When Machines Think Fast: A Behavioral Economics Test of AI

Keith Jacks Gamble[1*] and Patricia J. Hummel[2]

[1*]Jones College of Business, Middle Tennessee State University.
[2]School of Business, Lincoln Memorial University.

*Corresponding author(s). E-mail(s): Keith.Gamble@mtsu.edu;
Contributing authors: pjhummel01@gmail.com;

Abstract

We test whether large language models reproduce or correct for fourteen well-documented human behavioral biases. We administer each canonical experiment to ChatGPT and construct a pseudo-panel of responses from each iteration. The biases we examine span belief-formation tasks (anchoring, conjunction fallacy), preference-based tasks (framing effects, prospect-theoretic risk attitudes, present bias, mental accounting, sunk-cost reasoning, the endowment effect, the disposition effect, and naïve diversification), and intertemporal and social-choice problems (myopic loss aversion, save-more-tomorrow commitments, the house money effect, and the ultimatum game). On eight of fourteen tasks, the model behaves in a manner consistent with the rational benchmark. On the remaining six tasks (framing of the Asian disease problem, the conjunction fallacy, mental accounting, the house money effect, naïve diversification, and the endowment effect), the model exhibits a directional bias that mirrors the human pattern, but in every case the magnitude of the deviation is smaller than the human benchmark. We interpret these results as evidence that large language models can serve as effective debiasing decision aids in many settings. We show that residual biases inherited from training on human-generated text concentrate predictably in tasks involving valuation, framing, and intuitive judgments about likelihood.

Keywords: Behavioral economics, Generative AI, Large language models, Decision-making, Financial advice

JEL Classification: D03 , D81 , D91 , G41 , C91

1 Introduction

Few empirical regularities in economics have proven as robust, or as inconvenient, as the systematic deviations from rational choice catalogued by Tversky and Kahneman (1974), Kahneman and Tversky (1979), and Tversky and Kahneman (1981) and the long line of experimental work that followed. Framing effects, anchoring, the conjunction fallacy, mental accounting, the endowment effect, present bias, and a battery of other phenomena have been replicated across decades, populations, and incentive structures. They survive payoff manipulation, expert subjects, and even explicit warnings to subjects (Tversky and Kahneman, 1974). The persistence of these biases (as well as the apparent inability of conventional debiasing interventions to eliminate them) motivates a long-running search for tools and institutional designs that can help individuals make better decisions in domains ranging from retirement saving (Thaler and Benartzi, 2004) to medical choice.

Generative artificial intelligence has emerged as a candidate for precisely this role. By 2024, ChatGPT and comparable systems had become the most widely used decision-support tools in human history, deployed across contexts ranging from drafting employee benefits enrollment forms to providing investment guidance. If these systems offer advice that is itself rational or at least less biased than the unaided human, then they may provide effective decision aids that conventional educational interventions have failed to provide. If, instead, they reproduce or amplify the biases that are present in their training data, then their widespread adoption could further entrench those biases.

This paper provides systematic evidence on this important open question. We run fourteen of the most influential behavioral economics experiments on OpenAI's ChatGPT model via its application programming interface (API). We use the same wording that produced the canonical demonstration of a particular bias in human subjects. To capture the stochastic component of the model's responses, we run each experiment one hundred times with independent calls. This process allows us to construct a pseudo-panel of responses to the set of questions for each iteration. We also structure the questions to be answered binarily, such as "Yes," "No," or specific numerical values to minimize data cleaning. Our experiments span four classes of tasks: belief-formation problems in which the model must produce a numerical estimate or probability judgment; preference problems in which it must choose between lotteries or describe its willingness to pay; intertemporal choice problems; and social-choice problems, including the ultimatum game.

Throughout the paper, we evaluate the model against the rational benchmark of classical economic theory. For choice under risk, the rational benchmark is expected utility maximization with stable preferences, satisfying the axioms of completeness, transitivity, independence, and continuity (von Neumann and Morgenstern, 1944). For intertemporal choice, the rational benchmark is exponential discounting with a single discount rate applied consistently across horizons (Samuelson, 1937). For valuation, it requires that willingness-to-pay and willingness-to-accept coincide for inframarginal goods, that reservation prices be independent of the framing or context of purchase, and that portfolio allocations respond to the underlying risk-return properties of assets rather than to the investment menu presentation. For belief formation, it requires probability judgments that respect logical axioms and estimates that are insensitive to arbitrary, uninformative anchors. For social choice, the relevant benchmark in the ultimatum game is subgame-perfect equilibrium under self-interested preferences, in which the responder accepts any positive offer. A response is said to exhibit a bias when it departs systematically from one of these rational benchmarks; it is said to be rational when it conforms to the benchmarks.

