






# The emergence of economic rationality of GPT

Yiting Chen<sup>a,1,2</sup> , Tracy Xiao Liu<sup>b,1,2</sup>, You Shan<sup>c,1,2</sup> , and Songfa Zhong<sup>d,e,1,2</sup> 

Edited by Jose Scheinkman, Columbia University, New York, NY; received September 22, 2023; accepted November 13, 2023

As large language models (LLMs) like GPT become increasingly prevalent, it is essential that we assess their capabilities beyond language processing. This paper examines the economic rationality of GPT by instructing it to make budgetary decisions in four domains: risk, time, social, and food preferences. We measure economic rationality by assessing the consistency of GPT's decisions with utility maximization in classic revealed preference theory. We find that GPT's decisions are largely rational in each domain and demonstrate higher rationality score than those of human subjects in a parallel experiment and in the literature. Moreover, the estimated preference parameters of GPT are slightly different from human subjects and exhibit a lower degree of heterogeneity. We also find that the rationality scores are robust to the degree of randomness and demographic settings such as age and gender but are sensitive to contexts based on the language frames of the choice situations. These results suggest the potential of LLMs to make good decisions and the need to further understand their capabilities, limitations, and underlying mechanisms.

economic rationality | large language models | revealed preference analysis | decision-making

ChatGPT is a sophisticated chatbot application developed by OpenAI, which employs the state-of-the-art Generative Pre-trained Transformer model (hereafter referred to as “GPT”). As one of the most representative examples of large language models (LLMs), GPT uses transformer architecture and deep learning techniques to learn from vast web-based text corpora that contain 175 billion parameters (1, 2). Thanks to its massive volume of training data, GPT can generate human-like text with remarkable accuracy and fluency, to the extent that human evaluators find it difficult to distinguish GPT output from text written by humans (2). In addition to their natural language-generation capabilities, LLMs have demonstrated impressive abilities in a wide range of domains. For instance, they can generate computer code (3), engage in human-like conversations on various topics (4), solve university-level math problems (5), exhibit theory of mind ability (6), and possess psychological characteristics similar to humans (7, 8). LLMs have also shown their aptitude in performing high-level reasoning tasks (9). The impressive capabilities of LLMs reveal their remarkable potential, which can be likened to the emergence of a new species: “Homo silicus” (10). Because these achievements signify a major milestone in the development of LLMs, it is important that we understand how GPT performs in various high-level reasoning tasks.

Here, we present a study on the economic rationality of GPT. Rationality has been central to the methodological debate throughout various disciplines and is the fundamental assumption in economics and related social sciences. Here, we use a classic notion of economic rationality in revealed preference analysis that captures the extent to which a decision maker maximizes some well-behaved utility functions for the given budget constraints (11–17). Prior studies have computed rationality score based on choice data in risky, intertemporal, and social decision-making in laboratory environments (18–25) as well as expenditure data from survey and grocery stores in the field (26–29). Economic rationality has also been measured in children (23, 30), monkeys (31), rats, and pigeons (32). Moreover, it has been proposed as a measure of decision-making quality and linked to a wide range of economic outcomes, such as occupation, income, and wealth differences across individuals, and development gaps across countries (22, 33–39). Nevertheless, the rationality of GPT remains unexplored.

We instruct GPT to act as a decision maker to make budgetary decisions in choice environments with varying characteristics. The basic framework contains 25 decision tasks to allocate 100 points between two commodities with different prices, which is commonly used in experimental economics. The rationality of GPT is measured by the consistency of these 25 decisions with the generalized axiom of revealed preference (GARP), a necessary and sufficient condition under which a set of decisions are in accordance with utility maximization (11, 12, 14, 15). Therefore, a rationality score is derived from each group of 25 tasks within a given environment. Building on

## Significance

It is increasingly important to examine the capacity of large language models like Generative Pre-trained Transformer model (GPT) beyond language processing. We instruct GPT to make risk, time, social, and food decisions and measure how rational these decisions are. We show that GPT's decisions are mostly rational and even score higher than human decisions. The performance is affected by the way questions are framed, but not by settings of demographic information and randomness. Moreover, the estimated preference parameters of GPT, compared to those of human subjects, are slightly different and exhibit a substantially higher degree of homogeneity. Overall, these findings suggest that GPT could have the potential in assisting human decision-making, but more research is needed to fully assess their performance and underpinnings.

Author contributions: Y.C., T.X.L., Y.S., and S.Z. designed research; performed research; analyzed data; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

<sup>1</sup>Y.C., T.X.L., Y.S., and S.Z. contributed equally to this work.

<sup>2</sup>To whom correspondence may be addressed. Email: yitingchen26@gmail.com, liuxiao@sem.tsinghua.edu.cn, shany19@mails.tsinghua.edu.cn, or zhongsongfa@gmail.com.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2316205120/-/DCSupplemental>.

Published December 12, 2023.

this framework, we construct four environments by specifying the nature of the two commodities—two risky assets; two rewards with one for now and 1 for 1 mo later; two payments with one for the decision maker and one for another randomly paired subject; and two types of food with meat and tomatoes. Each environment is repeated 100 times, which generates 10,000 tasks for GPT. This design allows us to systematically measure GPT’s rationality in different choice domains. Moreover, we incorporate a series of variations in the randomness of GPT, the framing of decision tasks (40), the structure of the choice format (41), and the demographic settings of GPT (42). In order to compare the economic rationality between GPT and humans, we conduct a parallel experiment with 347 human subjects from a representative US sample.

We find that GPT demonstrates a high level of rationality in all four decision-making tasks concerning risk, time, social, and food, and it outperforms human subjects in the rationality score documented in both our human subject experiment and those reported in the literature. Furthermore, we find that GPT’s rationality scores are consistent across different demographic characteristics and invariant to the specification of the randomness of GPT. However, the level of rationality drops significantly when we employ a different price framing and when we use a discrete choice setting. These findings suggest that GPT obtains high rationality score but has some potential limitations in its decision-making abilities. Moreover, we estimate the preference parameters of GPT and human subjects. We find that the estimated preference parameters of GPT have some minor distinctions from human subjects and show a substantially higher degree of homogeneity.

Taken together, we use tools from revealed preference analysis and experimental economics to study increasingly capable artificial agents. There is growing interest in understanding these agents’ behavior (43) and ongoing debate about their performance compared with humans (44). Even though these artificial agents exhibit surprisingly excellent performance on many cognitive tests, some have expressed concern that such models are still far from achieving human-level understanding of language and semantics and exhibit considerable levels of behavioral bias (45, 46). We contribute to the understanding of the capacities and caveats of LLMs, by demonstrating that LLMs can act as if they are rational decision makers. We observe that rationality decreases when alternative price framing or discrete choice are used. This is line with some studies which show that GPT response can be highly sensitive to contexts (7, 10, 45, 47–49). Our study also highlights the need for more investigation and refinement of its decision-making mechanisms to ensure reliable and effective decision-making in various domains.

## Methods

We examine GPT’s decision-making in different environments using the public OpenAI application programming interface (API). Multiple GPT variants are accessible through this API. For our exercise, we focus on the GPT-3.5-Turbo, which powers ChatGPT and is the most popular, stable, and cost-effective model in the GPT family. We use APIs with Python instead of ChatGPT, since APIs enable us to adjust the parameters of the model and conduct massive experiments in an efficient manner.

Below, we describe how we ask GPT to “make decisions” by introducing the construction of prompts through which GPT returns a text in response to an input text. We then outline

multiple variations of our design to examine the robustness of our results.

### Design of the Baseline Condition.

**Instruct GPT to “make decisions”.** Each input prompt in GPT-3.5-turbo includes the specifications of a role (system, assistant, or user) and corresponding contents. We instruct GPT to make decisions in three steps. First, we specify the system’s role as “a human decision maker” and notify the system that “you should use your best judgment to come up with solutions that you like most”. Second, we explain the role of assistant with respect to the decision format: selecting a bundle of commodities from a standard budget line with varying prices, which will be explained in detail later, without requesting responses for any decision. This assists in storing information about the tasks. Afterward, we assign a series of decision-making tasks to the role of user in order to ask GPT to make decisions.

