unanimous decision is not affected.

D. The Trust Game. The Trust Game is a two-player game that investigates trust and reciprocity. In this game, the first player (the investor– the trustor) is given a sum of money and can invest any amount in the second player (the banker– the trustee). The amount sent is multiplied by the game host, and the banker then decides how much, if any, to return to the investor. This game is used to study trust, reciprocity, and social norms (14).

In the game, ChatGPT is also asked to play one of two roles: as an investor or as a banker. As an investor, it is asked to decide how much (from $0 to $100) to invest in the banker (which is expected to generate a profit), who may return the entire revenue or nothing to the investor. ChatGPT-3's decision follows a distribution that peaks at a moderate value of $30 ($min = 20, max = 60, \mu = 36.3, \sigma = 12.7$). When explicitly asked to explain the decision, its decision does not present a significant shift but shows a wider spread ($min = 20, max = 100, \sigma = 16.0, p = 0.439$). Different occupation roles also show an effect, where ChatGPT-3 makes a larger investment as a mathematician, a public relation specialist, or a journalist.

ChatGPT-4 acts as if it has significantly more trust in the banker ($p = 0.006$). Its decision follows a distribution peaked at investing half ($50) of the endowment, which also appears to be the maximum investment it makes. It tends to invest slightly more when specifically required to provide an explanation. The distributions are also mildly affected by the assumed occupations, whereas for a mathematician or an investment fund manager, the distribution is more spread-out, and it even occasionally invests the entire endowment. The difference is not significant except for being a public relation specialist ($p = 0.05$).

ChatGPT Plus invests less than ChatGPT-3, with a distribution peaked at $10 ($min = 0, max = 100, \mu = 19.6, \sigma = 23.0, p \ll 0.001$). The keywords standing out from the responses include *risk*, *return*, *desires*, *losing*, *minimizes*, *guarantee*, and *control*. Once again, ChatGPT Free's decision distribution is more similar to that of ChatGPT-3 ($mode = 30, min = 10, max = 60, \mu = 36.7, \sigma = 12.1, p = 0.792$).

In the second scenario, ChatGPT is asked to play the banker and decides what proportion of the total revenue is returned to the investor, which could range all the way from nothing ($0) to the entire revenue (original investment + 2x profit). We find that the two common strategies that ChatGPT-4 uses are to return 1) the original investment plus half of the profit, or 2) half of the revenue. The former returns more to the investor than the latter. For ChatGPT-4, the former is more frequently used, and requiring the explanation makes this strategy even more dominating. Among the assumed occupations, ChatGPT-4 returns more generously to the investor as a public relation specialist or an investment fund manager.

When the size of the investment increases, there is a shift in ChatGPT-4's decision. When the investment was $100 (out of $100), the second and less generous strategy (returning half revenue) becomes the new dominant (Fig. S6e).

ChatGPT-3's decision centers on the first strategy, with small probabilities of giving more generous (even the entire revenue) returns. This strategy is robust to the size of the investment. When the investment was $10, $50, and $100, ChatGPT-3's decision follows a distribution peaked at $20 ($\mu = 22.5, \sigma = 4.3$), $100 ($\mu = 112.5, \sigma = 27.5$), and $200 ($\mu = 239.2, \sigma = 40.7$) (Fig. S6f). Requesting an explanation makes the distribution even more centralized, but the extremely generous behavior (returning the whole revenue to the investor) does not vanish. Being a public relation specialist, a journalist, or an investment fund manager, ChatGPT-3 tends to pay back more generously than in the default role.

Being exposed to a different role influences ChatGPT's decision in the other role. When inquired about both scenarios (investor and banker) in the same session, ChatGPT's responses for the second role deviate from the distribution observed in the independent sessions. In particular, after playing the role of the banker, both ChatGPT-3's and ChatGPT-4's investments increased. After playing the role of the investor, both ChatGPT-3's and ChatGPT-4's decisions as the banker become more generous.

E. The Bomb Risk Game. In the Bomb Risk Game, a player decides how many out of 100 boxes to collect, one of which contains a 'bomb.' Earnings increase linearly with the number of boxes a player decides to open, but drop to zero if the 'bomb' is hit. The game is designed to measure risk attitudes (15).

In this game, ChatGPT is asked to open a number of boxes out of 100, among which a bomb is randomly placed. If the box that contains the bomb is not opened, then the player earns points that equal the number of boxes they open. If the box that contains the bomb is opened, the bomb explodes and the player gets zero points. In every new round, a new set of 100 boxes is provided and the bomb is placed at random. Opening 50 boxes is where the expectation of points gain is maximized, and the decision at each round should be independent of the decisions/results in previous rounds.

