PNAS RESEARCH ARTICLE | ECONOMIC SCIENCES OPEN ACCESS

A Turing test of whether AI chatbots are behaviorally similar to humans

Qiaozhu Mei[a,1], Yutong Xie[a], Walter Yuan[b], and Matthew O. Jackson[c,d,1]

Contributed by Matthew O. Jackson; received August 12, 2023; accepted January 4, 2024; reviewed by Ming Hsu, Juanjuan Meng, and Arno Riedl

We administer a Turing test to AI chatbots. We examine how chatbots behave in a suite of classic behavioral games that are designed to elicit characteristics such as trust, fairness, risk-aversion, cooperation, etc., as well as how they respond to a traditional Big-5 psychological survey that measures personality traits. ChatGPT-4 exhibits behavioral and personality traits that are statistically indistinguishable from a random human from tens of thousands of human subjects from more than 50 countries. Chatbots also modify their behavior based on previous experience and contexts "as if" they were learning from the interactions and change their behavior in response to different framings of the same strategic situation. Their behaviors are often distinct from average and modal human behaviors, in which case they tend to behave on the more altruistic and cooperative end of the distribution. We estimate that they act as if they are maximizing an average of their own and partner's payoffs.

AI | chatbot | behavioral games | Turing test | personality

Significance

As AI interacts with humans on an increasing array of tasks, it is important to understand how it behaves. Since much of AI programming is proprietary, developing methods of assessing AI by observing its behaviors is essential. We develop a Turing test to assess the behavioral and personality traits exhibited by AI. Beyond administering a personality test, we have ChatGPT variants play games that are benchmarks for assessing traits: trust, fairness, risk-aversion, altruism, and cooperation. Their behaviors fall within the distribution of behaviors of humans and exhibit patterns consistent with learning. When deviating from mean and modal human behaviors, they are more cooperative and altruistic. This is a step in developing assessments of AI as it increasingly influences human experiences.

As Alan Turing foresaw to be inevitable, modern AI has reached the point of emulating humans: holding conversations, providing advice, writing poems, and proving theorems. Turing proposed an intriguing test: whether an interrogator who interacts with an AI and a human can distinguish which one is artificial. Turing called this test the "imitation game" (1), and it has become known as a Turing test.

Advancements in large language models have stirred debate. Discussions range from the potential of AI bots to emulate, assist, or even outperform humans, e.g., writing essays, taking the SAT, writing computer programs, giving economic advice, or developing ideas, (2–5), to their potential impact on labor markets (6) and broader societal implications (7, 8). As some roles for AI involve decision-making and strategic interactions with humans, it is imperative to understand their behavioral tendencies before we entrust them with pilot or co-pilot seats in societal contexts, especially as their development and training are often complex and not transparent (9). Do AIs choose similar actions or strategies as humans, and if not how do they differ? Do they exhibit distinctive personalities and behavioral traits that influence their decisions? Are these strategies and traits consistent across varying contexts? A comprehensive understanding of AI's behavior in generalizable scenarios is vital as we continue to integrate them into our daily lives.

We perform a Turing test of the behavior of a series of AI chatbots. This goes beyond simply asking whether AI can produce an essay that looks like it was written by a human (10) or can answer a set of factual questions, and instead involves assessing its behavioral tendencies and "personality." In particular, we ask variations of ChatGPT to answer psychological survey questions and play a suite of interactive games that have become standards in assessing behavioral tendencies, and for which we have extensive human subject data. Beyond eliciting a "Big Five" personality profile, we have the chatbots play a variety of games that elicit different traits: a dictator game, an ultimatum bargaining game, a trust game, a bomb risk game, a public goods game, and a finitely repeated Prisoner's Dilemma game. Each game is designed to reveal different behavioral tendencies and traits, such as cooperation, trust, reciprocity, altruism, spite, fairness, strategic thinking, and risk aversion. The personality profile survey and the behavioral games are complementary as one measures personality traits and the other behavioral tendencies, which are distinct concepts; e.g., agreeableness is distinct from a tendency to cooperate. Although personality traits are predictive of various behavioral tendencies (11, 12), including both dimensions provides a fuller picture.

In line with Turing's suggested test, we are the human interrogators who compare the ChatGPTs' choices to the choices of tens of thousands of humans who faced the same surveys and game instructions. We say an AI passes the Turing test if its responses cannot be statistically distinguished from randomly selected human responses.

Reviewers: M.H., University of California, Berkeley; J.M., Peking University; and A.R., Universiteit Maastricht.

Competing interest statement: The human game-playing data used were shared from MobLab, a for-profit educational platform. The data availability is an in-kind contribution to all authors, and the data are available for purposes of analysis reproduction and extended analyses. W.Y. is the CEO and Co-founder of MobLab. M.O.J. is the Chief Scientific Advisor of MobLab and Q.M. is a Scientific Advisor to MobLab, positions with no compensation but with ownership stakes. Y.X. has no competing interest.

[1]To whom correspondence may be addressed. Email: qmei@umich.edu or jacksonm@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2313925121/-/DCSupplemental.

Published February 22, 2024.

We find that the chatbots' behaviors are generally within the support of those of humans, with only a few exceptions. Their behavior is more concentrated than the full distribution of humans. However, we are comparing two chatbots to tens of thousands of humans, and so a chatbot's variation is within subject and the variation in the human distribution is across subjects. The chatbot variation may be similar to what a single individual would exhibit if repeatedly queried. We do an explicit Turing test by comparing an AI's behavior to a randomly selected human behavior, and ask which is the most likely to be human based on a conditional probability calculation from the data. The behaviors are generally indistinguishable, and ChatGPT-4 actually outperforms humans on average, while the reverse is true for ChatGPT-3. There are several games in which the AI behavior is picked more likely to be human most of the time, and others where it is not. When they do differ, the chatbots' behaviors tend to be more cooperative and altruistic than the median human, including being more trusting, generous, and reciprocating.

In that vein, we do a revealed-preference analysis in which we examine the objective function that best predicts AI behavior. We find that it is an even average of own and partner's payoffs. That is, they act as if they are maximizing the total payoff of both players rather than simply their own payoff. Human behavior also is optimized with some weight on the other player, but the weight depends on the preference specification and humans are more heterogeneous and less well predicted.