Moreover, to confirm that GPT has understood the instructions, we ask three testing questions, in which we either directly ask it to recall the decision format or ask about the consequence of certain decision scenarios. For each question, we simulate 25 times and GPT constantly provides correct answers. This confirms that GPT understands the decision environment. Detailed prompts to instruct GPT and obtain GPT responses are provided in *SI Appendix*.

**Decision task.** GPT decision tasks follow a typical budgetary experiment, in which a decision maker (DM) is endowed with 100 points to select a bundle of commodities, commodity  $A$  and commodity  $B$ . The prices of the two commodities are based on different exchange rates between points and payoffs. Thus, a decision  $i$  obtains a tuple  $(p^i, x^i)$  whereby a DM selects a bundle  $(x_A^i, x_B^i)$  under the prices  $(p_A^i, p_B^i)$ . Since measuring rationality requires a collection of such decisions, we include 25 tasks with randomly generated prices (22). After that, we measure the economic rationality of these 25 decisions  $(p^i, x^i)_{i=1}^{25}$ , based on the extent to which there exist some well-behaved utility functions to rationalize them.

To measure rationality across different preference domains, we vary the commodities in the decision tasks. In the first domain, the two commodities are specified as two contingent securities, in which the decisions capture the DM’s risk preference (21). In the second domain, the two commodities are rewards for today and 1 mo later, which are designed to examine the DM’s time preference (19). In the third domain, the two commodities are payoffs for the DM and another randomly matched subject, and thus the allocation captures the DM’s social preference (18, 25). Finally, in the fourth domain, the two commodities are the amount of meat and tomatoes, which captures the DM’s food preference (23).

We incorporate four preference domains of decisions, each consisting of 25 tasks. To examine GPT’s consistency in behavior, we simulate this process 100 times, resulting in 10,000 tasks for GPT. We refer to these 10,000 tasks, the 100 GPT observations in each preferences domain, as the baseline condition. A detailed description of tasks and parameters for prices are provided in *SI Appendix*. We set the temperature parameter to 0 (see the explanation below) and keep the default values for all other parameters.

**Design of Conditions with Variations.** To enrich our understanding of GPT’s economic rationality, based on the baseline condition, we introduce variations in the temperature and the

decision tasks. We also include demographic information in the text of the prompt as explained below.

**Variations of temperature.** Temperature plays a critical role in regulating the level of stochasticity and creativity in the responses generated by GPT (50). It ranges from 0 to 1, with a higher number indicating higher randomness. We set the temperature to be 0 in the baseline condition, in which the model gives deterministic answers (7, 9, 10). Some studies on GPT incorporate the variation in temperature to investigate the impact of randomness in creating text (3, 51). Following their practice, we conduct two additional sets of conditions, with the parameter set to be 0.5 and 1.

**Variations of decision task.** We design two variations of decision tasks to change the framing of prices and to switch from continuous to discrete choice, respectively. A detailed description is provided in *SI Appendix*.

In the baseline condition, we use “1 point = X units of commodity” to present price information, which is used in many existing experiments with human subjects (18, 19, 33, 35). In the price framing condition, we change it to “Y points = 1 unit of commodity”, which is an alternative framing used in the experimental literature (52). Since the budget sets remain constant, this allows us to examine whether framing affects the rationality of GPT.

In the baseline condition, the DM makes choices under the continuous budget sets. In the discrete choice condition, we change these to discrete choices: The DM is presented with 11 discrete options chosen from the budget line and is asked to choose one of them rather than directly choose from the budget line (37, 53). Specifically, the third prompt changes to the following: “In this round, there are 11 options, which are  $(A_0, B_0)$ ,  $(A_1, B_1)$ , ..., and  $(A_{10}, B_{10})$ . Please only tell me your best option in every round”. This allows us to examine whether rationality of GPT is robust to the change from continuous to discrete choice sets.

**Response to demographic information.** We also investigate whether the rationality exhibited by GPT varies with the embedded demographic information. To achieve this, we include demographic information which varies in gender, age, education level, and minority group status. We change the input content of the system’s role in GPT to be “I want to you to act as a [demographic] decision maker, ...”. Variations are gender: “female decision maker” versus “male decision maker”; age: “young child decision maker” versus “elderly decision maker”; education: “decision maker with an elementary school education” versus “decision maker with a college education”; and minority: “Asian decision maker” versus “African American decision maker”. By doing so, we can examine whether GPT is responsive to demographic information and whether it performs differently under different individual characteristics. The responsiveness, if any, is relevant to the discussion about algorithm bias (42).

**Design of the Human Experiment.** To obtain a better understanding of the behavior of GPT, we also conduct a human subject experiment with identical decision tasks, in which 347 human subjects from a representative US sample are randomly assigned to the baseline, price framing, and discrete choice conditions.\* We keep the experimental instructions between human subjects and GPT as similar as possible. *SI Appendix* provides the design and instructions of this pre-registered human experiment (AEARCTR-0011750). This experiment was approved by The

Institutional Review Board of Finance and Economics Experimental Laboratory in The Wang Yanan Institute of Studies in Economics, Xiamen University (FEEL230701), and all subjects provided informed consent before they started the experiment. *SI Appendix, Table S1* shows the demographic characteristics of our human subjects.

### Revealed Preference Analysis.

**Generalized axiom of revealed preference.** Consider a DM who selects a bundle  $x^i \in \mathbb{R}_+^K$  from a budget line  $\{x : p^i \cdot x \leq p^i \cdot x^i, p^i \in \mathbb{R}_{++}^k\}$ . A dataset  $\mathcal{O} = (p^i, x^i)_{i=1}^N$  represents a collection of  $N$  decisions made by the DM. We say that a utility function  $U : \mathbb{R}_+^k \rightarrow \mathbb{R}$  rationalizes the dataset  $\mathcal{O}$  if for every bundle  $x^i$ , we have:

$$U(x^i) \geq U(x) \text{ for all } x \in \mathbb{R}_+^K \text{ s.t. } p^i \cdot x \leq p^i \cdot x^i.$$

Let  $\mathcal{X} = \{x^i\}_{i=1}^N$  be the set of bundles selected by the DM. We say that  $x^i$  is directly revealed to be preferred to  $x^j$ , denoted by  $x^i \succ^* x^j$ , if the DM chooses  $x^i$  when  $x^j \in \mathcal{X}$  is affordable (i.e.,  $p^i \cdot x^j \leq p^i \cdot x^i$ ). We denote  $\succ^*$  as the relation of directly strictly revealed preference. We denote  $\succ^{**}$  as the transitive closure of  $\succ^*$ , which refers to the revealed preferred relation.

A utility function is well behaved if it is continuous, concave, and strictly increasing. Afriat’s theorem (11, 14) states that a dataset  $\mathcal{O}$  can be rationalized by a well-behaved utility function if and only if the dataset obeys the generalized axiom of revealed preference (GARP):

$$\text{for all } x^i \text{ and } x^j, x^i \succ^{**} x^j \text{ implies } x^j \not\succeq^* x^i.$$

Apart from GARP, two closely related notions are the weak axiom of revealed preference (WARP): for all  $x^i$  and  $x^j$  in a dataset  $\mathcal{O}$ ,  $x^i \succ^* x^j$  implies  $x^j \not\succeq^* x^i$ , and the strong axiom of revealed preference (SARP): for all  $x^i$  and  $x^j$  in a dataset  $\mathcal{O}$ ,  $x^i \succ^{**} x^j$  implies  $x^j \not\succeq^{**} x^i$ , which works by exploiting transitivity. In our setting with two goods, checking WARP is equivalent to checking SARP (54). In our discrete setting, (23) shows that a locally non-satiated, strictly monotonic, continuous, and concave utility may violate GARP and demonstrates the need to use the assumption of strong monotonicity (see also ref. 55 for discussions).