We have three main findings. First, on most tasks (eight of fourteen), ChatGPT behaves in a manner essentially indistinguishable from the rational benchmark and substantially more rational than the human subjects in the original studies. The model accepts the favorable single-shot lottery in Samuelson's (1963) myopic loss aversion problem in every instance, exhibits no measurable present bias in Thaler's (1981) intertemporal choice, chooses the certain payoff in both the gain and loss frames of Kahneman and Tversky's (1979) prospect theory problem (and is therefore internally consistent), shows no sunk-cost effect in Thaler's (1980) basketball ticket problem, exhibits no anchoring in the United Nations estimation task of Tversky and Kahneman (1974), expresses fully cooperative save-more-tomorrow commitments at the same rate as immediate-saving commitments, and exhibits no disposition effect in a standard investment vignette.

Second, the large language model (LLM) tested is not uniformly rational. On six of fourteen tasks, namely the framing of the Asian disease problem, the conjunction fallacy in the Linda problem, mental accounting in the beer-on-the-beach scenario, the house money effect, naïve diversification, and the endowment effect, the model produces choices consistent with well-documented human biases. In every case the magnitude of the deviation from rationality is smaller than the human benchmark in the original study. Where 50 percent of human subjects reverse their choice across frames in Tversky and Kahneman's (1981) Asian disease problem, only 9 percent of trials with ChatGPT do so. Where 85 percent of human subjects in Tversky and Kahneman (1983) commit the conjunction fallacy, the rate among trials with ChatGPT is 71 percent.

Third, the human behavioral biases that remain exhibit a recognizable pattern. The six tasks on which ChatGPT exhibits human-like biases share rich contextual descriptions. Three tasks involve the valuation of context-dependent goods, services, or outcomes (the beer, the mug, the asset allocation). The other three tasks involve the categorical processing of information whose rich contextual description drives the response (Linda, the Asian disease frames, the house money domain). In contrast, tasks that admit a clean expected-utility calculation are solved correctly. This pattern is consistent with a model that has been trained to simulate strong procedural knowledge of normative economic principles, but when given richly described scenarios, the resulting output still bears the imprint of the human-generated training data.

Our findings contribute to a rapidly growing literature on the behavior of large language models in economic experiments. Filippas et al (2023) introduce the framework of using language models as *homo silicus*, arguing that LLMs can be treated as implicit computational models of human behavior; he replicates several classic findings using GPT-3. Chen et al (2023) document that GPT-3.5 satisfies the generalized axiom of revealed preference at higher rates than human subjects in budget-allocation tasks, though it remains sensitive to the linguistic frame of the choice problem. Hagendorff et al (2023) report that the cognitive reflection test errors observed in earlier OpenAI models largely vanish in ChatGPT-3.5 and ChatGPT-4. Keshavarz et al (2026) also find that LLMs exhibit classic behavioral biases in their outcomes when applied to financial decision-making scenarios. Most closely related to our work, Bini et al (2026) administer a comprehensive battery of behavioral experiments across multiple LLM families and find that preference-based tasks elicit more human-like behavior in larger models while belief-based tasks are increasingly solved rationally. Our main contribution is to quantify both the direction and the magnitude of any deviation against the original human benchmark in a tightly controlled within-model comparison across fourteen canonical experiments. Further, we organize our results around the question

that motivates downstream policy: should we expect generative AI to debias the decisions of the people who use it?

The remainder of the paper proceeds as follows. Section 2 describes the experimental protocol. Section 3 presents results for each of the fourteen biases, organized into four substantive groups. Section 4 synthesizes the findings, develops a typology that distinguishes the biases the model corrects from those it inherits, and discusses implications for the use of generative AI as a decision-support tool. Section 5 concludes.

2 Experimental Design

2.1 Selection of Biases

We select fourteen behavioral phenomena that satisfy three criteria. First, each is established in a study now considered canonical in the behavioral economics literature, with the original effect documented at conventional levels of statistical and economic significance. Second, the experimental stimulus is sufficiently self-contained that it can be administered to a language model with a single prompt, without requiring elaborate context-setting or interactive manipulation. Third, the set spans the principal categories of bias identified in the surveys of Camerer et al (2004) and Thaler (2016): heuristics in belief formation, deviations from expected utility under risk, intertemporal preference anomalies, mental accounting and reference-dependent valuation, and social preferences. Table 1 lists the fourteen biases together with their original sources and the rational or human benchmark against which we evaluate the model's responses.

2.2 Protocol

For each of the fourteen biases, we construct a prompt that reproduces, as closely as a written prompt permits, the exact stimulus used in the original human study. Where the original study presents two or more between-subject conditions (for example, the gain and loss frames of the Asian disease problem), we administer each condition separately, treating each as a distinct experiment. The prompt is preceded by a brief role specification that instructs the model to respond as a decision-maker. We submit each prompt 100 times to OpenAI's ChatGPT model via the API, with independent calls and the default temperature setting that allows, but does not require, variability in responses. The resulting collection of approximately 2,800 observations across all biases and conditions forms our primary dataset.

We deliberately keep the prompt minimal. Recent work has shown that LLM responses can be made dramatically more or less rational through prompt engineering; for example, chain-of-thought prompts, role assignments as a financial expert, and explicit instructions to be rational all systematically affect outcomes (Chen et al, 2023; Bini et al, 2026). Our interest is in the model's default behavior when consulted by a typical user who does not engage in such optimization. The relevant counterfactual for policy is the experience of an ordinary person asking ChatGPT for help with a decision, not the experience of a researcher who has tuned the prompt to elicit particular reasoning.