There are two other dimensions on which we compare AI and human behavior. The first is whether context and framing matter, as they do with humans. For example, when we ask them to explain their choices or tell them that their choices will be observed by a third party, they become significantly more generous. Their behavior also changes if we suggest that they act as if they were faced with a partner of a gender, or that they act as if they were a mathematician, legislator, etc. The second dimension is that humans change their behaviors after they have experienced different roles in a game. The chatbots also exhibit significant changes in behaviors as they experience different roles in a game. That is, once they have experienced the role of a "partner" in an asymmetric game, such as a trust game or an ultimatum game, their behavior shifts significantly.

Finally, it is worth noting that we observe behavioral differences between the versions of ChatGPT that we test, so that they exhibit different personalities and behavioral traits.

1. Methods and the Turing Test Design

We conduct interactive sessions, prompting AI chatbots to participate in classic behavioral economics games and respond to survey questions using the same instructions as given to human subjects. We compare how the chatbots behave to how humans behave and also estimate which payoff function best predicts the chatbots' behaviors.

We examine the widely-used AI chatbot: ChatGPT developed by OpenAI. We primarily evaluate two specific versions of ChatGPT: the API version tagged as GPT-3.5-Turbo (referred to as ChatGPT-3) and the API version based on GPT-4 (denoted as ChatGPT-4). We also include the subscription-based Web version (Plus), and the freely available Web version (Free) for comparison (see *SI Appendix*, section 1.A for more description of the chatbots).

The human subject data are derived from a public Big Five Test response database and the MobLab Classroom economics experiment platform, both spanning multiple years and more than 50 countries, encompassing 108,314 subjects (19,719 for the Big Five Test, and 88,595 for the behavioral economics games, who are mostly college and high school students). Details about the human datasets, including the demographics of the subjects are included in the (*SI Appendix*, section 1.B; and see also Lin et al. (13) who provide additional background details about the human data which cover North America, Europe, and Asia).

We administer the OCEAN Big Five questionnaire to each chatbot to create a personality profile. Following this, we ask each chatbot what actions they would choose in a suite of six games designed to illuminate various behavioral traits (and fuller details appear in *SI Appendix*, section 1.A):

(i) A Dictator Game—given an endowment of money, one player (the dictator) chooses how much of the money to keep and how much to donate to a second player. This involves altruism (14, 15).
(ii) An Ultimatum Game—given an endowment of money, one player (the proposer) offers a split of the money to a second player (the responder) who either accepts the split or rejects it in which case neither player gets anything. This involves fairness and spite (14).
(iii) A Trust Game—given an endowment of money, one player (the investor) decides how much of the money to keep and passes the remainder to a second player (the banker), which is then tripled. The banker decides how much of that tripled revenue to keep and returns the remainder to the investor. This involves trust, fairness, altruism, and reciprocity (16).
(iv) A Bomb Risk Game—a player chooses how many boxes out of 100 to open and the player is rewarded for each opened box but loses everything if a randomly placed bomb is encountered. This involves risk aversion (17).
(v) A Public Goods Game—given an endowment of money, a player chooses how much of the money to keep and how much to contribute to a public good and receives half of the total amount donated to the public good by all four players. This involves free-riding, altruism, and cooperation (18).
(vi) A finitely repeated Prisoners Dilemma Game—in each of five periods two players simultaneously choose whether to "cooperate" or "defect," yielding the highest combined payoff if both cooperate but with one player getting a better payoff if they defect. This involves cooperation, reciprocity, and strategic reasoning (19–22).

Each chatbot answers each survey question and plays each role in each game 30 times in individual sessions. As we cannot pay the chatbots, we ask how they would behave in each role in each game. Details about how the chatbots' responses are collected can be found in *SI Appendix*, section 1.A.

2. Results

A. Personality Profiles of the AIs. Fig. 1 provides a summary of the chatbots' Big Five personality profiles and compares them with the human distribution. We illustrate the behaviors of ChatGPT-3 and ChatGPT-4 specifically, as the Free Web version exhibits similarities to ChatGPT-3,* and the Plus version aligns closely with ChatGPT-4.† More detailed results, including those of the two Web-based versions, can be found in *SI Appendix*, section 3.

*https://openai.com/blog/chatgpt, retrieved 08/02/2023.
†https://openai.com/gpt-4, retrieved 08/02/2023.

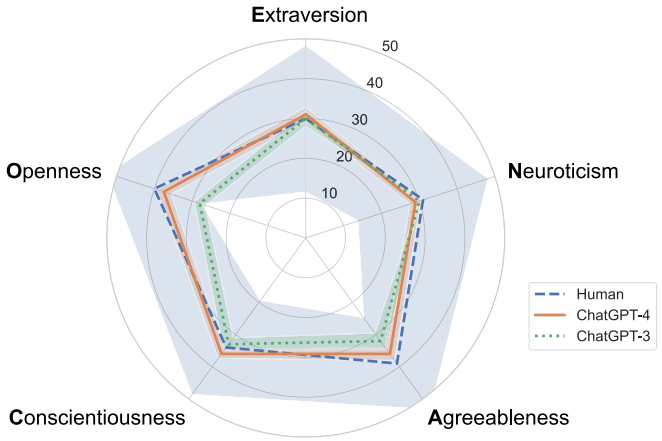

Fig. 1. "Big Five" personality profiles of ChatGPT-4 and ChatGPT-3 compared with the distributions of human subjects. The blue, orange, and green lines correspond to the median scores of humans, ChatGPT-4, and ChatGPT-3 respectively; the shaded areas represent the middle 95% of the scores, across each of the dimensions. ChatGPT's personality profiles are within the range of the human distribution, even though ChatGPT-3 scored noticeably lower in Openness.

The personality traits of ChatGPT-3 and ChatGPT-4, as derived from their responses to the OCEAN Big Five questionnaire, are depicted in Fig. 1. Comparing humans and chatbots, ChatGPT-4 exhibits substantial similarity to the human respondents across all five dimensions in terms of the median scores. ChatGPT-3 likewise demonstrates comparable patterns in four dimensions but displays a relatively lower score in the dimension of openness. Particularly, on extroversion, both chatbots score similarly to the median human respondents, with ChatGPT-4 and ChatGPT-3 scoring higher than 53.4% and 49.4% of human respondents, respectively. On neuroticism, both chatbots exhibit moderately lower scores than the median human. Specifically, ChatGPT-4 and ChatGPT-3 score higher than 41.3% and 45.4% of humans, respectively. As for agreeableness, both chatbots show lower scores than the median human, with ChatGPT-4 and ChatGPT-3 surpassing 32.4% and 17.2% of humans, respectively. While for conscientiousness, both chatbots fluctuate around the median human, with ChatGPT-4 and ChatGPT-3 scoring higher than 62.7% and 47.1% of human respondents. Both chatbots exhibit lower openness than the median human, with ChatGPT-3's being notably lower. On this dimension, ChatGPT-4 and ChatGPT-3 score higher than 37.9% and 5.0% of humans, respectively.