**Rationality score.** Afriat’s theorem provides a powerful tool for analyzing choice behavior. A popular approach for measuring the departure from rationality is the critical cost efficiency index (CCEI) proposed by Afriat (12). A subject has a CCEI  $e \in [0, 1]$  if  $e$  is the largest number with a well-behaved  $U$  that rationalizes the data set for every  $x^i \in \mathcal{X}$ :

$$U(x^i) \geq U(x) \text{ for all } x \in \mathbb{R}_+^K \text{ s.t. } p^i \cdot x \leq e \cdot p^i \cdot x^i.$$

A CCEI of 1 indicates passing GARP perfectly. A CCEI less than 1—say, 0.95—indicates that there is a utility function for which the chosen bundle  $x^i$  is preferred to any bundle that is cheaper than  $x^i$  for more than 5%. Put differently, the CCEI can be viewed as the amount by which a budget constraint must be relaxed in order to remove all violations of GARP, because the DM can achieve her utility targets by spending less money (12, 15). We compute CCEI to obtain a score of rationality for each domain with 25 decisions.

In the revealed preference literature, there are several other indices to score rationality (departure from GARP). These indices include the Houtman–Maks index (HMI) (56), money pump

\*Variations of temperature are inapplicable among human beings, while variations of demographics can be naturally obtained in a representative sample.

index (MPI) (28), and minimum cost index (MCI) (29). We also compute these indices and report the results as robustness checks.

**Structural Estimation for Preferences.** In addition to rationality score, we further examine the underlying preferences using structural estimation.

**Risk and time preferences estimation.** In the domain of risk preference, suppose that the DM chooses the contingent security  $(x_A, x_B)$ , we denote  $x_1 = \max\{x_A, x_B\}$  as the high outcome and  $x_2 = \min\{x_A, x_B\}$  as the low outcome. In the domain of time preference, suppose that the DM chooses the payment schedule  $(x_A, x_B)$ , we denote  $x_1 = x_A$  as the payment for today and  $x_2 = x_B$  as the payment for 1 mo later. For these two domains, we assume that the underlying utility function is given by:

$$U(x_1, x_2) = \alpha u(x_1) + (1 - \alpha)u(x_2),$$

where the utility function  $u(z) = \begin{cases} \frac{1}{\rho} z^\rho, & \rho \leq 1 (\rho \neq 0) \\ \ln(z), & \rho = 0 \end{cases}$  and

$\alpha \in [0, 1]$ . For risk preference,  $\alpha$  captures the decision weight placed on the better outcome (24, 57). When  $\alpha = 0.5$ , we have a standard expected utility function and when  $\alpha > 0.5$  ( $\alpha < 0.5$ ), the better outcome is over(under)-weighted relative to the objective probability of 0.5. The parameter  $\rho$  captures risk attitude with the parameter  $\theta = 1 - \rho$  being the Arrow–Pratt measure of relative risk aversion.<sup>†</sup> For time preference,  $\alpha$  captures the weight placed on the payment today (19). When  $\alpha > 0.5$  ( $\alpha < 0.5$ ), it corresponds to positive (negative) time preference. The parameter  $\rho$  is the curvature of the period function. When  $\rho = 1$ , the DM allocates all expenditure to the time period with lower price, and as  $\rho$  decreases, the DM is more desired to smooth payments across periods.

**Social and food preferences estimation.** Regarding social preference, suppose that the DM chooses the allocation  $(x_A, x_B)$ , we denote  $x_1 = x_A$  as the payment for self and  $x_2 = x_B$  as the payment for the other. In the domain of food preference, assuming that the DM chooses the bundle  $(x_A, x_B)$ , we denote  $x_1 = x_A$  as the consumption of meat and  $x_2 = x_B$  as the consumption of tomatoes. Moreover, we assume that the underlying utility function is a member of the CES family and is given by:

$$U(x_1, x_2) = [\alpha x_1^\rho + (1 - \alpha)x_2^\rho]^{\frac{1}{\rho}},$$

where  $\rho \leq 1$  and  $\alpha \in [0, 1]$ . For social preference, the parameter  $\alpha$  captures the weight placed on the self's payment relative to the other's payment.  $\alpha = 1$  implies pure selfishness,  $\alpha = 0.5$  indicates fair-mindedness, and  $\alpha = 0$  refers to pure altruistic (18, 25).  $\rho$  represents the curvature of the indifference curves, which measures equality efficiency orientation.  $\rho = 1$  indicates that the two payments are perfectly substitute with  $U(x_1, x_2) = \alpha x_1 + (1 - \alpha)x_2$ , which means that the DM is efficiency orientated. When  $\rho \rightarrow 0$ , the utility function approaches the Cobb–Douglas utility function, and shares of expenditures to self and to the other are constant. When  $\rho \rightarrow -\infty$ , it approaches to the Leontief utility function  $\min\{\alpha x_1, (1 - \alpha)x_2\}$ , which implies that the two payments are perfectly complemented and the DM is equality orientated (18, 25). In a similar vein, the parameter  $\alpha$  in the food preference domain captures the weight placed on

<sup>†</sup>In our budget set, there is no difference between  $\rho > 1$  and  $\rho = 1$ , because the DM will choose corner solutions when  $\rho \geq 1$ . Therefore, our estimation is conditional on  $\rho \leq 1$  in all the four preference domains. *SI Appendix* provides further details about the estimation of corner solutions.

meat relative to tomatoes and the parameter  $\rho$  represents the curvature of the indifference curves as that for social preference. We provide further details on estimation methods in *SI Appendix*.

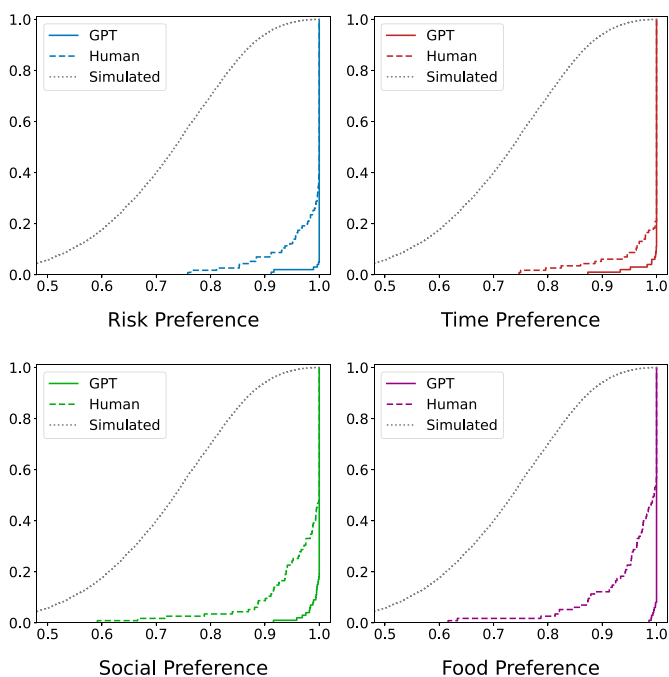
## Results

In this section, we first present the results from the baseline condition, and then report whether and how the results change with the variations in the decision tasks.

### Results from the Baseline Condition.

**Rationality score.** Fig. 1 presents the cumulative distributions of CCEI—the rationality score—for each of the four preference domains. We find that 95, 89, 81, and 92 out of 100 GPT observations for risk, time, social, and food preferences exhibit no violations of GARP; that is, CCEI equals 1. The average CCEI is 0.998, 0.997, 0.997, and 0.999 for risk, time, social, and food preferences, respectively. Meanwhile, in our human experiment, the average CCEI among human subjects is 0.980, 0.985, 0.967, and 0.963 for risk, time, social, and food preferences, respectively. Fig. 1 displays a consistent trend that GPT outperforms human subjects in terms of rationality. In each of the four preference domains, GPT obtains higher CCEI than human subjects ( $P < 0.01$ , two-sided two-sample  $t$ -tests). In addition, we summarize studies in the revealed preference literature. *SI Appendix, Fig. S1* plots CCEI values documented in prior studies, which range from 0.81 to 0.99 with an average of 0.918. Consistently, we find that CCEI of GPT also surpasses those of human subjects in all domains ( $P < 0.01$ , two-sided one-sample  $t$ -tests).