2.3 Comparison with the Human Benchmark

For each bias we report two quantities: the proportion of ChatGPT trials that exhibit the bias-consistent response, and the comparable proportion in the original human study. We then characterize the result in one of three ways. We say that the model *exhibits no bias* when the proportion is statistically and economically indistinguishable from what a rational benchmark would predict. We say that the model *exhibits an attenuated bias* when the direction of the response matches the human pattern, but the magnitude is smaller. We say that the model *exhibits a bias of comparable magnitude* when the proportion is statistically indistinguishable from the human rate.

3 Results

We organize the fourteen experiments into four groups. Section 3.1 covers framing and reference-dependent valuation. Section 3.2 includes belief-formation tasks. Section 3.3 covers intertemporal and risk-related preferences. Section 3.4 includes social and portfolio choices. Within each section we present the evidence bias-by-bias in the order in which the canonical sources appear in the literature.

3.1 Framing and Reference-Dependent Valuation

3.1.1 The Asian Disease Problem (Tversky and Kahneman, 1981)

The original demonstration of framing in risky choice asks subjects to choose between a sure option and a risky option in a public-health scenario. In the gain frame, Program A saves 200 of 600 lives for sure, while Program B offers a one-third chance of saving all 600. In the loss frame, Program C kills 400 of 600 for sure, while Program D offers a two-thirds chance that all 600 die. The two frames describe identical lotteries. Tversky and Kahneman (1981) find that 72 percent of subjects choose A in the gain frame and 22 percent choose C in the loss frame, a fifty-percentage-point preference reversal that has since been replicated many times (Kühberger, 1998).

ChatGPT chooses Program A in the gain frame in 100 of 100 trials and chooses Program C in the loss frame in 91 of 100 trials. Comparing the rates of choosing the certain option, the model's preference reversal across frames is 9 percentage points, against the 50-percentage-point reversal in the human study. The model exhibits an attenuated framing effect in the same direction as humans, and the magnitude of the bias is approximately one-fifth of the human benchmark. Even in the loss frame, the model retains a strong preference for the certain option, suggesting that it does not exhibit the reflection of risk attitudes across the gain-loss boundary, the central prediction of prospect theory.

3.1.2 Mental Accounting (Thaler, 1985)

Thaler's (1985) beer-on-the-beach scenario asks the respondent to imagine lying on a beach and to state the maximum price the respondent would pay for a beer purchased by a friend. In one condition the beer comes from a fancy resort hotel; in the other it comes from a small, run-down grocery store. The beer itself is identical, and the respondent will consume it on the beach in either case. Standard theory of consumer choice predicts identical reservation prices. Thaler (1985) reports substantially higher reservation prices when the beer comes from the fancy hotel, on the order of $1.50 against $0.95 in the original 1985 sample.

ChatGPT exhibits the same pattern. The mean reservation price across 100 trials in the fancy resort condition is approximately $7.50; in the run-down grocery store condition, it is approximately $5.00. The dollar gap is larger than in the original study; however, the proportional gap is similar and the direction is identical. The model is therefore committed to mental accounting in the sense Thaler intended: the perceived appropriateness of a price depends on the context of purchase even when consumption is held fixed. This result is particularly striking because the model has no reputation, no preference for premium settings, and is not subject to any social pressure, all reasons that a human may be driven to pay a higher price. This bias is best understood as a learned association between contextual descriptors and price norms, encoded during training on human-written text.

3.1.3 The House Money Effect (Thaler and Johnson, 1990)

Thaler and Johnson (1990) show that subjects who have just won $15 are substantially more willing to accept a fair gamble of plus or minus $4.50 than are subjects facing a choice with the same final payoff distribution but no preceding windfall. In their data, 77 percent of subjects take the gamble after a windfall, while only 44 percent take the equivalent compound gamble in the no-windfall framing. The result is a violation of the independence axiom of expected utility theory and a leading piece of evidence for the proposition that perceived ownership of money affects subsequent risk attitudes.

We present both framings to ChatGPT. In the windfall framing, the model declines the gamble in 79 percent of trials, essentially the mirror image of the human result. In the integrated framing, the model prefers the certain $15 in 99 percent of trials. The directions match the human pattern in the sense that the model is more likely to take the gamble after a windfall (21 percent) than in the integrated framing (1 percent), but the levels are dramatically different from those of human subjects. In this context the model is much more risk averse on net than the typical human subject. Whether this represents a separate kind of bias or a successful internalization of the rational benchmark is a question we return to in Section 4.

