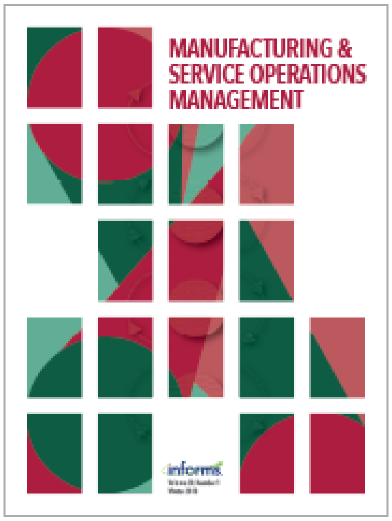


Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?

Yang Chen; , Samuel N. Kirshner; , Anton Ovchinnikov; , Meena Andiappan; , Tracy Jenkin

To cite this article:
Yang Chen; , Samuel N. Kirshner; , Anton Ovchinnikov; , Meena Andiappan; , Tracy Jenkin (2025) A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?. Manufacturing & Service Operations Management 27(2):354-368. https://doi.org/10.1287/msom.2023.0279

informs.
https://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 27, No. 2, March–April 2025, pp. 354–368
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?

Yang Chen,[a,*] Samuel N. Kirshner,[b] Anton Ovchinnikov,[c,d] Meena Andiappan,[e,f] Tracy Jenkin[c,g]

[a] Ivey Business School, Western University, London, Ontario N6G 0N1, Canada; [b] University of New South Wales Business School, University of New South Wales, Sydney, New South Wales 2052, Australia; [c] Smith School of Business, Queen's University, Kingston, Ontario K7L 3N6, Canada; [d] INSEAD, 77300 Fontainebleau, France; [e] DeGroote School of Business, McMaster University, Hamilton, Ontario L8S 4M4, Canada; [f] Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario M5T 3M6, Canada; [g] Vector Institute, Toronto, Ontario M5G 0C6, Canada

*Corresponding author

Contact: ychen@ivey.ca, https://orcid.org/0000-0001-8535-7815 (YC); s.kirshner@unsw.edu.au, https://orcid.org/0000-0002-2604-2001 (SNK); anton.ovchinnikov@queensu.ca, https://orcid.org/0000-0001-5972-2217 (AO); meena.andiappan@mcmaster.ca, https://orcid.org/0000-0002-4713-508X (MA); tracy.jenkin@queensu.ca, https://orcid.org/0000-0002-1639-9322 (TJ)

Received: May 21, 2023
Revised: May 31, 2024; November 18, 2024
Accepted: November 27, 2024
Published Online in Articles in Advance: January 31, 2025

https://doi.org/10.1287/msom.2023.0279

Abstract. ***Problem definition***: Large language models (LLMs) are being increasingly leveraged in business and consumer decision-making processes. Because LLMs learn from human data and feedback, which can be biased, determining whether LLMs exhibit human-like behavioral decision biases (e.g., base-rate neglect, risk aversion, confirmation bias, etc.) is crucial prior to implementing LLMs into decision-making contexts and workflows. To understand this, we examine 18 common human biases that are important in operations management (OM) using the dominant LLM, ChatGPT. ***Methodology/results***: We perform experiments where GPT-3.5 and GPT-4 act as participants to test these biases using vignettes adapted from the literature ("standard context") and variants reframed in inventory and general OM contexts. In almost half of the experiments, Generative Pre-trained Transformer (GPT) mirrors human biases, diverging from prototypical human responses in the remaining experiments. We also observe that GPT models have a notable level of consistency between the standard and OM-specific experiments as well as across temporal versions of the GPT-3.5 model. Our comparative analysis between GPT-3.5 and GPT-4 reveals a dual-edged progression of GPT's decision making, wherein GPT-4 advances in decision-making accuracy for problems with well-defined mathematical solutions while simultaneously displaying increased behavioral biases for preference-based problems. ***Managerial implications***: First, our results highlight that managers will obtain the greatest benefits from deploying GPT to workflows leveraging established formulas. Second, that GPT displayed a high level of response consistency across the standard, inventory, and non-inventory operational contexts provides optimism that LLMs can offer reliable support even when details of the decision and problem contexts change. Third, although selecting between models, like GPT-3.5 and GPT-4, represents a trade-off in cost and performance, our results suggest that managers should invest in higher-performing models, particularly for solving problems with objective solutions.