To confirm that our chosen parameters have sufficient power to measure rationality, we adopt the test proposed by Bronars (58) as a benchmark, in which we generate simulated subjects



**Fig. 1.** Cumulative distributions of the CCEI values. This figure consists of four subplots for four preference domains. Each subplot depicts a cumulative distribution function (CDF) plot, which shows the proportion of CCEI values less than or equal to a specific threshold. The light dotted lines represent simulated subjects, the dark dashed lines represent human subjects, and the solid lines represent GPT observations.

by uniformly drawing random allocations along each of the budget lines and examine their rationality. We find that 99.9% of simulated subjects violate GARP. Fig. 1 shows the cumulative distributions of CCEI of simulated subjects, which are lower than both GPT observations and human subjects. We also conduct the power analysis using the predictive success (59), the Selten score (29), as well as bootstrapping from the sample of subjects (18). We show that the chosen parameters have the power to detect rationality violations, in support of the empirical validity of our study (see *SI Appendix* for more information).

In addition to CCEI, we calculate other indices to measure rationality including the Houtman–Maks index (HMI) (56), money pump index (MPI) (28), and minimum cost index (MCI) (29), and construct cumulative distribution plots for each index of GPT observations, human subjects, and simulated subjects in *SI Appendix*, Figs. S2–S5. Consistent with the observations based on CCEI, results from these indices show that GPT observations exhibit a high level of rationality across the four preference domains and surpass those of human subjects across all domains ( $P < 0.1$ , two-sided two-sample  $t$ -tests).

**Downward-sloping demand.** While GPT exhibits a high level of rationality, it is possible that its decisions are simply clustered at the corners or in certain areas. To address such concern, we examine whether GPT behavior respects the property of downward-sloping demand, a fundamental principle in the analysis of consumer behavior whereby the demand for a commodity decreases with its price (21, 25, 60).

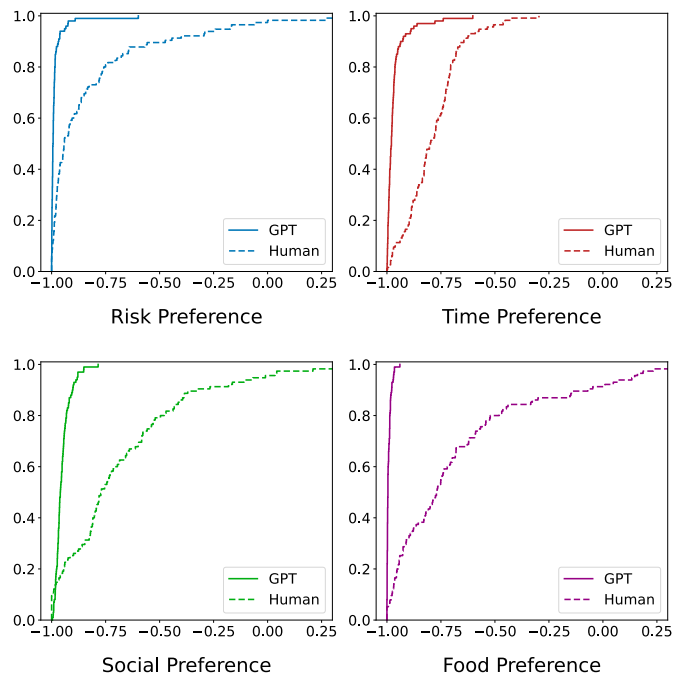
We measure the degree of compliance with downward-sloping demand for GPT observations and human subjects. This principle requires that when the relative price of a commodity increases, the consumer should not increase its consumption. More specifically, we measure whether each DM's decisions respect this principle by calculating the Spearman's correlation coefficient of  $\ln(x_A/x_B)$  and  $\ln(p_A/p_B)$  (60). A negative correlation indicates an appropriate response to price fluctuations, and zero or positive correlation indicates no response or irregular response to price changes. Note that  $\ln(x_A/x_B)$  is not defined in the corners. We adjust corner choices by a small constant, 0.1% of the budget, in each choice (60). We plot the cumulative distribution of the Spearman's correlation coefficients of  $\ln(x_A/x_B)$  and  $\ln(p_A/p_B)$  as a proxy for the degree of downward-sloping demand for each of the four preference domains in Fig. 2.

For GPT observations, the coefficients for risk, time, social, and food preferences have a mean of  $-0.984$ ,  $-0.966$ ,  $-0.951$ , and  $-0.992$ , while these are  $-0.826$ ,  $-0.788$ ,  $-0.681$ , and  $-0.673$  for human subjects, respectively. Overall, GPT is more responsive to price changes than human subjects in each preference domain ( $P < 0.01$ , two-sided two-sample  $t$ -tests). Fig. 2 further illustrates that GPT observations always have negative Spearman's correlation coefficients, while human has a lower proportion having negative Spearman's correlation coefficients (96.1% on average). This strengthens our findings based on the rationality score and suggests that GPT is more capable of making reasonable responses to the changes in prices than human subjects.

In addition, for each GPT observation, *SI Appendix*, Figs. S6–S9 provide comprehensive visual representations by showing scatter diagrams and fitted lines of the shares of quantities  $x_A/(x_A + x_B)$  and the log-price ratio  $\ln(p_A/p_B)$ .

**Preference estimation.** Since choices of GPT and human subjects are mostly consistent with well-behaved utility functions, we proceed to estimate the underlying risk, time, social, and food preferences.<sup>‡</sup> In total, we have eight estimated parameters:

<sup>‡</sup>We omit GPT or human individuals with a CCEI score below 0.95 (15).

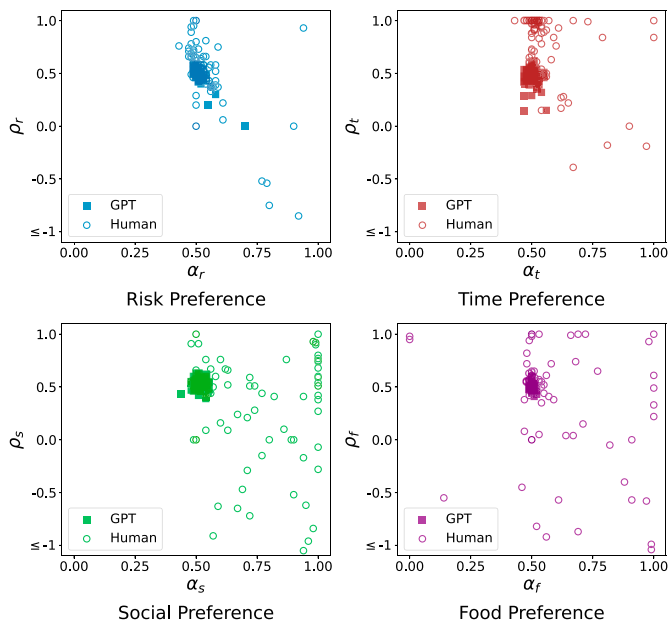


**Fig. 2.** Cumulative distributions of the Spearman's correlation coefficient of  $\ln(x_A/x_B)$  and  $\ln(p_A/p_B)$ . This figure contains four subplots for four preference domains. The dashed (solid) lines represent human subjects (GPT observations).

decision weight of the better outcome ( $\alpha_r$ ) and utility curvature ( $\rho_r$ ) for risk preference, weight of today ( $\alpha_t$ ) and utility curvature ( $\rho_t$ ) for time preference, weight for self's payment ( $\alpha_s$ ) and utility curvature ( $\rho_s$ ) for social preference, weight for meat ( $\alpha_f$ ) and utility curvature ( $\rho_f$ ) for food preference. We first estimate the preference parameters at the aggregate level by pooling all responses of GPT observations and human subjects, respectively (*SI Appendix*, Table S2). Results show that, compared to human subjects, GPT is closer to an expect-utility maximizer ( $\alpha_r$ : 0.618 vs. 0.508 for Human vs. GPT) and has a more linear utility curve ( $\rho_r$ : 0.335 vs. 0.488) in risk preference; is more patient ( $\alpha_t$ : 0.513 vs. 0.504) and has a less linear utility curve ( $\rho_t$ : 0.466) in time preference; is more other-regarding ( $\alpha_s$ : 0.735 vs. 0.512) and more efficiency-orientated ( $\rho_s$ : 0.330 vs. 0.520) in social preference; and is less fond of meat ( $\alpha_f$ : 0.583 vs. 0.501) and more efficiency-orientated ( $\rho_f$ : 0.386 vs. 0.491) in food preference. Similar patterns can be observed in the individual-level estimations, in which we estimate preference parameters for each GPT decision maker and human subject, as shown in Fig. 3 and *SI Appendix*, Table S3. Moreover, the scatter plots of human subjects are more dispersed, which suggests a significantly higher level of preference heterogeneity among human subjects than GPT observations.