When comparing the two chatbots, we find that ChatGPT-4 has higher agreeableness, higher conscientiousness, higher openness, slightly higher extraversion, and slightly lower neuroticism than ChatGPT-3, consistent with each chatbot having a distinct personality.

B. The Games and the Turing Test. We perform a formal Turing test as follows. Consider a game and role, for instance, the giver in the Dictator Game. We randomly pick one action from the chatbot's distribution and one action from the human distribution. We then ask, which action "looks more typically human?" Specifically, we ask which of the two actions is more likely under the human distribution. If AI picks an action that is very rare under the human distribution then it is likely to lose in the sense that the human's play will often be estimated to be more likely under the human distribution. If AI picks the modal human action then it will either be estimated as being more likely under the human distribution or else tie.[‡]

The results appear in Fig. 2. As a benchmark, we also report what happens when two humans are matched against each other. In that case, there should be equal wins and losses (up to variations due to taking only 10,000 draws). We see that overall (on average) ChatGPT-4 is actually picked as human or ties significantly more often than a random human, while ChatGPT-3 is picked as human less often than a random human. In this particular sense, ChatGPT-4 would pass this Turing test, while ChatGPT-3 would fail it.

The results vary nontrivially across games. ChatGPT-4 does better than or comparably to humans in all games except in the Prisoner's Dilemma (where it cooperates most of the time and the human mode is to defect) and as the Investor role in the Trust Game (in which it generally invests half while humans tend to be more extreme one way or the other). ChatGPT-3 does well in a few games, but is outperformed by humans in 6 of the 8 games, and overall.

C. Comparisons of ChatGPTs' Behaviors to Humans' on a Variety of Dimensions. We also look at distributions of behaviors in more detail across games by comparing the distribution of an AI's responses to the distribution of human responses. Note that a human distribution is mostly obtained from one observation per human, so its variation is between subjects. Variation in an AI distribution is obtained from the same chatbot, so it is within subject. Thus, the fact that the distributions differ is not informative, but the following information about the distributions is useful to note.

Human players' actions generally exhibit multiple peaks and nontrivial variance, indicating the presence of varied behavioral patterns across the population. In most games, the responses of ChatGPT-4 and ChatGPT-3 are not deterministic when the same games are repeated (except for ChatGPT-4 in the Dictator game and in the Ultimatum Game as the proposer) and adhere to certain distributions. Typically, the distributions produced by the chatbots encompass a subset of the modes observed in the corresponding human distributions. As illustrated in Fig. 3, ChatGPT-3 makes decisions that result in usually single-mode, and moderately skewed distributions with nontrivial variance. Conversely, ChatGPT-4's decisions form more concentrated distributions.

Next, we examine in more detail some of the behavioral traits that have been associated with the suite of games we use.

C.1. Altruism. In games that involve distributional concerns, the chatbots act more generously to the other player than the human median. In particular, they display increased generosity: in the Dictator Game (Fig. 3*A*), as the proposer in the Ultimatum Game (Fig. 3*B*), and as the banker in the Trust Game (Fig. 3*E*), and as a contributor in the Public Goods Game (Fig. 3*F*). Note that from the perspective of maximizing the player's own payoff, the most beneficial strategies would be to give \$0 to the other player in the Dictator Game, return \$0 to the investor as the banker in the Trust Game, and contribute \$0 in the Public Goods

[‡]Alternatively, one instead could also use the AI distribution and do relative Bayesian updating, and assign posterior probabilities of being human vs. AI taking into account the action's relative likelihood under each of the distributions. That is less in the spirit of what Turing described as it would require the interrogator to have knowledge about the AI behavior, but also an interesting question. In a case in which AI plays a tighter distribution, even if the modal human action, such Bayesian updating would pick out AI more often. For example, if AI always plays the modal human action and humans vary their action, then in our test AI always wins or ties, while under Bayesian updating with precise knowledge of AI behavior it would always lose.

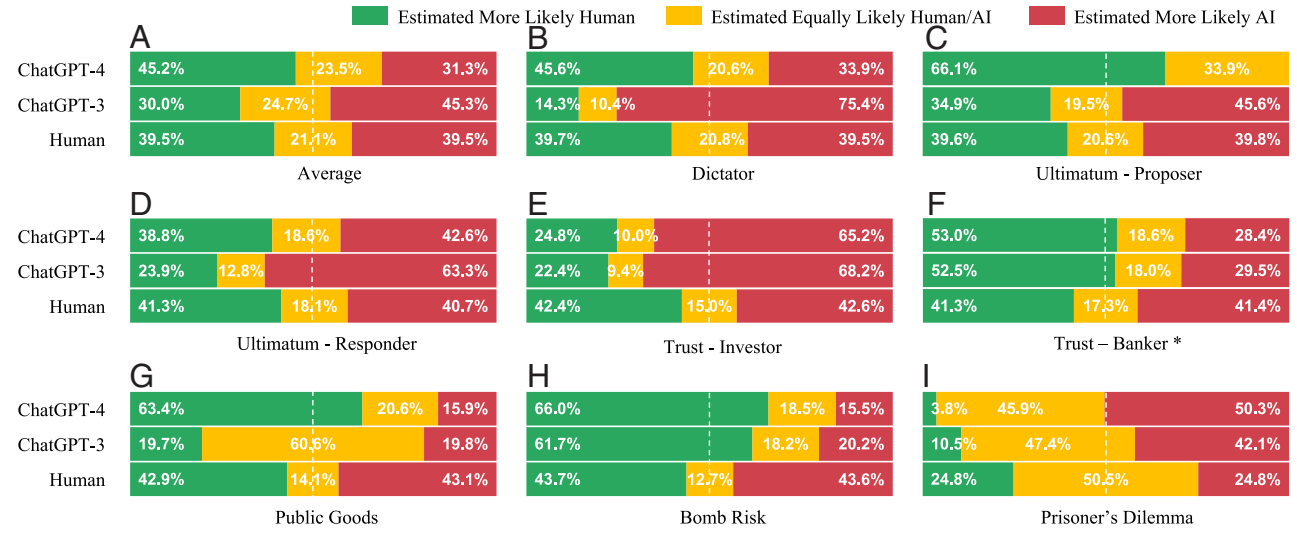

Fig. 2. The Turing test. We compare a random play of Player A (ChatGPT-4, ChatGPT-3, or a human player, respectively) and a random play of a second Player B (which is sampled randomly from the human population). We compare which action is more typical of the human distribution: which one would be more likely under the human distribution of play. The green bar indicates how frequently Player A's action is more likely under the human distribution than Player B's action, while the red bar is the reverse, and the yellow indicates that they are equally likely (usually the same action). (*A*): average across all games; (*B–I*): results in individual games. ChatGPT-4 is picked as more likely to be human more often than humans in 5/8 of the games, and on average across all games. ChatGPT-3 is picked as or more likely to be human more often than humans in 2/8 of the games and not on average.