Funding: This work was supported by the Social Sciences and Humanities Research Council of Canada [Grant SSHRC 430-2019-00505]. The authors also gratefully acknowledge the Smith School of Business at Queen's University for providing funding to support Y. Chen's postdoctoral appointment.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/msom.2023.0279.

Keywords: large language models • decision biases • ChatGPT • behavioral operations management

1. Introduction

Large language models (LLMs) are massive artificial intelligence (AI) algorithms that process and generate text. As of late 2024, OpenAI's ChatGPT was the undisputed leader among commercial LLMs, attracting over 3 billion monthly visits worldwide; in comparison, Google's Gemini saw 300 million visits, whereas Perplexity, Anthropic, and Microsoft's CoPilot were under 100 million each.[1] This pattern hardly changed from late 2023, when ChatGPT captured 72% of the global text generative AI user base.[2] Popularized through the website and mobile applications, ChatGPT has conversational capabilities with broad general knowledge and remarkable problem-solving abilities, even in operations management (OM) tasks (Terwiesch 2023). Although some companies prohibit ChatGPT's use, CNBC[3] reports that "[h]alf of the companies … said they are using ChatGPT," and *Forbes* highlights[4] the specific ways in which Generative Pretrained Transformer (GPT) could be used in operational tasks: "[w]ith ChatGPT, retailers can manage inventory levels by analyzing sales data and predicting demand. This can help retailers avoid overstocking or running out of products, so they reduce costs and keep customers happier." More generally, Gartner predicts[5] that "[b]y 2026, over 100 million people will engage with robo-colleagues (synthetic virtual colleagues) to contribute to enterprise work."

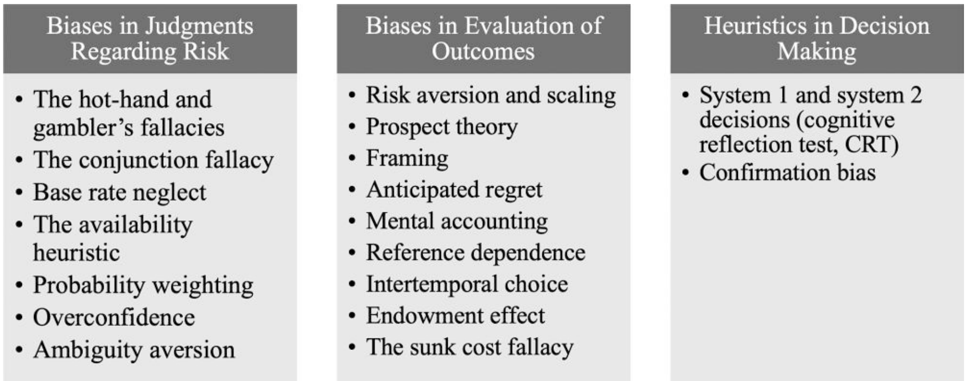
There is also ample evidence that consumers use ChatGPT for a broad range of tasks ranging from personal and shopping advice to business ventures and financial planning. For example, Motley Fool Money reports[6] that "54% of Americans use ChatGPT for personal finance recommendations." More generally, a survey by PWC[7] indicated that 44% of consumers would use chatbots for product information, whereas a third would use them for alerts, such as product availability, and personalized communication. Finally, industry reports suggest that value chain partners (manufacturers, retailers, buyers, suppliers, etc.) also use or plan to integrate GPT into various operational tasks. For example, PWC reports[8] that almost 40% of manufacturers plan to invest in GPT over the next 18 months. Multinational leaders, like Walmart, use LLMs for negotiating prices and selecting vendors.[9]

Taken together, LLMs, like ChatGPT, will inevitably impact operational decision making: either directly when managers use AI or indirectly by impacting consumer and/or supplier behavior. The question of how the behaviors of managers, consumers, and value chain partners impact operations is a central tenet of *behavioral operations*. With a few exceptions, most of the existing literature studies the behaviors of people and finds that people exhibit systematic decision biases, which can be strategically incorporated into the design of operational systems. However, as LLMs, like ChatGPT, become advisors or even delegates in various operational tasks, it becomes important to understand their biases: whether they act rationally, mirror human biases, or exhibit biases entirely different from human decision making.