**Results from the Conditions with Variations.** We examine variations in the temperature, decision tasks, and demographic information. Fig. 4 presents the mean CCEI values and 95% CIs across variations, and *SI Appendix*, Fig. S24 shows the mean Spearman's correlation coefficients of  $\ln(x_A/x_B)$  and  $\ln(p_A/p_B)$  and their 95% CIs. We report these results in detail below.

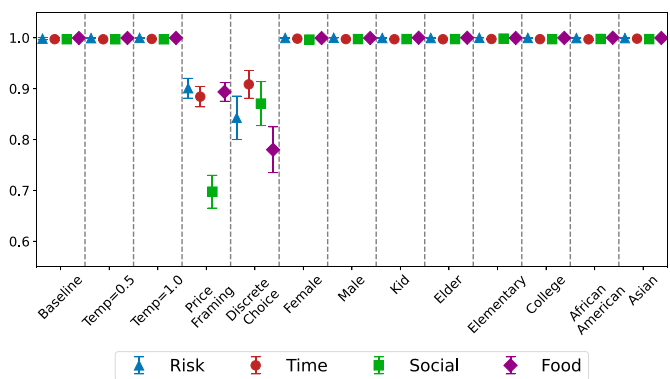
**Insensitive to variations in temperature.** When the temperature increases from 0 to 0.5 and 1, there is a higher number of invalid responses, namely, GPT does not provide an answer to the specified question (invalid response rate is 4.7% for temperature of 0.5 and 9.8% for temperature of 1). Therefore, we analyze the data conditional on those providing valid answers. We find that



**Fig. 3.** Scatter plots of estimated parameters. This figure contains four subplots for four preference domains. Each hollow circle (solid square) points represent a human subject (a GPT observation).

as the randomness increases, the level of rationality is similar to that in the baseline condition (Fig. 4). For each temperature and each preference domain, we plot the cumulative distributions of the CCEI values of GPT observations and simulated subjects for Bronars' test in *SI Appendix, Fig. S10* and the cumulative distributions of the Spearman's correlation coefficients in *SI Appendix, Fig. S11*. These findings suggest that randomness increases the stochasticity and creativity in language presentations of GPT, but not the rationality score.

There are no significant differences for the estimated Spearman's correlation coefficients of  $\ln(x_A/x_B)$  and  $\ln(p_A/p_B)$  between the baseline and the higher temperature (*SI Appendix, Fig. S24*) at the 10% level (two-sided two-sample *t*-tests). Similarly, the mean of estimated preference parameters is statistically indifferent to changes in temperature. However, the SDs of some parameters increase with temperature ( $\rho_r, \rho_t, \rho_s, \rho_f$ ;  $P < 0.01$ , two-sample Levene tests), which suggests that high temperature may generate greater heterogeneity in the behavior of GPT.



**Fig. 4.** Mean CCEI values of GPT observations across different variations. This figure displays the average CCEI values and 95% CIs for GPT observations under different conditions: baseline, temperature of 0.5, temperature of 1, price framing, and discrete choices, and various demographic settings.

**Sensitive to variation in the decision tasks.** First, we compare the baseline and the price framing conditions. Changing the price framing significantly reduces GPT's rationality level in all four tasks (Fig. 4). Remarkably, the average CCEI for risk preference declines to 0.901, with 34% exhibiting a CCEI below 0.9. These values are 0.884 (48%), 0.698 (88%), and 0.894 (49%) for time, social, and food preferences, respectively.<sup>§</sup> In each preference domain, CCEI values are significantly higher in the baseline condition than in the price framing condition ( $P < 0.01$ , two-sided two-sample *t*-tests). Moreover, the downward-sloping demand property is impaired under the alternative price framing, with the key Spearman's correlation coefficients being  $-0.053$ ,  $-0.116$ ,  $0.267$ , and  $-0.499$  for risk, time, social, and food preferences, respectively (*SI Appendix, Fig. S24*). *SI Appendix, Figs. S13–S16* show the disordered responses of GPT observations to price changes in the price framing condition, which appear to be flatter compared to those in the baseline condition.

In *SI Appendix, Figs. S17 and S18*, we display the CDFs of CCEI (Spearman's correlation coefficients) in both the GPT experiment and the human experiment. We find that the alternative price framing also reduces the rationality level and the downward-sloping demand property in the human subjects experiment ( $P < 0.05$ , two-sided two-sample *t*-tests in risk and time preferences). However, the figures suggest that these reductions are larger in the GPT experiment than in the human experiment, which is further verified in OLS regression analyses (*SI Appendix, Table S4*).

Second, we compare the baseline and the discrete choice conditions. When we present GPT with a set of 11 options, we also observe a decrease in rationality levels for discrete choices of GPT observations for all four tasks in Fig. 4 (risk: 0.998 vs. 0.843,  $P < 0.01$ ; time: 0.997 vs. 0.908,  $P < 0.01$ ; social: 0.997 vs. 0.871,  $P < 0.01$ ; food: 0.999 vs. 0.780,  $P < 0.01$ , two-sided two-sample *t*-tests). Additionally, 51%, 32%, 33%, and 55% of GPT observations demonstrate a CCEI below 0.9 in risk, time, social, and food preferences, respectively. *SI Appendix, Figs. S19–S22* show the demand curves of GPT observations, which exhibit significantly more corner solutions. Consistently, the Spearman's correlation coefficients are  $-0.589$ ,  $-0.497$ ,  $-0.519$ , and  $-0.533$  for risk, time, social, and food preferences (*SI Appendix, Fig. S24*;  $P < 0.01$  when compared to the baseline condition, two-sided two-sample *t*-tests). These suggest that GPT is less responsive to price changes under discrete choices than continuous choices.

*SI Appendix, Figs. S23 and S24* shows the CDFs of the CCEI (Spearman's correlation coefficients) in baseline and discrete choice conditions in the GPT experiment and human experiment. Human subjects' rationality level and the downward-sloping demand property reduce in the discrete setting, compared to the baseline condition ( $P < 0.05$ , two-sided two-sample *t*-tests in risk and time preferences). As shown in the figures, these reductions are larger in the GPT experiment than in the human experiment. We also verify this observation through OLS regression analyses (*SI Appendix, Table S4*). These results suggest that GPT's decision-making is more significantly affected by both the framing of prices and discrete choices than human subjects.

**Insensitive to demographic information.** Comparing the baseline condition and variations in demographics in the GPT experi-

<sup>§</sup> Given the low level of rationality exhibited by GPT in the price framing condition, we have difficulty in determining that GPT's decisions are consistent with a well-behaved utility function. Therefore, we refrain from adopting the preference estimation approach under this condition (15). The situation is identical in the discrete choice condition as described below.

ment, we find that CCEI values, Spearman's correlation coefficients of  $\ln(x_A/x_B)$  and  $\ln(p_A/p_B)$ , and estimated preference parameters are all insensitive to variations of demographic factors embedded in the prompts to request responses from GPT (Fig. 4, *SI Appendix*, Fig. S24 and Tables S2 and S3).