Game. Even though these strategies are chosen by a significant portion of human players, they were never chosen by the chatbots.

ChatGPT's decisions are consistent with some forms of altruism, fairness, empathy, and reciprocity rather than maximization of its personal payoff. To explore this in more detail, we calculate the own 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's 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 *SI Appendix*, Table S1.

In particular, ChatGPT-4 obtains a higher own payoff than human players in the Ultimatum Game and a lower own payoff in all other games. In all seven games, it yields a higher partner payoff. Moreover, it achieves the highest combined payoff for both players in five out of seven games, the exceptions being the Dictator game and the Trust Game as the banker (where the combined payoffs are constant).

These findings are indicative of ChatGPT-4's increased level of altruism and cooperation compared to the human player distribution. ChatGPT-3 has a more mixed payoff pattern. For example, although it yields a lower own payoff in the Trust Game and the Public Goods Game compared to ChatGPT-4, it achieves the highest partner payoff and combined payoff in the Public Goods Game, as well as the highest partner payoff in the Trust game as the banker.

C.2. Fairness. ChatGPT-3 typically proposes a more favorable deal to the other player in games where the outcome depends on the other player's approval (i.e., in the Ultimatum Game) compared to when it does not (i.e., in the Dictator Game), mirroring behavior observed in the human data. In contrast, ChatGPT-4 consistently prioritizes fairness in its decision-making process. This is evidenced by its equal split of the endowment, whether acting as a dictator or as a proposer in the Ultimatum Game, particularly when asked to explain its decisions (*SI Appendix*, Fig. S6*A*).

In the Ultimatum Game as responder, less than a fifth of human players are willing to accept as low as $1 as proposed by the other player (Fig. 3*C*), despite this being the dominant strategy for the responder in the game. Interestingly, this forms the most common response of ChatGPT-4. However, there is another peak at $50, which is close to the modal human response and corresponds to the fair split.

C.3. Trust. Generally speaking, ChatGPT-4 displays more "trust" in the banker (the first/other player) compared to ChatGPT-3, by investing a higher proportion of the endowment, as shown in Fig. 3*D*. This is more trust than exhibited by humans, except for a group that invests their entire endowment. Both chatbots also tend to invest more in public goods projects than human players, as shown in Fig. 3*F*.

C.4. Cooperation. ChatGPT's first action is most often cooperative in the Prisoner's Dilemma Game (Fig. 3*H*). In particular, ChatGPT-4's strategy in the first round is substantially more cooperative than human players, with a large majority (91.7%) of sessions opting to cooperate, as opposed to 45.1% of human players. ChatGPT-3's strategy lies somewhere in between, with 76.7% choosing to cooperate. Both ChatGPT-3 and ChatGPT-4 are also more cooperative than human players in the Public Goods Game (Fig. 3*F*).

C.5. Tit-for-Tat. While chatbots exhibit a higher cooperative tendency in the Prisoner's Dilemma Game than the typical human subject, their cooperation is not unconditional. As shown in Fig. 4*A*, if the other player cooperates in the first round, ChatGPT-4's decision remains consistent in the following round. On the other hand, around half of the ChatGPT-3's sessions that chose defection in the first round switched to cooperation in the second round. A small proportion of the cooperative sessions also switch to defection, mimicking similar behavior observed among human subjects.

When the other player defects in the first round, however, all previously cooperative sessions of ChatGPT-4 switch to