3.1.4 The Endowment Effect (Kahneman et al, 1990)

The classical Cornell coffee-mug experiments document that subjects randomly endowed with a mug demand substantially more to part with it than subjects without a mug are willing to pay to acquire one. In the original study, the median willingness-to-accept is approximately $5.25 against a willingness-to-pay of $2.25 to $2.75, with the mug retailing for $6.00 at the time of the original experiment. Standard economic theory predicts that there is no gap between willingness-to-pay and willingness-to-accept in this context. Note that we update our model prompt to reflect the current retail price of $14.99 for a Cornell coffee mug.

Our experiment elicits a striking pattern. When endowed with the mug, the model splits almost evenly between accepting $14.99 and refusing to sell at any price (encoded as $0). When asked the maximum it would pay, the model essentially refuses to pay anything: 98 percent of trials return $0 and only 2 percent return $14.99. The mean willingness-to-accept therefore far exceeds the mean willingness-to-pay, in the same direction as the human result. The pattern is, however, structurally different from the human one: rather than producing intermediate values that cluster around a reference point, the model is bimodal, oscillating between the salient retail anchor and zero. We interpret this as the model executing two competing rules: respect the listed retail price; refuse a transaction whose subjective value is unclear,

rather than computing a continuous valuation. Whatever the mechanism, the qualitative endowment effect appears in the model's outcomes.

3.2 Belief-Formation Tasks

3.2.1 Anchoring (Tversky and Kahneman, 1974)

The original demonstration of anchoring asks subjects to estimate the percentage of African countries in the United Nations after first being asked whether the percentage is greater or less than a randomly assigned anchor (10 percent or 65 percent). The median estimates are 25 and 45 percent, respectively. The fact that an arbitrary anchor with no informational content shifts the median estimate by twenty percentage points has been replicated in dozens of contexts and is regarded as one of the most robust phenomena in cognitive psychology (Furnham and Boo, 2011).

ChatGPT is unmoved by the anchor. Across 100 trials in each anchor condition, the model returns an estimate of 28 percent every time, regardless of whether the anchor presented is 10 percent or 65 percent. The constancy is striking: the model does not appear to use the anchor as a cue at all. Instead, it generates what is the best estimate of the underlying quantity from its training data unmoved by the anchor. Where humans are anchored by 20 percentage points, the model is anchored by 0. Notably, this experiment is also one in which the model's tendency to generate varied responses vanishes entirely. We return to the implications of this consistent rationality on numerical estimation tasks in Section 4.

3.2.2 The Conjunction Fallacy (Tversky and Kahneman, 1983)

In the canonical Linda problem, subjects read a description of a 31-year-old, single, outspoken philosophy major who is concerned with social justice and has participated in anti-nuclear demonstrations. They are then asked to rank the likelihood of various propositions about Linda. Most relevant are "Linda is a bank teller" (proposition F) and "Linda is a bank teller and is active in the feminist movement" (proposition H). Probability theory requires that $P(F)$ be at least as large as $P(F \cap H)$, since H is a strict subset of F. Tversky and Kahneman (1983) report that 85 percent of respondents nevertheless rank H as more likely than F, identifying the representativeness heuristic as the underlying mechanism.

ChatGPT commits the conjunction fallacy on 70.7 percent of valid trials (one of the 100 trials returns only proposition G and is excluded from the denominator). The direction of the bias matches the human pattern, and the magnitude is somewhat smaller. The persistence of the fallacy in 71 percent of trials is, however, substantial. We interpret this result as evidence that the representativeness pattern is encoded so deeply in the human-written training data that the model reproduces it. Hagendorff et al (2023) document a similar pattern of failure on the cognitive reflection test in earlier model generations and improvement in later ones, suggesting that progress on this kind of bias has been ongoing.

3.3 Risk and Intertemporal Preferences

3.3.1 Myopic Loss Aversion (Samuelson, 1963; Benartzi and Thaler, 1995)

Samuelson's (1963) celebrated colleague refuses a single play of a 50-50 gamble offering plus $200 or minus $100 but reports that he would accept 100 plays of the same gamble. The combination is irrational under

any expected-utility specification with smooth preferences, since acceptance of the compound gamble implies acceptance of each component (Rabin, 2000). The pattern is one of the leading pieces of evidence for myopic loss aversion as defined by Benartzi and Thaler (1995).

ChatGPT accepts the single gamble in 100 of 100 trials and accepts the compound gamble of 100 plays in 100 of 100 trials. Both responses are consistent with rational expected utility maximization, and the model exhibits no inconsistency between the two. This result is among the cleanest in our sample: the model not only computes the expected value correctly in each isolated case but is also internally coherent across the two scenarios in a way that the typical human subject is not.

3.3.2 Present Bias (Thaler, 1981)

Thaler (1981) reports that the median respondent is indifferent between $15 today and $20 in one month (an implicit one-month discount rate of 33 percent) but becomes markedly more patient when both options are shifted into the future. Hyperbolic discounting models, including the quasi-hyperbolic specification of Laibson (1997), capture this pattern as time inconsistency arising from a fixed discounting on all future consumption.

ChatGPT chooses $20 in one month over $15 today in 100 of 100 trials and chooses $20 in six months over $15 in five months in 100 of 100 trials. The model exhibits no present bias and no time inconsistency. It is more patient than the typical human subject and is internally consistent across the two framings. As with myopic loss aversion, this is a case where the model performs the expected-value calculation cleanly and is unmoved by the salience of immediate gratification.