These are in contrast to results of our human experiment (*SI Appendix*, Tables S5 and S6) and prior studies where rationality score and preference have been shown to differ across demographic groups (22, 28, 61). The fact that GPT's decision-making process remains consistent across demographic variables suggests that GPT does not exhibit algorithmic bias in terms of decision-making quality, which provides a measure of reassurance regarding its fairness and consistency across diverse user groups.

## Discussion

We conduct a study to assess the rationality of GPT, a popular large language model, using revealed preference analysis. Our findings demonstrate that GPT is able to display a high level of rationality in decision-making related to risk, time, social, and food preferences. We also observe that increasing the randomness of GPT does not significantly impact its performance. Furthermore, our analysis reveals that the level of rationality of GPT remains constant across different demographic characteristics, which indicates that it does not exhibit an algorithm bias. However, we observe a significant drop in rationality when we use a less standard presentation of prices or change the choice set from continuous to discrete. This suggests that GPT may have limitations in terms of sensitivity to contexts and frames.

Our study contributes to the ongoing discussions of the performance of GPT in various domains; these include reasoning, logic, math, language processing, and identifying factual errors (45). In addition to cognitive techniques and practical skills, some researchers have explored whether GPT can exhibit human-like decision-making abilities or perceive others' thoughts (6, 10). Our study adds to these parallel studies by subjecting GPT to traditional decision-making tasks and employing a set of measures to systematically describe its behavior. Our work aligns with recent calls to study machine behavior to "reap their benefits and minimize their harms" (43). By providing insights into GPT's decision-making capacity, we can better understand how to optimize its performance and address potential limitations.

Our study is situated within the growing literature on AI-based decision support tools. Many researchers have explored the usefulness of leveraging AI in various decision-making domains, such as bail decisions (62); clinical diagnosis (63); work arrangements (64); stock price forecasts (65); job recruitment (66); product or content consumption (67, 68); and mathematics development (69). Unlike these algorithms, which require data input and training, GPT is a language-based model that provides a direct question-and-answer service for normal users. Given its high level of rationality in decision-making across various domains, our study proves the potential of GPT as a general AI-based decision-support tool. The user-friendly interface and versatility of GPT render it a promising option for individuals and organizations seeking easy-to-use AI-based advice.

Our paper makes contributions to the literature on rationality and experimental methods. First, we demonstrate the effectiveness of experimental economics methods in studying choice behavior of artificial intelligence (43), which adds earlier studies of children (23, 30), monkeys (31), rats, and pigeons (32). Second, our work highlights the potential of large language

models like GPT to streamline experimental research and yield new data and insights (10). Finally, studying the choice behavior of artificial intelligence can provide an important benchmark for understanding natural intelligence. For example, our understanding of how LLMs make decisions could help reveal general principles that govern both language intelligence and decision intelligence (70). By synthesizing insights from these various domains, our paper offers a unique perspective on the nature of rationality and broadens the methods that can be used to study it.

As an initial assessment of the economic rationality of GPT, our study has several limitations. First, our study examines the choice behavior of GPT but does not explore the mechanisms that underlie our observations. For example, we find that GPT responses are highly sensitive to contexts and frames. This may be due to the reflection of biases presented in the existing data (71, 72), the insufficient training of texts of the alternative contexts and frames (3, 5), or the tendency for LLMs to exploit spurious correlations or statistical irregularities in the data set under dissimilar tasks (73). In particular, ref. 74 suggests that a significant source of LLMs bias originates from a corpus-based heuristic using the relative frequencies of words. The "50-50 split" or "equal split" are high-frequency texts in allocation settings, and GPT can adapt this corpus-based heuristic and exhibit the tendency to choose the midpoint under an "unfamiliar" task with the alternative price framework. Similarly, "all or nothing" can be high-frequency texts under the presentation of options context, so GPT exhibits the tendency to choose the first or last option under an "unfamiliar" discrete choice condition. Recent studies have documented similar patterns in different environments (7, 10, 48, 49, 75).

In addition, our study reveals that demographic factors do not significantly impact GPT's rationality or estimated preference parameters. This contrasts with the majority of the empirical literature, including our human subject experiment, where demographic factors often play a significant role. The lack of responsiveness to demographics aligns with the concept of hyperaccuracy distortion (76), which refers to the distortion resulting from the extensive efforts to align LLMs with human ethics such as the censorship of demographic information to reduce and prevent problematic outputs. In conclusion, with some speculative conjectures, we leave it to future studies to explore the mechanisms that underlie GPT's choice behavior and open the black-box of this technology.

Second, we focus on economic rationality as defined by revealed preference analysis, whereas rationality is often defined more broadly in the literature to include various decision rules and heuristics (40, 77, 78). Third, we use a simple experimental environment with only two commodities to present budgetary decisions. However, studying rationality in more realistic settings, such as shopping behavior in a supermarket and portfolio choices in the financial markets would be more challenging yet important. Our study shows that economic rationality can emerge in GPT when decision contexts are simple and framed in specific ways. Future research is needed to investigate the broader applications of artificial intelligent agents as they continue to evolve.

## Materials and Methods

**Preference Utility Estimation.** Once the DM's CCEI is sufficiently to justify treating the data as generated utility functions of well-behaved behavior, another interest is to estimate the preference parameters of the DM. Our estimations will be made for each DM separately on the assumption of the underlying utility

function commonly employed found in the prior literature to capture risk, time, social, and food preferences.

**Risk preference.** For the DM's choice of contingent securities  $x : (x_A, x_B)$ , referring to ref. 24, we assume  $x_1$  is the better outcome,  $x_1 = \max\{x_A, x_B\}$ , and  $x_2$  is the worse outcome,  $x_2 = \min\{x_A, x_B\}$ . We estimate the risk preference of the DM by a functional form of disappointment aversion (DA) introduced by ref. 57:

$$U(x_1, x_2) = \alpha u(x_1) + (1 - \alpha)u(x_2),$$

where  $u(z)$  is the CRRA utility function:

$$u(z) = \begin{cases} \frac{1}{\rho} z^\rho & \rho \leq 1 (\rho \neq 0) \\ \ln(z) & \rho = 0. \end{cases}$$

The parameter  $\alpha \in [0, 1]$  is the weight placed on the better outcome. For  $\alpha > 0.5$  ( $\alpha < 0.5$ ), the better (worse) outcome is over weighted relative to the objective probability (of 0.5). If  $\alpha = 0.5$ , we have a standard expected utility decision maker. The parameter  $\theta = 1 - \rho \geq 0$  is the Arrow-Pratt measure of relative risk aversion.

**Time preference.** Following the method of ref. 19, we use a time separable period utility function to estimate the DM's time preference from intertemporal decisions of the payoff between today and 1 mo later  $x : (x_A, x_B)$ , with the assumption that background income is zero:

$$U(x_A, x_B) = \alpha u(x_A) + (1 - \alpha)u(x_B),$$

where  $u(z)$  is the period utility function:

$$u(z) = \begin{cases} \frac{1}{\rho} z^\rho & \rho \leq 1 (\rho \neq 0) \\ \ln(z) & \rho = 0. \end{cases}$$

The parameter  $\alpha \in [0, 1]$  is the weight placed on the today. For  $\alpha > 0.5$  ( $\alpha < 0.5$ ), today is over weighted relative to 1 mo later. The parameter  $\rho \leq 1$  is the curvature of the period utility function.