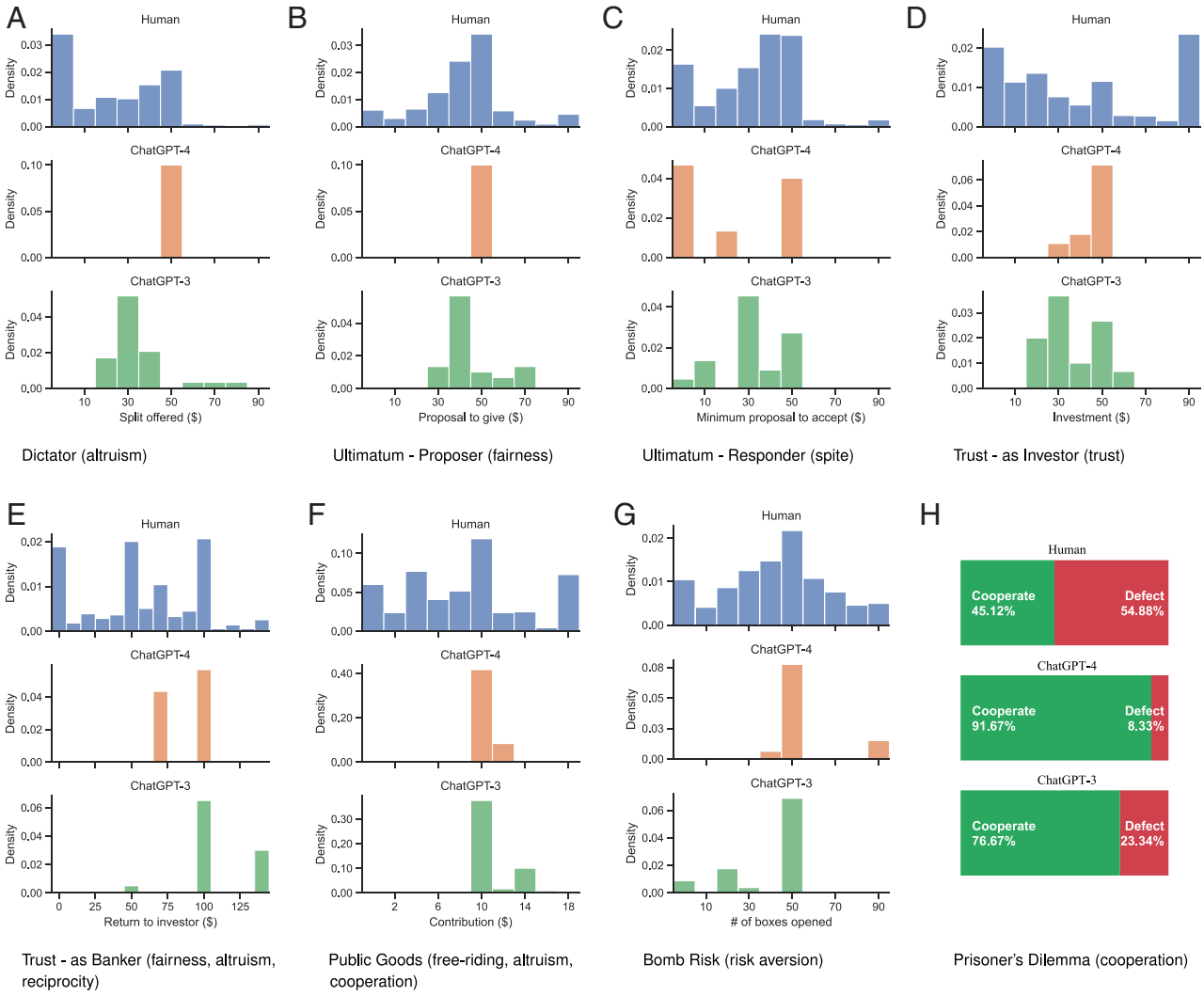

Fig. 3. Distributions of choices of ChatGPT-4, ChatGPT-3, and human subjects in each game: (*A*) Dictator; (*B*) Ultimatum as proposer; (*C*) Ultimatum as responder; (*D*) Trust as investor; (*E*) Trust as banker; (*F*) Public Goods; (*G*) Bomb Risk; (*H*) Prisoner's Dilemma. Both chatbots' distributions are more tightly clustered and contained within the range of the human distribution. ChatGPT-4 makes more concentrated decisions than ChatGPT-3. Compared to the human distribution, on average, the AIs make a more generous split to the other player as a dictator, as the proposer in the Ultimatum Game, and as the Banker in the Trust Game, on average. ChatGPT-4 proposes a strictly equal split of the endowment both as a dictator or as the proposer in the Ultimatum Game. Both AIs make a larger investment in the Trust Game and a larger contribution to the Public Goods project, on average. They are more likely to cooperate with the other player in the first round of the Prisoner's Dilemma Game. Both AIs predominantly make a payoff-maximization decision in a single-round Bomb Risk Game. Density is the normalized count such that the total area of the histogram equals 1.

defection, showcasing a play that would be similar to a "Tit-for-Tat" pattern as illustrated in Fig. 4*B*. This pattern is also observed in human players and ChatGPT-3, although to a lesser but still majority extent. There are additional dynamics for further study in repeated game settings, as the chatbots often revert to cooperation even if the other player continues to defect (*SI Appendix*, Fig. S8).

C.6. Risk aversion. The chatbots also differ in their exhibited risk preferences. In the Bomb Risk Game (Fig. 5), both ChatGPT-3 and ChatGPT-4 predominantly opt for the expected payoff-maximizing decision of opening 50 boxes. This contrasts with the more varied human decisions, which include a distinct group of extreme subjects who only open one box.

Interestingly, the chatbots' decisions in this game are influenced by the outcomes of previous rounds, despite their independence. If the bomb exploded in a prior round, ChatGPT-3 tends to adopt a more risk-averse behavior by opting to open fewer boxes—a trend mirrored, to a lesser extent, in human data. Meanwhile, the preferred decision of ChatGPT-4 remains constant, albeit with higher variance.

In instances where the bomb did not explode, the decisions of both ChatGPT-3 and ChatGPT-4 converge and revert to the expected payoff-maximizing option. Overall, ChatGPT-4 displays a consistent and neutral risk preference. ChatGPT-3, however, tends toward risk aversion, especially in unexpected contexts—a pattern that is also observed when it acts as the investor in the Trust Game, where it makes the lowest investment on average.

D. Revealed-Preferences. Given the observations above, especially regarding fairness, cooperation, and altruism, we perform a systematic analysis by inferring which preferences would rationalize the AIs' behaviors. This enables one to make predictions out of sample, and so we estimate an objective function that best predicts AI behavior. In particular, just as is done with humans, we estimate which utility function predicts decisions as if it were being maximized. This can then be used in future analyses to predict AI behavior in new settings.

First, we consider a utility function that is a weighted average of the two players' payoffs:

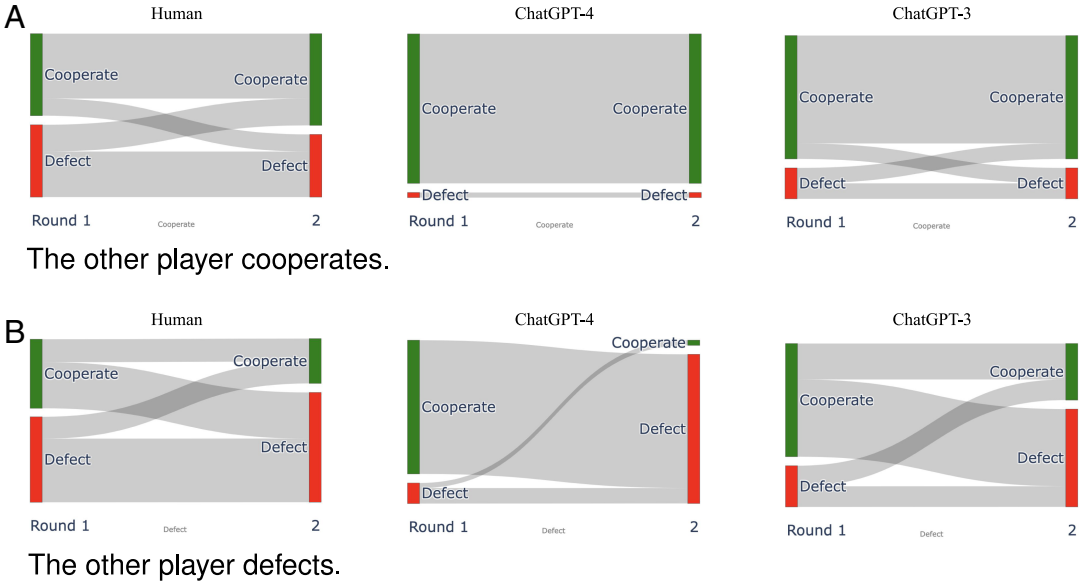

Fig. 4. ChatGPT's dynamic play in the Prisoner's Dilemma Game. ChatGPT-4 exhibits a higher tendency to cooperate compared to ChatGPT-3, which is significantly more cooperative than human players. The tendency persists when the other player cooperates. On the other hand, both chatbots apply a one-round Tit-for-Tat strategy when the other player defects. The other player's (first round) choice is observed after Round 1 play and before Round 2 play: (*A*) the other player cooperates; (*B*) the other player defects.

$$b \times \text{Own Payoff} + (1 - b) \times \text{Partner Payoff},$$

for some $b \in [0, 1]$. Purely selfish preferences correspond to $b = 1$ and purely selfless-altruistic preferences correspond to $b = 0$, and maximizing the total payoff of both players corresponds to $b = 1/2$.