3.3.3 Risk Preferences over Gains and Losses (Kahneman and Tversky, 1979)

The fourfold pattern of risk attitudes that motivates prospect theory predicts risk aversion over moderate-probability gains and risk seeking over moderate-probability losses. Kahneman and Tversky (1979) report that 80 percent of subjects prefer a sure $3,000 to an 80 percent chance of $4,000, while only 8 percent of the same subjects prefer a sure loss of $3,000 to an 80 percent chance of losing $4,000. The reflection across the gain-loss boundary is one of the central empirical pillars of prospect theory.

ChatGPT chooses the sure $3,000 in 100 of 100 trials in the gain frame and chooses the sure $3,000 loss in 100 of 100 trials in the loss frame. Strictly speaking, the model is risk-averse in both domains. This preference pattern is consistent with concave utility over wealth and inconsistent with prospect theory's S-shaped value function. Compared to the human pattern, the model differs by not becoming risk-seeking in the domain of losses.

3.3.4 The Sunk Cost Fallacy (Thaler, 1980)

Thaler (1980) presents subjects with two scenarios: in one, the subject has paid $40 for a basketball ticket and faces a snowstorm on game day; in the other, the subject is given the ticket for free and faces the same snowstorm. The marginal cost of attending is identical in both. The standard sunk-cost fallacy is that subjects are substantially more likely to attend when they have paid for the ticket, treating the sunk cost as relevant to the marginal decision when rationally it should make no difference.

ChatGPT chooses not to attend in 100 of 100 trials in both conditions. The model treats the sunk cost as irrelevant, as the rational benchmark requires, and is internally consistent across the two scenarios. We note that the model's choice not to attend differs from the modal human response in the paid-ticket condition (to attend). The model's response is consistent with rational treatment of marginal cost.

3.3.5 The Save More Tomorrow Plan (Thaler and Benartzi, 2004)

Thaler and Benartzi (2004) document that 78 percent of employees who decline to increase their retirement contributions today nevertheless agree to commit to increasing them after their next raise. The asymmetry is evidence of present bias and self-control problems, since the loss in current consumption from the future commitment is, viewed prospectively, identical to the loss in current consumption from an immediate increase.

ChatGPT agrees to commit to a 3 percent retirement contribution today in 100 of 100 trials and agrees to the same commitment after the next raise in 100 of 100 trials. The model exhibits no asymmetry: it accepts the commitment in both framings. As with the present-bias result of Thaler (1981), the model behaves as if it has internalized a commitment device for the rational behavior, never invoking the present bias and self-control problem that motivates the human pattern.

3.3.6 The Disposition Effect (Odean, 1998)

Odean (1998) documents that individual investors realize gains at substantially higher rates than they realize losses, a pattern inconsistent with rational tax-motivated trading and consistent with the prediction of prospect theory that investors are reluctant to acknowledge losses. We adapt the standard disposition test to the language model. In our version a stock is up or down 10 percent and the investor must choose to sell, hold, or buy more.

ChatGPT chooses "hold and wait" in 100 of 100 trials in the down condition and "hold and wait" in 97 of 100 trials in the up condition (with three trials choosing "buy more"). The model never sells in either condition; therefore, the proportion of gains realized minus the proportion of losses realized (Odean's standard measure of the disposition effect) is exactly zero. The model exhibits no disposition effect. We caution that this result is partly mechanical: the model's strong default toward "hold" eliminates the asymmetry by collapsing both numerators. But the rational benchmark, properly understood, also calls for the absence of an asymmetry between sells of winners and losers, and on that benchmark the model performs well.

3.4 Social and Portfolio Choice

3.4.1 The Ultimatum Game (Güth et al, 1982)

The ultimatum game is the leading laboratory tool for measuring deviations from the self-interested benchmark in bilateral bargaining. The proposer divides a fixed sum and the responder accepts or rejects; rejection yields zero to both. The subgame-perfect equilibrium under selfish preferences is for the proposer to offer the smallest positive amount and for the responder to accept. Empirical play converges instead on offers near 50 percent and rejection rates of 40 to 50 percent for offers below 20 percent (Camerer, 2003).

When ChatGPT is asked to play as the proposer with $100, it offers exactly $50 in essentially every trial. When asked to play as the responder, it accepts offers as low as $1 in essentially every trial. The proposer's behavior matches the modal human pattern; the responder's behavior matches the rational benchmark of selfish maximization. Taken together, the two behaviors are inconsistent with one another in the sense that no single utility function over self and other consequences can rationalize both. We interpret the pattern as evidence that the model has learned to apply role-specific norms rather than to reason from a coherent set of social preferences.