**Social preference.** For capturing DM's social preferences in choice  $x : (x_A, x_B)$ , the payoff of self and the other, following the method of refs. 18 and 25, we assume that  $U(x_A, x_B)$  is a member of the CES family. It is given by

$$U(x_A, x_B) = [\alpha (x_A)^\rho + (1 - \alpha) (x_B)^\rho]^{1/\rho}.$$

The CES specification is very flexible, spanning a range of well-behaved utility. The parameter  $\alpha \in [0, 1]$  represents the relative weight on the payoff for the self of the DM.  $\alpha = 0.5$  indicates fair-mindedness, whereas  $\alpha = 1$  indicates pure selfishness and  $\alpha = 0$  indicates pure selflessness. The parameter  $\rho \leq 1$  represents the curvature of the indifference curves (equality-efficiency tradeoffs, an important implication in social preference).  $\rho \rightarrow 0$  indicates a Cobb-Douglas function. When  $\rho \rightarrow 1$ , the utility approaches perfect substitutes:  $\alpha x_A + (1 - \alpha)x_B$ ; when  $\rho \rightarrow -\infty$ , the utility approaches Leontief:  $\min\{\alpha x_A, (1 - \alpha)x_B\}$ .  $\rho$  represents "equality-efficiency" trade offs.  $\rho > 0$  ( $\rho < 0$ ) indicates toward efficiency (increasing total payoffs) because the expenditure decreases when the relative price increases.

**Food preference.** For DM's choice  $x : (x_A, x_B)$  in food preference, the quantities of meat, and the quantities of tomatoes, we assume that  $U(x_A, x_B)$  is a member of the CES family. It is given by:

$$U(x_A, x_B) = [\alpha (x_A)^\rho + (1 - \alpha) (x_B)^\rho]^{1/\rho}.$$

The parameter  $\alpha \in [0, 1]$  represents the relative weight on the utility of meat. The parameter  $\rho \leq 1$  represents the curvature of the indifference curves.  $\rho \rightarrow 0$  indicates a Cobb-Douglas function, which implies that expenditures on meat and tomatoes are equal to fractions  $\alpha$  and  $1 - \alpha$ , respectively.  $\rho > 0$  ( $\rho < 0$ ) indicates a decrease in the relative price of meat and tomatoes ( $p_A/p_B$ ) lowers (raises) the fraction of tomatoes in total expenditure.

**Econometric specification.** For the four utility functions above, the first-order conditions at the optimal choice  $(x_A, x_B)$ , given  $(p_A, p_B)$ <sup>4</sup>, can be written as follows:

$$\ln(x_A/x_B) = \frac{1}{\rho - 1} \left[ \ln(p_A/p_B) + \ln \frac{1 - \alpha}{\alpha} \right].$$

Because  $\ln(x_1/x_2)$  is not well defined in corner solutions, referring to the method of refs. 25, 38, and 39, the demand function is given by:

$$x_A = \left[ \frac{g}{(p_A/p_B)^r + g} \right] \frac{E}{p_A},$$

where  $E$  is the expenditure,  $r = \rho/(1 - \rho)$ , and  $g = [\alpha/(1 - \alpha)]^{1/(1-\rho)}$ . This generates the following econometric specification:  $\frac{p_A x_A}{E} = \frac{g}{(p_A/p_B)^r + g}$ .

Note again that expenditure shares are bounded between zero and one. We can generate estimates of  $g$  and  $r$  using nonlinear tobit maximum likelihood and use this to infer the values of the underlying  $\alpha$  and  $\rho$  for four preference domains.

**Data, Materials, and Software Availability.** Code and data for the current study are publicly available through <https://www.dropbox.com/scl/fo/572ptz57vjs5cqkczj9l/h?rlkey=hpsgdb6ghdzsvj35mfdnwc&dl=0> (79).

**ACKNOWLEDGMENTS.** T.X.L. gratefully acknowledges financial support from National Natural Science Foundation of China (72222005) and Tsinghua University (No. 2022Z04W01032). S.Z. gratefully acknowledges financial support from the Singapore Ministry of Education (Academic Research Fund Tier 1) and the National University of Singapore (Dean's Chair Grant).

Author affiliations: <sup>a</sup>Department of Economics, Lingnan University, Hong Kong, China HKG; <sup>b</sup>Department of Economics, School of Economics and Management, National Center for Economic Research at Tsinghua University, Tsinghua University, Beijing 100084, China; <sup>c</sup>Department of Economics, School of Economics and Management, Tsinghua University, Beijing 100084, China; <sup>d</sup>Department of Economics, Hong Kong University of Science and Technology, Hong Kong, China HKG; and <sup>e</sup>Department of Economics, National University of Singapore, Singapore 117570, Singapore

<sup>4</sup> $(x_1, x_2)$  and  $(p_1, p_2)$  in risk preference.

1. A. Vaswani et al., Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 1–11 (2017).
2. T. Brown et al., Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
3. M. Chen et al., Evaluating large language models trained on code. arXiv [Preprint] (2021). <http://arxiv.org/abs/2107.03374> (Accessed 25 November 2023).
4. Z. Lin et al., Cairo: An end-to-end empathetic chatbot. *Proc. AAAI Conf. Artif. Intell.* **34**, 13622–13623 (2020).
5. I. Drori et al., A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2123433119 (2022).
6. M. Kosinski, Theory of mind may have spontaneously emerged in large language models. arXiv [Preprint] (2023). <http://arxiv.org/abs/2302.02083> (Accessed 25 November 2023).
7. M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218523120 (2023).
8. P. S. Park, P. Schoenegger, C. Zhu, Artificial intelligence in psychology research. arXiv [Preprint] (2023). <http://arxiv.org/abs/2302.07267> (Accessed 25 November 2023).
9. T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541 (2023).

10. J. J. Horton, Large language models as simulated economic agents: What can we learn from homo silicus? NBER Working Paper (2023).
11. S. N. Afriat, The construction of utility functions from expenditure data. *Int. Econ. Rev.* **8**, 67–77 (1967).
12. S. N. Afriat, Efficiency estimation of production functions. *Int. Econ. Rev.* **13**, 568–598 (1972).
13. P. A. Samuelson, A note on the pure theory of consumer's behaviour. *Economica* **5**, 61–71 (1938).
14. H. R. Varian, The nonparametric approach to demand analysis. *Econometrica* **50**, 945–973 (1982).
15. H. R. Varian, Goodness-of-fit in optimizing models. *J. Econ.* **46**, 125–140 (1990).
16. C. P. Chambers, F. Echenique, *Revealed Preference Theory* (Cambridge University Press, 2016).
17. H. Nishimura, E. A. Ok, J. K. H. Quah, A comprehensive approach to revealed preference theory. *Am. Econ. Rev.* **107**, 1239–1263 (2017).
18. J. Andreoni, J. Miller, Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* **70**, 737–753 (2002).
19. J. Andreoni, C. Sprenger, Estimating time preferences from convex budgets. *Am. Econ. Rev.* **102**, 3333–3356 (2012).
20. D. Ahn, S. Choi, D. Gale, S. Kariv, Estimating ambiguity aversion in a portfolio choice experiment. *Quant. Econ.* **5**, 195–223 (2014).