We estimate which b best predicts behavior. Consider the distribution of play from the human distribution. Given that distribution of partner play, for every $b \in [0, 1]$ there is a best-response payoff: the best possible expected utility that the player could earn across actions if their utility function was described by b. Then we can see what action they choose, and see what fraction of that best possible expected payoff they earn when that is matched against the human distribution of partner play. The difference (in proportional terms) is the error. We average that squared error across the distribution of actions that the chatbot (or human) plays in that game. We then look at the average squared error across all plays, and select the $b \in [0, 1]$ that minimizes that mean-squared error. The results as a function of b are reported in Fig. 6.

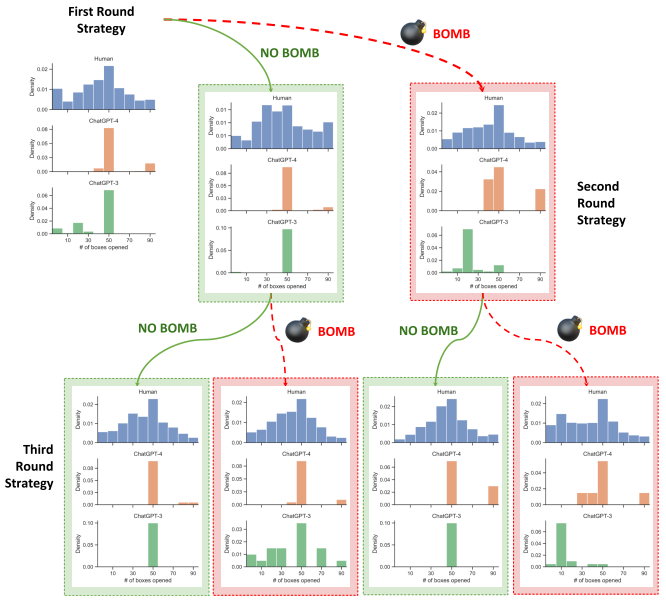

Fig. 5. ChatGPT-4 and ChatGPT-3 act as if they have particular risk preferences. Both have the same mode as human distribution in the first round or when experiencing favorable outcomes in the Bomb Risk Game. When experiencing negative outcomes, ChatGPT-4 remains consistent and risk-neutral, while ChatGPT-3 acts as if it becomes more risk-averse.

For the linear specification (above), the errors for both the chatbots are minimized at $b = 0.5$, and those for humans are minimized at a nearby point $b = 0.6$ (see *SI Appendix*, section 2.B for per-game estimates). ChatGPT-4's behavior exhibits the smallest error in that case, while the humans' behavior is the most varied, exhibits the highest errors, and is the least well-predicted of the three by $b = 0.5$.

The estimated b varies across games, with the best fit being selfish ($b = 1.0$) in the Ultimatum Game, but being centered around $b = 0.5$ in the other games (*SI Appendix*, Fig. S4). We also perform a multinomial logistic discrete choice analysis and estimate the best fitting b's by each game and find similar results (*SI Appendix*, Table S2).

We also note that a linear specification does not fully capture preferences for relative payoffs as, for example, when $b = 0.5$ how a total payoff is allocated is inconsequential. Instead, if one works with a constant elasticity of substitution (CES) utility function (23) of the form

$$\left(b \times (\text{Own Payoff})^{1/2} + (1 - b) \times (\text{Partner Payoff})^{1/2}\right)^2,$$

then relative allocations across the two players are more distinguished. For this specification, we see the human error curve shift to have the weight that minimizes errors be more selfish, and we see more distinction between all three of ChatGPT-4,