When first exposed to the game, both ChatGPT-3's and ChatGPT-4's decisions peak at opening 50 boxes (ChatGPT-3: $\mu = 39.8, \sigma = 16.3$, ChatGPT-4: $\mu = 57.6, \sigma = 17.4$), demonstrating a rational and risk neural pattern of behavior. Despite the mode, the distribution of ChatGPT-4 is significantly more risk-loving than that of ChatGPT-3 ($p < 0.001$), with 13.8% of the chance of opening a maximum number (99) of boxes. When playing the game for multiple rounds, its decision in the following round is influenced by the outcome of the previous rounds. As long as the bomb does not explode in the previous round, ChatGPT-3's decision in the second round becomes even more concentrated on 50 boxes ($\mu = 48.8, \sigma = 7.7$). When the bomb does explode in the first round, ChatGPT-3 tends to be more conservative in the second round, with a new mode of opening 25 boxes ($\mu = 26.4, \sigma = 10.8$). If the bomb explodes again in the second round, it decides to open even fewer boxes ($mode = 10, \mu = 14.8, \sigma = 11.0$) in the third round. If the bomb exploded in the second round but not in the first round, the average number of boxes opened in the third round drops from 50 to 43.1 even though the mode is still 50 ($mode = 50, \sigma = 25.0$).

F. Public Goods. In the Public Goods game, each participant is given an initial endowment and can contribute any portion to a public good project. The total amount raised for the project is then multiplied by a factor and distributed equally among all participants, regardless of their individual contributions. This setup allows for the exploration of altruistic contributions, the free-rider problem, and social dilemmas (16).

In the game, ChatGPT is asked to decide how much money from 0-$20 to contribute to a public project as one of four participants, the personal payoff of which is the sum of the amount not invested and 50% of the group contribution. It therefore makes a profit when the contribution from the rest of the group is greater than its contribution. It is the most common for both ChatGPT-3 and ChatGPT-4 to invest half of the endowment into the public goods project. When ChatGPT-3 receives a greater payoff than the original endowment, it tends to increase its contribution in the next rounds despite whether the other players made a larger or smaller contribution (or received a higher payoff). ChatGPT-4 on the other hand, makes a consistent contribution in most sessions despite the contribution of other players. In sessions where its contributions do increase over rounds, the increases tend to be smaller than those of ChatGPT-3.

G. Prisoner's Dilemma. The Prisoner's Dilemma is a fundamental game in game theory and behavioral economics, illustrating the tension between individual and collective rationality. The game highlights that individuals, acting in their own self-interest, can end up with worse outcomes than if they had cooperated (17–20).

In our version from MobLab, the framing is as follows. Two players are separately deciding whether to play the 'push' (cooperate) or 'pull' (defect) card. If both push, they collectively earn a higher payoff; if one pulls and the other pushes, the defector gets all the payoff; if both pull, they both receive less payment. The payoff matrix is displayed in Table S3.

Table S3. Payoff matrix of the Prisoner's Dilemma game. The first numbers in cells are the payoffs of Player A, and the second numbers are the payoffs of Player B.

		Player B	
		Push	**Pull**
Player A	**Push**	\$400, \$400	\$0, \$700
	Pull	\$700, \$0	\$300, \$300

In the five-round game (Fig. S8, for the first round, ChatGPT attempts to cooperate with the other player in a majority of the sessions (21/30 for ChatGPT-3, and 29/30 for ChatGPT-4). When the other player chooses to defect, however, the majority of its decisions in the next round quickly turn into "Pull" (defect), consistent with the 'tit for tat' pattern. Such a "punishing" strategy does not last if the other player continues to play "Pull." Instead, the ratio of "Push" bounces back to 25/30 for ChatGPT-4 and 14/30 for ChatGPT-3, once again trying to incentivize the other player to coordinate.

When the other player plays two "Push" cards (chooses to cooperate) in a row in the third and fourth rounds, both ChatGPT-3's and ChatGPT-4's decisions become relatively stable and end up cooperating in 14/30 of the instances (ChatGPT-3) and 25/30 of the instances (ChatGPT-4) after all five rounds.