3.4.2 Naïve Diversification (Benartzi and Thaler, 2001)

Benartzi and Thaler (2001) document that retirement-plan participants tend to allocate contributions equally across the funds offered using a 1/n heuristic rather than allocating according to the underlying risk-return profile of the funds. They report a 56 percent equity allocation when participants choose between a stock and a bond fund, an 80 percent equity allocation when they choose between a stock and a balanced (50-50) fund, and only 20 percent equity when they choose between a balanced and a bond fund. The pattern is incompatible with mean-variance optimization for any single set of preferences.

We present ChatGPT with two analogous portfolio problems. In the menu of two stock funds and one bond fund, the modal allocation is 40 percent to each stock fund and 20 percent to the bond fund. In the menu of one stock fund and two bond funds, the modal allocation is 50 percent to the stock fund and 25 percent to each bond fund. Both allocations are consistent with the 1/n heuristic, in the sense that funds within the same asset class receive equal weight. The resulting equity exposure depends on the structure of the menu (80 percent in the first case, 50 percent in the second). The model is therefore subject to the same partition-dependence that drives the original finding of Benartzi and Thaler. The implication for retirement-plan design is the same as in the human case: the menu of funds shapes the resulting portfolio, and a model deployed to advise retirement savers will reproduce that dependence.

Table 1 Summary of Fourteen Behavioral Economics Experiments Administered to ChatGPT (100 Trials per Condition)

Bias	Source	Human Benchmark	ChatGPT Result	Bias Inherited?
Framing (Asian disease)	Tversky & Kahneman 1981	72% A; 22% C (50pp reversal)	100% A; 91% C (9pp reversal)	Yes, attenuated
Myopic loss aversion	Samuelson 1963	Reject single, accept 100	Accept both (100%/100%)	No
Present bias	Thaler 1981	Indifferent at \$15 vs \$20/mo	\$20 in both framings (100%)	No
Risk reflection	Kahneman & Tversky 1979	80% risk-averse gain; 8% risk-averse loss	100% sure in both	No,(over-cautious)
Mental accounting	Thaler 1985	Higher WTP at fancy hotel	\$7.50 vs \$5.00 (50% gap)	Yes, attenuated
Sunk cost fallacy	Thaler 1980	Attend if paid; not if free	Do not attend in either (100%)	No
Anchoring	Tversky & Kahneman 1974	25% vs 45% by anchor	28% in both anchors	No
Ultimatum game	Güth et al. 1982	Offer ∼50%; reject < 20%	Offer \$50; accept \$1	No,(responder rational)
Conjunction fallacy	Tversky & Kahneman 1983	85% rank H > F	70.7% rank H > F	Yes, attenuated
Save More Tomorrow	Thaler & Benartzi 2004	Future > present saving	Accept both at 100%	No
Disposition effect	Odean 1998	Realize gains > losses by 5pp	Hold in both (100%/97%)	No
House money effect	Thaler & Johnson 1990	77% gamble after windfall; 44% otherwise	21% vs 1% gamble	Yes, attenuated
Naïve diversification	Benartzi & Thaler 2001	1/n allocation; partition dependence	80% vs 50% equity by menu	Yes, comparable
Endowment effect	Kahneman, Knetsch, Thaler 1990	WTA > WTP by ∼2x	WTA \$7.50; WTP \$0.30	Yes, attenuated

Notes: Each row reports the canonical human benchmark (column 3) and the corresponding ChatGPT result from 100 trials per condition (column 4). Green shading in the final column indicates that the model exhibits no economically meaningful bias. Yellow shading indicates that the model exhibits a directional bias matching the human pattern; in all such cases the magnitude is attenuated relative to the human benchmark. WTA denotes willingness-to-accept; WTP denotes willingness-to-pay; pp denotes percentage points.

4 Discussion: A Typology of Inherited and Corrected Biases

ChatGPT exhibits no measurable bias on eight of fourteen tasks: myopic loss aversion, present bias, the gain-loss reflection of prospect theory, the sunk cost fallacy, anchoring, the Save More Tomorrow asymmetry, the disposition effect, and the responder side of the ultimatum game. On the remaining six tasks, it produces choices in the same direction as the documented human bias, with magnitudes that are attenuated. These six tasks showing attenuated bias include framing of the Asian disease problem, the conjunction fallacy, mental accounting, the house money effect, naïve diversification, and the endowment effect. We see no case in our sample in which the model is more biased than the human benchmark.

4.1 A Conjecture About the Pattern

The six tasks that elicit the inherited biases share a common structure. They demand that the model produce a contextual valuation or a judgment about likelihood that depends on the surface features of

a richly described scenario. The mug, the beer, the windfall, the menu of funds, the linguistic frame of the disease problem, and Linda's biography all introduce information that should be irrelevant under the strict normative theory but that drives both human and model responses. By contrast, the nine tasks the model handles correctly tend to admit a clean reduction to expected value or to a simple consistency check across two framings. Computing the expected value of a 50-50 gamble, computing the present value of a future payoff, computing the marginal cost of attending a basketball game, or computing the relative likelihood of a conjunction are tasks for which a generic procedural rule, applied without reference to context, produces the right answer. The model has learned to simulate procedure.