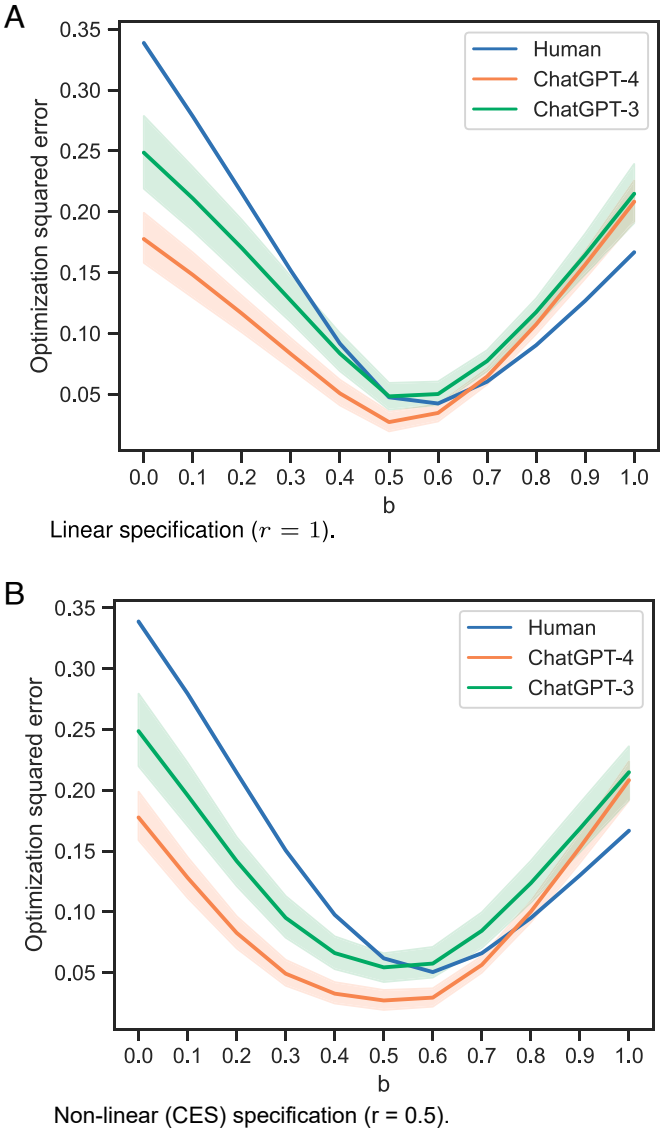

Fig. 6. Mean squared error of the actual distribution of play relative to the best-response payoff, when matched with a partner playing the human distribution for possible preferences indexed by b. The average is across all games. The errors are plotted for each possible b, the weight on own vs partner payoff in the utility function. $b = 1$ is the purely selfish (own) payoff, $b = 0$ is the purely selfless/altruistic (partner) payoff, and $b = 0.5$ is the overall welfare (average) payoff, and other bs are weighted averages of own and partner payoffs. Both chatbots' behaviors are best predicted by $b = 0.5$, and those of humans are best predicted by $b = 0.6$; they best predict ChatGPT-4's behavior and have higher errors in the other cases. (*A*) The *Top* panel is for utility = $b \times$ Own Payoff $+ (1-b) \times$ Partner Payoff. (*B*) The *Bottom* panel is for CES preferences: utility $= \left(b \times (\text{Own Payoff})^{1/2} + (1-b) \times (\text{Partner Payoff})^{1/2}\right)^2$.

ChatGPT-3, and the humans. This also carries over game by game as shown in *SI Appendix*, Fig. S5.

E. Framing and Context. Human behavior can be significantly altered by framing (e.g., ref. 24). We examine whether AI behavior also varies with how a given strategic setting is framed. We find, that 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.

SI Appendix has detailed prompts used for framing the AI (*SI Appendix*, section 1.A.3), and it also presents distributions of behaviors (*SI Appendix*, section 2.C). Here are some examples of how the framing matters.

When ChatGPT-3 is asked to explicitly explain its decision or when it is aware that the Dictator Game is witnessed by a third-party observer (a game host), it demonstrates significantly greater generosity as the dictator (*SI Appendix*, Fig. S6*A*).

In the Ultimatum Game, when ChatGPT-4 is made aware of the gender of the proposer (regardless of what it is), its decision as the responder moves away from the dominant strategy of accepting any proposal and starts demanding higher splits on average (*SI Appendix*, Fig. S6*B*), even though we do not observe a specific gender effect.

In the Trust Game (*SI Appendix*, Fig. S6 *D–F*), as the size of the potential investment is increased, ChatGPT-4's strategy as the banker shifts from returning the original investment plus an even split of the profit (which equals a doubled investment) to evenly splitting the entire revenue (which is a tripled investment). By contrast, ChatGPT-3 tends to make a more generous return to the investor when the potential investment is larger.

ChatGPT's decisions are also impacted when they are asked to play the games as if they are from a given occupation, altering their default role as a helpful assistant (*SI Appendix*, section 1.A.3). For instance, in the Ultimatum Game as the responder (*SI Appendix*, Fig. S6*C*), when ChatGPT-4 is prompted to play as a mathematician, its decision shifts toward the dominant strategy, agreeing to accept as low as $1 in most cases. Conversely, when prompted to be a legislator, its decisions align more with what is traditionally considered "fair": demanding $50 in the majority of cases.

F. Learning. One last thing we investigate is the extent to which the chatbots' behaviors change as they gain experience in different roles in a game, *as if* they were learning from such experience. This is something that is true of humans (e.g., ref. 25).

In games with multiple roles (such as the Ultimatum Game and the Trust Game), the AIs' decisions can be influenced by previous exposure to another role. For instance, if ChatGPT-3 has previously acted as the responder in the Ultimatum Game, it tends to propose a higher offer when it later plays as the proposer, while ChatGPT-4's proposal remains unchanged (*SI Appendix*, Fig. S7*A*). Conversely, when ChatGPT-4 has previously been the proposer, it tends to request a smaller split as the responder (*SI Appendix*, Fig. S7*B*).

Playing the banker's role in the Trust Game, especially when the investment is large, also influences ChatGPT-4 and ChatGPT-3's subsequent decisions as the investor, leading them to invest more (*SI Appendix*, Fig. S7*C*). Similarly, having played the investor first also influences the AIs' subsequent decisions as the banker, resulting in both ChatGPT-3 and ChatGPT-4 returning more to the investor (*SI Appendix*, Fig. S7*D*).

Our analyses of learning and framing are far from systematic, and it would be interesting to compare how the effects of context change AI behavior to how context changes human behavior. For example, it would be interesting to see how chatbots act when asked to assume the role of a specific gender, demographic group, or personality profile.

3. Discussion

We have sidestepped the question of whether artificial intelligence can think (26–28), which was a central point of Turing's original essay (1), but we have performed a test along the lines of what

he suggested. We have found that AI and human behavior are remarkably similar. Moreover, not only does AI's behavior sit within the human subject distribution in most games and questions, but it also exhibits signs of human-like complex behavior such as learning and changes in behavior from role-playing. On the optimistic side, when AI deviates from human behavior, the deviations are in a positive direction: acting as if it is more altruistic and cooperative. This may make AI well-suited for roles necessitating negotiation, dispute resolution, or caregiving, and may fulfill the dream of producing AI that is "more human than human." This makes them potentially valuable in sectors such as conflict resolution, customer service, and healthcare.

The observation that ChatGPT's, especially ChatGPT-4's, behavior is more concentrated and consistent evokes both optimism and apprehension. This is similar to what might happen if a single human were compared to the population. However, the chatbots are used in technologies that interact with huge numbers of others and so this narrowness has consequences. Positively, its rationality and constancy make AI highly attractive for various decision-making contexts and make it more stable and predictable. However, this also raises concerns regarding the potential loss of diversity in personalities and strategies (compared to the human population), especially when put into new settings and making important new decisions.

Our work establishes a straightforward yet effective framework and benchmark for evaluating chatbots and other AI as they are rapidly evolving. This may pave the way for a new field in AI behavioral assessment. The AI that we tested was not necessarily programmed to pass this sort of Turing test, and so that raises the question of when and how AI that is designed to converse with and inform humans, and is trained on human-generated data, necessarily behaves human-like more broadly. That could also help in advancing our understanding of why humans exhibit certain traits. Most importantly, the future will tell the extent to which AI enhances humans rather than substituting for them (29, 30).