References

1. Goldberg LR (1993) The structure of phenotypic personality traits. *American psychologist* 48(1):26.
2. McCrae RR, Costa Jr PT (1997) Personality trait structure as a human universal. *American psychologist* 52(5):509.
3. Roccas S, Sagiv L, Schwartz SH, Knafo A (2002) The big five personality factors and personal values. *Personality and social psychology bulletin* 28(6):789–801.
4. McCrae RR, Costa Jr PT (2008) The five-factor theory of personality. *Theoretical Perspectives*.

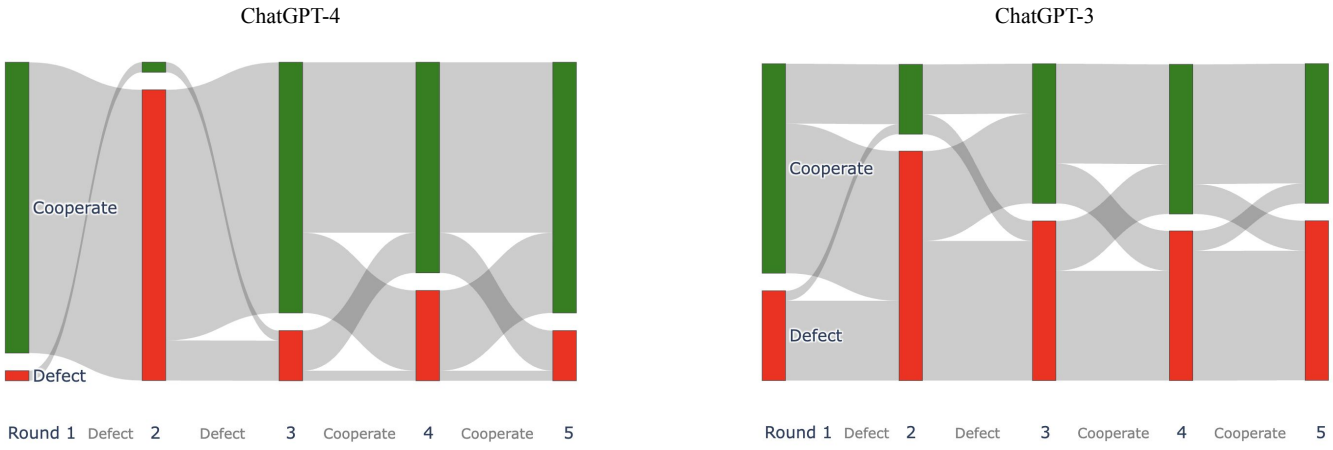


Fig. S8. Five-round Prisoner's Dilemma Game. A large proportion of AI decisions switch from cooperation to defection if the other player defects in the first round. However, a significant portion reverts back to cooperation in the third round even if the other player continues to defect. The proportion of cooperation becomes relatively stable in the following rounds. We executed 5 rounds in total. Partner actions in each round (1-4) are observed after the player's action and recorded below each panel.

5. Goldberg LR (1992) The development of markers for the big-five factor structure. *Psychological assessment* 4(1):26.
6. Eloundou T, Manning S, Mishkin P, Rock D (2023) Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.
7. Lin PH, et al. (2020) Evidence of general economic principles of bargaining and trade from 2,000 classroom experiments. *Nature Human Behaviour* 4(9):917–927.
8. OpenAI (2023) Gpt-4 technical report.
9. Ouyang L, et al. (2022) Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35:27730–27744.
10. McFadden D (1963) Constant elasticity of substitution production functions. *The Review of Economic Studies* 30(2):73–83.
11. McFadden D, , et al. (1973) Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*.
12. Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization* 3(4):367–388.
13. Forsythe R, Horowitz JL, Savin NE, Sefton M (1994) Fairness in simple bargaining experiments. *Games and Economic behavior* 6(3):347–369.
14. Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. *Games and economic behavior* 10(1):122–142.
15. Crosetto P, Filippin A (2013) The "bomb" risk elicitation task. *Journal of risk and uncertainty* 47:31–65.
16. Andreoni J (1995) Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review* pp. 891–904.
17. Von Neumann J, Morgenstern O (1947) *Theory of games and economic behavior, 2nd rev.* (Princeton university press).
18. Schelling TC (1958) The strategy of conflict. prospectus for a reorientation of game theory. *Journal of Conflict Resolution* 2(3):203–264.
19. Rapoport A, Chammah AM (1965) *Prisoner's dilemma: A study in conflict and cooperation*. (University of Michigan press) Vol. 165.
20. Andreoni J, Varian H (1999) Preplay contracting in the prisoners' dilemma. *Proceedings of the National Academy of Sciences* 96(19):10933–10938.