We do not claim this typology is sharp. The conjunction fallacy in the Linda problem is, in principle, a clean inequality between probabilities, yet the model gets it wrong 71 percent of the time. We suspect the explanation is that the descriptive content of Linda's biography is so semantically connected to feminism in the training corpus that the model treats the inferential task as one of narrative fit rather than probability comparison. This finding is consistent with Hagendorff et al's (2023) interpretation of LLM cognitive errors as a form of intuitive reasoning (labeled system one by Kahneman) that the model has only partially learned to override. Improvements across model generations on this and similar tasks are plausibly the result of training procedures that increasingly reward the application of procedural knowledge over narrative association.

4.2 Implications for the Use of AI as a Decision-Support Tool

The headline finding of our study is that the model's default response is more rational than the user's would be on the majority of classical decision problems. A user contemplating the basketball-ticket dilemma, the present-bias intertemporal choice, the myopic loss aversion lottery, or the responder problem of the ultimatum game would receive better advice from ChatGPT than the modal human peer would offer. This result is consistent with the broader finding of Chen et al (2023) that GPT-3.5 satisfies the generalized axiom of revealed preference at higher rates than humans in budget-allocation tasks.

However, the biases that are reproduced in our findings can be consequential. Mental accounting, the endowment effect, naïve diversification, and the framing effect are precisely the biases that motivate the design of retirement plans, insurance markets, and medical decision aids. A user who consults a generative AI to help structure a retirement portfolio may encounter the same partition-dependence that motivates the Benartzi and Thaler critique of 401(k) menus. A user who consults the model to negotiate a sale price for an item they own may get different results depending on irrelevant context provided. A user who frames a public-health choice in terms of lives lost rather than lives saved will receive systematically different advice. These residual biases concentrate in domains where the welfare implications are high.

Two qualifications are important. First, our experiments use a minimal prompt designed to mimic the casual user. There is now evidence that prompting techniques reduce or eliminate many of these biases (Bini et al, 2026). The biases we report should therefore be interpreted as default behavior rather than as an upper bound on the model's capabilities. Second, our experiments use a single model with a single set of training and tuning choices. Bini et al (2026) document substantial variation across model families and across model generations, with some biases attenuating and others persisting as models scale. Our headline result that current ChatGPT is more rational than the typical human, but exhibits residual biases in valuation and framing tasks, is consistent with the broader pattern in this literature.

4.3 Why Are Some Biases Inherited?

A natural question is why the language model, which is in principle capable of computing expected values and applying probability axioms, retains any of the biases at all. Our reading of the evidence suggests three mechanisms. First, the training corpus is generated by humans whose own writing reflects the biases under study; passages that describe the value of a beer in a fancy hotel will systematically associate the higher price with the more luxurious context. The model learns the association as a fact about how humans write about value, and it reproduces the association when asked. Second, biases that depend on the categorical structure of the problem, namely, the partition into stock funds and bond funds, the framing as gains or losses, and the description of Linda's interests, operate through the model's processing of natural language. Third, post-training fine-tuning that rewards conventionally helpful responses may reinforce certain choices that are themselves biased. The strong default toward "hold and wait" in the disposition vignettes, for instance, may reflect a fine-tuning preference for cautious advice that happens to coincide with the rational benchmark in this case but might not in others. Since we find that LLMs replicate the biases that follow from framing, perhaps their responses will vary relative to the user's own interpretation of the problem. For example, an enthusiastic risk-taker may interact with ChatGPT and receive an equally as enthusiastic and optimistic reply. Given the dependency on the prompt itself, it would follow that the level of risk management the LLM exhibits is relative to that of the prompt it receives, or what it can perceive from it, even if the objective answer is very different.

4.4 Relation to the Prior Literature

Our findings line up with the emerging consensus in the LLM-and-economics literature. Filippas et al (2023) argue that LLMs can be treated as *homo silicus* whose responses qualitatively mirror those of human subjects in classic economic games. Our results bear out this characterization in the six experiments where the model exhibits a directional human-like bias, but they qualify it in two important respects: the magnitudes are smaller than the human benchmarks, and on the majority of tasks the model is closer to *homo economicus* than to *homo sapiens*. Chen et al (2023) reach a closely related conclusion using a different methodology: GPT-3.5 satisfies the generalized axiom of revealed preference at substantially higher rates than human subjects in budget-allocation tasks, with sensitivity primarily to the linguistic frame of the choice problem. Hagendorff et al (2023) report that the cognitive reflection test errors of older models largely vanish in ChatGPT-3.5 and ChatGPT-4. The most comprehensive recent study, Bini et al (2026), administers a battery of behavioral experiments across multiple LLM families and finds that preference-based tasks elicit more human-like behavior in larger models while belief-based tasks are increasingly solved rationally, a pattern that maps closely onto our results and interpretation of these findings.