21. S. Choi, R. Fisman, D. Gale, S. Kariv, Consistency and heterogeneity of individual behavior under uncertainty. *Am. Econ. Rev.* **97**, 1921–1938 (2007).
22. S. Choi, S. Kariv, W. Müller, D. Silverman, Who is (more) rational? *Am. Econ. Rev.* **104**, 1518–1550 (2014).
23. W. T. Harbaugh, K. Krause, T. R. Berry, GARP for kids: On the development of rational choice behavior. *Am. Econ. Rev.* **91**, 1539–1545 (2001).
24. Y. Halevy, D. Persitz, L. Zrill, Parametric recoverability of preferences. *J. Polit. Econ.* **126**, 1558–1593 (2018).
25. R. Fisman, S. Kariv, D. Markovits, Individual preferences for giving. *Am. Econ. Rev.* **97**, 1858–1876 (2007).
26. R. Blundell, M. Browning, I. Crawford, Nonparametric Engel curves and revealed preference. *Econometrica* **71**, 205–240 (2003).
27. R. Blundell, M. Browning, I. Crawford, Best nonparametric bounds on demand responses. *Econometrica* **76**, 1227–1262 (2008).
28. F. Echenique, S. Lee, M. Shum, The money pump as a measure of revealed preference violations. *J. Polit. Econ.* **119**, 1201–1223 (2011).
29. M. Dean, D. Martin, Measuring rationality with the minimum cost of revealed preference violations. *Rev. Econ. Stat.* **98**, 524–534 (2016).
30. I. Brocas, J. D. Carrillo, T. D. Combs, N. Kodaverdian, The development of consistent decision-making across economic domains. *Games Econ. Behav.* **116**, 217–240 (2019).
31. M. K. Chen, V. Lakshminarayanan, L. R. Santos, How basic are behavioral biases? Evidence from capuchin monkey trading behavior. *J. Polit. Econ.* **114**, 517–537 (2006).
32. J. H. Kagel *et al.*, Experimental studies of consumer demand behavior using laboratory animals. *Econ. Inquiry* **13**, 22–38 (1975).
33. J. Banks, L. Carvalho, F. Perez-Arce, Education, decision making, and economic rationality. *Rev. Econ. Stat.* **101**, 428–441 (2019).
34. A. W. Cappelen, S. Kariv, E. Ø. Sørensen, B. Tungodden, The development gap in economic rationality of future elites. *Games Econ. Behav.* **142**, 866–878 (2023).
35. L. S. Carvalho, S. Meier, S. W. Wang, Poverty and economic decision-making: Evidence from changes in financial resources at payday. *Am. Econ. Rev.* **106**, 260–284 (2016).
36. R. Fisman, P. Jakiela, S. Kariv, Distributional preferences and political behavior. *J. Public Econ.* **155**, 1–10 (2017).
37. H. B. Kim, S. Choi, B. Kim, C. Pop-Eleches, The role of education interventions in improving economic rationality. *Science* **362**, 83–86 (2018).
38. J. Li, W. H. Dow, S. Kariv, Social preferences of future physicians. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E10291–E10300 (2017).
39. J. Li, L. P. Casalino, R. Fisman, S. Kariv, D. Markovits, Experimental evidence of physician social preferences. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2112726119 (2022).
40. D. Kahneman, Maps of bounded rationality: Psychology for behavioral economics. *Am. Econ. Rev.* **93**, 1449–1475 (2003).
41. D. McFadden, Economic choices. *Am. Econ. Rev.* **91**, 351–378 (2001).
42. S. Corbett-Davies, S. Goel, The measure and misuse of fairness: A critical review of fair machine learning. *arXiv [Preprint]* (2018). <http://arxiv.org/abs/1808.00023> (Accessed 25 November 2023).
43. I. Rahwan *et al.*, Machine behaviour. *Nature* **568**, 477–486 (2019).
44. M. Mitchell, D. C. Krakauer, The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2215907120 (2023).
45. A. Borji, A categorical archive of chatGPT failures. *arXiv [Preprint]* (2023). <http://arxiv.org/abs/2302.03494> (Accessed 25 November 2023).
46. Y. Chen, M. Andiappan, T. Jenkin, A. Ovchinnikov, A manager and an AI walk into a bar: Does chatGPT make biased decisions like we do? Available at SSRN 4380365 (2023).
47. K. Mahowald *et al.*, Dissociating language and thought in large language models: A cognitive perspective. *arXiv [Preprint]* (2023). <http://arxiv.org/abs/2301.06627> (Accessed 25 November 2023).
48. E. Jones, J. Steinhart, Capturing failures of large language models via human cognitive biases. *Adv. Neural Inf. Process. Syst.* **35**, 11785–11799 (2022).
49. J. Brand, A. Israeli, D. Ngwe, Using GPT for market research. Available at SSRN 4395751 (2023).
50. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016).
51. M. Bommarito II, D. M. Katz, GPT takes the bar exam. *arXiv [Preprint]* (2022). <http://arxiv.org/abs/2212.14402> (Accessed 25 November 2023).
52. A. C. Drichoutis, R. M. Nayga Jr, Economic rationality under cognitive load. *Econ. J.* **130**, 2382–2409 (2020).
53. M. Chen, T. X. Liu, Y. Shan, S. Zhong, Y. Zhou, The consistency of rationality measures. Working Paper (2023).
54. H. Rose, Consistency of preference: The two-commodity case. *Rev. Econ. Stud.* **25**, 124–125 (1958).
55. M. Polisson, J. K. H. Quah, Revealed preference in a discrete consumption space. *Am. Econ. J.: Microecon.* **5**, 28–34 (2013).
56. M. Houtman, J. Maks, Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve Methoden* **19**, 89–104 (1985).
57. F. Gul, A theory of disappointment aversion. *Econometrica* **59**, 667–686 (1991).
58. S. G. Bronars, The power of nonparametric tests of preference maximization. *Econometrica* **55**, 693–698 (1987).
59. T. K. Beatty, I. Crawford, How demanding is the revealed preference approach to demand? *Am. Econ. Rev.* **101**, 2782–2795 (2011).
60. F. Echenique, T. Imai, K. Saito, Approximate expected utility rationalization. *J. Euro. Econ. Assoc.* **21**, 1821–1864 (2023).
61. H. M. Von Gaudecker, A. Van Soest, E. Wengström, Heterogeneity in risky choice behavior in a broad population. *Am. Econ. Rev.* **101**, 664–694 (2011).
62. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, Human decisions and machine predictions. *Quart. J. Econ.* **133**, 237–293 (2018).
63. S. Mullainathan, Z. Obermeyer, Solving medicine's data bottleneck: Nightingale open science. *Nat. Med.* **28**, 897–899 (2022).
64. Y. Kawaguchi *et al.*, Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions. *arXiv [Preprint]* (2021). <http://arxiv.org/abs/2106.04492> (Accessed 25 November 2023).
65. A. Lopez-Lira, Y. Tang, Can chatGPT forecast stock price movements? Return predictability and large language models *arXiv [Preprint]* (2023). <http://arxiv.org/abs/2304.07619> (Accessed 25 November 2023).
66. J. J. Horton, The effects of algorithmic labor market recommendations: Evidence from a field experiment. *J. Labor Econ.* **35**, 345–385 (2017).
67. G. Adomavicius, J. C. Bockstedt, S. P. Curley, J. Zhang, Effects of online recommendations on consumers' willingness to pay. *Inf. Syst. Res.* **29**, 84–102 (2018).
68. K. Agrawal, S. Athey, A. Kanodia, E. Palikot, Personalized recommendations in EdTech: Evidence from a randomized controlled trial. *arXiv [Preprint]* (2022). <http://arxiv.org/abs/2208.13940> (Accessed 25 November 2023).
69. A. Davies *et al.*, Advancing mathematics by guiding human intuition with AI. *Nature* **600**, 70–74 (2021).
70. T. J. Sejnowski, Large language models and the reverse Turing test. *Neural Comput.* **35**, 309–342 (2023).
71. P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, K. Kersting, Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **4**, 258–268 (2022).
72. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
73. T. McCoy, E. Pavlick, T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, L. Márquez, Eds. (Association for Computational Linguistics, 2019), pp. 3428–3448.
74. N. McKenna *et al.*, Sources of hallucination by large language models on inference tasks. *arXiv [Preprint]* (2023). <http://arxiv.org/abs/2305.14552> (Accessed 25 November 2023).
75. P. Brookins, J. M. DeBacker, Playing games with GPT: What can we learn about a large language model from canonical strategic games? Available at SSRN 4493398 (2023).
76. G. V. Aher, R. I. Arriaga, A. T. Kalai, "Using large language models to simulate multiple humans and replicate human subject studies" in *International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, Eds. (PMLR, 2023), pp. 337–371.
77. H. A. Simon, Rational decision making in business organizations. *Am. Econ. Rev.* **69**, 493–513 (1979).
78. R. H. Thaler, Behavioral economics: Past, present, and future. *Am. Econ. Rev.* **106**, 1577–1600 (2016).
79. Y. Chen, T. X. Liu, Y. Shan, S. Zhong, Data for "The emergence of economic rationality of GPT." Dropbox. <https://www.dropbox.com/scl/fo/572ptz57vjis5cqkczj9l/h?rlkey=hpsgdb6ghdzsvj35mafndwc&dl=0>. Deposited 6 November 2023.