In terms of limitations, given that our human data are collected from students, it is important to expand the reference population in further analyses. The games we have chosen are prominent ones, but one can imagine expanding the suite of analyses included in a Turing test, and also tailoring such tests to the specific tasks that are entrusted to different versions of AI. In addition, the chatbots tested here are just one of a growing number, and a snapshot at a specific point in time of a rapidly evolving form of AI. Thus, the results should not be taken as broadly representative, but instead should be taken as illustrative of a testing approach and what can be learned about particular instances of AI.

Data, Materials, and Software Availability. Processed data files, ChatGPT responses, and code have been deposited in Github (https://github.com/yutxie/ChatGPT-Behavioral) (31).

ACKNOWLEDGMENTS. The reviewers inspired us to include additional analyses that have strengthened the paper. We also thank ZhiHong Jian, research economist at MobLab, for his essential support in organizing the MobLab data from human players. This research was deemed not regulated by the University of Michigan IRB (HUM00232017).

Author affiliations: [a]School of Information, University of Michigan, Ann Arbor, MI 48109; [b]MobLab, Pasadena, CA 91107; [c]Department of Economics, Stanford University, Stanford, CA 94305; and [d]External Faculty, Santa Fe Institute, Santa Fe, NM 87501

Author contributions: Q.M., Y.X., W.Y., and M.O.J. designed research; Q.M., Y.X., and W.Y. performed research; Q.M., Y.X., and M.O.J. analyzed data; and Q.M., Y.X., W.Y., and M.O.J. wrote the paper.

1. A. M. Turing, Computing machinery and intelligence. *MIND: Quart. Rev. Psychol. Philos.* **54**, 433–460 (1950).
2. K. Warwick, *Turing Test Success Marks Milestone in Computing History* (University or Reading Press Release, 2014), p. 8.
3. S. Bubeck *et al*., Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv [Preprint] (2023). http://arxiv.org/abs/2303.12712 (Accessed 28 December 2023).
4. K. Girotra, L. Meincke, C. Terwiesch, K. T. Ulrich, Ideas are dimes a dozen: large language models for idea generation in innovation. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4526071. Accessed 28 December 2023.
5. Y. Chen, T. X. Liu, Y. Shan, S. Zhong, The emergence of economic rationality of GPT. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2316205120 (2023).
6. T. Eloundou, S. Manning, P. Mishkin, D. Rock, GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv [Preprint] (2023). http://arxiv.org/abs/2303.10130 (Accessed 28 December 2023).
7. P. Lee, S. Bubeck, J. Petro, Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New Engl. J. Med.* **388**, 1233–1239 (2023).
8. L. J. Trautman, W. G. Voss, S. Shackelford, How we learned to stop worrying and love AI: Analyzing the rapid evolution of generative pre-trained transformer (GPT) and its impacts on law, business, and society. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4516154. Accessed 28 December 2023.
9. R. Bommasani *et al*., The foundation model transparency index. arXiv [Preprint] (2023). http://arxiv.org/abs/2310.12941 (Accessed 28 December 2023).
10. K. Elkins, J. Chun, Can GPT-3 pass a writer's Turing Test? *J. Cult. Analyt.* **5** (2020).
11. B. W. Roberts, Back to the future: Personality and assessment and personality development. *J. Res. Person.* **43**, 137–145 (2009).
12. M. Almlund, A. L. Duckworth, J. J. Heckman, T. Kautz, "Personality psychology and economics" in *Handbook of the Economics of Education*, E. A. Hanushek, S. J. Machin, L. Woessmann, Eds. (Elsevier, Amsterdam, 2011), pp. 1–181.
13. P. H. Lin *et al*., Evidence of general economic principles of bargaining and trade from 2,000 classroom experiments. *Nat. Hum. Behav.* **4**, 917–927 (2020).
14. W. Güth, R. Schmittberger, B. Schwarze, An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Organ.* **3**, 367–388 (1982).
15. R. Forsythe, J. L. Horowitz, N. E. Savin, M. Sefton, Fairness in simple bargaining experiments. *Games Econ. Behav.* **6**, 347–369 (1994).
16. J. Berg, J. Dickhaut, K. McCabe, Trust, reciprocity, and social history. *Games Econ. Behav.* **10**, 122–142 (1995).
17. P. Crosetto, A. Filippin, The "bomb" risk elicitation task. *J. Risk Uncert.* **47**, 31–65 (2013).
18. J. Andreoni, Cooperation in public-goods experiments: Kindness or confusion? *Am. Econ. Rev.* **85**, 891–904 (1995).
19. J. Von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, ed. 2, 1947).
20. T. C. Schelling, The strategy of conflict. Prospectus for a reorientation of game theory. *J. Conf. Res.* **2**, 203–264 (1958).
21. A. Rapoport, A. M. Chammah, *Prisoner's Dilemma: A Study in Conflict and Cooperation* (University of Michigan Press, 1965), vol. 165.
22. J. Andreoni, H. Varian, Preplay contracting in the prisoners' dilemma. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10933–10938 (1999).
23. D. McFadden, Constant elasticity of substitution production functions. *Rev. Econ. Stud.* **30**, 73–83 (1963).
24. A. Tversky, D. Kahneman, "Rational choice and the framing of decisions" in *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers*, B. Karpak, S. Zionts, Eds. (Springer, Berlin, Heidelberg, 1989), pp. 81–126.
25. V. Benndorf, C. Moellers, H. T. Normann, Experienced vs. inexperienced participants in the lab: Do they behave differently? *J. Econ. Sci. Assoc.* **3**, 12–25 (2017).
26. M. Mitchell, How do we know how smart AI systems are? *Science* **381**, eadj5957 (2023).
27. P. Butlin *et al*., Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv [Preprint] (2023). http://arxiv.org/abs/2308.08708 (Accessed 28 December 2023).
28. N. Shapira *et al*., Clever hans or neural theory of mind? Stress testing social reasoning in large language models. arXiv [Preprint] (2023). http://arxiv.org/abs/2305.14763 (Accessed 28 December 2023).
29. I. Rahwan *et al*., Machine behaviour. *Nature* **568**, 477–486 (2019).
30. E. Brynjolfsson, The Turing trap: The promise & peril of human-like artificial intelligence. *Daedalus* **151**, 272–287 (2022).
31. Q. Mei, Y. Xie, W. Yuan, M. O. Jackson, Data and code for "A Turing test of whether AI chatbots are behaviorally similar to humans." GitHub. https://github.com/yutxie/ChatGPT-Behavioral. Accessed 28 December 2023.