5 Conclusion

We administer fourteen canonical behavioral economics experiments to ChatGPT one hundred times each and compare the resulting distributions of choices to the human benchmarks reported in the original studies. On eight of fourteen tasks the model behaves rationally and is more rational than the human subjects in the source experiments. On six of fourteen tasks the model exhibits a directional bias that matches the human pattern, with magnitudes that are in every case attenuated relative to the human

benchmark. The biases the model inherits concentrate on tasks involving valuation, framing, and the categorical processing of richly described scenarios. The biases the model corrects for involve tasks that admit a clean expected-value or consistency calculation.

These findings have mixed implications for the policy question that motivates the study. On one hand, generative AI does appear to be a useful debiasing aid in the majority of decision domains and is especially valuable on the intertemporal-choice and risky-choice problems that have proven resistant to conventional educational debiasing. On the other hand, the biases that survive are concentrated in exactly the domains (retirement saving, framing of medical and policy choices, valuation of owned assets) where the welfare consequences of biased decisions are largest and where institutional design has long sought to compensate for human cognitive limitations. The expectation that generative AI will fully substitute for institutional debiasing is therefore unwarranted. The expectation that it will substantially improve average decision quality is, by contrast, well supported by our results.

There is much research yet to be done on the effectiveness of LLMs to aid human decision making. The response of the inherited biases to systematic prompt engineering (chain-of-thought instructions, role assignments as a financial advisor, explicit invocation of the rational benchmark) should be quantified for the same set of canonical experiments. Researchers should also explore the heterogeneity of these patterns across model families. Fine-tuning regimes should be mapped to identify which design choices most reduce inherited biases. Field evidence on the actual prompts that ordinary users supply to generative AI, and on the contextual richness of those prompts relative to our minimal protocol, would be powerful to produce better estimates of welfare consequences of using these models. An essential question for future research is whether a typical user, asking the typical question, is helped or hindered. Our results suggest that on average the answer is helped, with some important exceptions.

References

Benartzi S, Thaler RH (1995) Myopic loss aversion and the equity premium puzzle. Quarterly Journal of Economics 110(1):73–92

Benartzi S, Thaler RH (2001) Naive diversification strategies in defined contribution saving plans. American Economic Review 91(1):79–98

Bini P, Cong LW, Huang X, et al (2026) Behavioral economics of AI: LLM biases and corrections. NBER Working Paper 34745, National Bureau of Economic Research

Camerer CF (2003) Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton

Camerer CF, Loewenstein G, Rabin M (eds) (2004) Advances in Behavioral Economics. Princeton University Press, Princeton

Chen Y, Liu TX, Shan Y, et al (2023) The emergence of economic rationality of GPT. Proceedings of the National Academy of Sciences 120(51):e2316205120. https://doi.org/10.1073/pnas.2316205120

Filippas A, Horton JJ, Manning BS (2023) Large language models as simulated economic agents: What can we learn from Homo Silicus? arXiv preprint arXiv:2301.07543, https://doi.org/10.48550/arXiv.2301.07543

Furnham A, Boo HC (2011) A literature review of the anchoring effect. Journal of Socio-Economics 40(1):35–42

Güth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization 3(4):367–388

Hagendorff T, Fabi S, Kosinski M (2023) Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science 3(10):833–838

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. Econometrica 47(2):263–291

Kahneman D, Knetsch JL, Thaler RH (1990) Experimental tests of the endowment effect and the Coase theorem. Journal of Political Economy 98(6):1325–1348

Keshavarz J, Seagraves C, Sirmans S (2026) Artificially biased intelligence: Does AI think like a human investor? Ssrn working paper, SSRN, https://doi.org/10.2139/ssrn.5998954

Kühberger A (1998) The influence of framing on risky decisions: A meta-analysis. Organizational Behavior and Human Decision Processes 75(1):23–55

Laibson D (1997) Golden eggs and hyperbolic discounting. Quarterly Journal of Economics 112(2):443–478

von Neumann J, Morgenstern O (1944) Theory of Games and Economic Behavior. Princeton University Press, Princeton

Odean T (1998) Are investors reluctant to realize their losses? Journal of Finance 53(5):1775–1798

Rabin M (2000) Risk aversion and expected-utility theory: A calibration theorem. Econometrica 68(5):1281–1292

Samuelson PA (1937) A note on measurement of utility. Review of Economic Studies 4(2):155–161

Samuelson PA (1963) Risk and uncertainty: A fallacy of large numbers. Scientia 98:108–113

Thaler RH (1980) Toward a positive theory of consumer choice. Journal of Economic Behavior & Organization 1(1):39–60

Thaler RH (1981) Some empirical evidence on dynamic inconsistency. Economics Letters 8(3):201–207

Thaler RH (1985) Mental accounting and consumer choice. Marketing Science 4(3):199–214

Thaler RH (2016) Behavioral economics: Past, present, and future. American Economic Review 106(7):1577–1600

Thaler RH, Benartzi S (2004) Save more tomorrow: Using behavioral economics to increase employee saving. Journal of Political Economy 112(S1):S164–S187

Thaler RH, Johnson EJ (1990) Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. Management Science 36(6):643–660

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. Science 185(4157):1124–1131

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211(4481):453–458

Tversky A, Kahneman D (1983) Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review 90(4):293–315