Two counteracting forces may influence the biases of LLMs, like ChatGPT. On one hand, LLMs may be less biased in business decision making as after all, they are computer models; they lack emotions, do not suffer the same cognitive limitations as humans, and process information differently from the human brain. On the other hand, LLMs, such as ChatGPT, are trained on human data with human inputs. First, in the pretraining step, GPT learns from a vast collection of human language materials and picks up grammar, facts, and reasoning abilities as well as biases.[10] Second, in the fine-tuning step, a scoring model is trained with data generated by human reviewers on their ranking of potential ChatGPT outputs according to OpenAI guidelines. This methodology, called reinforcement learning with human feedback,[11] addresses concerns regarding GPT's responses to political and controversial topics by having human reviewers further train ChatGPT; however, it may also introduce bias. Which of the two forces will prevail in various tasks where the behavioral tendencies impact operations is unclear.

To investigate the biases of LLMs, we use the dominant model, GPT, to examine decision making across 18 common human biases identified as most relevant to operational decision making in *The Handbook of Behavioral Operations* chapter "Biases in individual decision-making" by Davis (2018). We borrow the classification of biases regarding risk judgments, the evaluation of outcomes, and heuristics in decision making; see Figure 1. We conducted our study in two phases, which we label as time 1 and time 2. At time 1 (in January and February of 2023 (i.e., very shortly after ChatGPT's initial release on November 30, 2022)), using one of the earliest versions of ChatGPT, we collected responses that examined the 18 biases using reference studies to which we refer as the *standard* context. These tests were done manually as

Figure 1. List of Behavioral Decision Biases Tested with GPT



Biases in Judgments Regarding Risk

- The hot-hand and gambler's fallacies
- The conjunction fallacy
- Base rate neglect
- The availability heuristic
- Probability weighting
- Overconfidence
- Ambiguity aversion

Biases in Evaluation of Outcomes

- Risk aversion and scaling
- Prospect theory
- Framing
- Anticipated regret
- Mental accounting
- Reference dependence
- Intertemporal choice
- Endowment effect
- The sunk cost fallacy

Heuristics in Decision Making

- System 1 and system 2 decisions (cognitive reflection test, CRT)
- Confirmation bias

Source. Adapted from Davis (2018) with modifications.

the application programming interface (API) was not available at the time. At time 2 (in October to December of 2023), with the availability of the API, we conducted a more comprehensive set of studies. First, for each bias, in addition to testing the standard context, we created two reframed *operations management* variants of the standard problems. These variants (one inventory related and one general) enable us to examine GPT's cross-context decision-making consistency and derive subsequent insights for OM researchers and practitioners. Second, for each scenario, we also compare two GPT models: one based on GPT-3.5, which at the time, was the most widely used and accessible free version of ChatGPT, and one based on GPT-4, which was accessible through the premium subscription. The analysis of our time 2 studies is the central focus of our research. However, the comparisons between time 1 and time 2 *versions* of the same GPT-3.5 model as well as the time 2 comparison between GPT-3.5 and GPT-4 *models* enable an investigation of the comprehensive trajectory of GPT's behavioral "evolution," shedding light on how model capabilities impact behavioral biases and producing insights into a possible future of LLM decision making in business contexts.

We focus on the foundational, individual biases outlined in Figure 1 as opposed to the prototypical operational problems (like the newsvendor problem) that combine multiple biases because we believe that it is important to understand the microfoundations of AI behavior first. This is no different from human behavior, where the existing behavioral theories (e.g., prospect theory) are leveraged to generate hypotheses and make predictions about outcomes and performance in operational contexts. However, these theories were developed over decades of foundational research on human decision making in economics and psychology. Because LLMs, like ChatGPT, are both novel and black boxes, there is currently limited insight into their decision-making processes. Our research aims to bridge this gap by examining GPT's decision making with respect to the foundational human biases relevant to OM.

With this concentrated focus on GPT's behavioral biases in operational decision making, we have several key findings.

1. GPT, particularly GPT-4, displays decision-making patterns that vary by bias/task. In subjective scenarios (e.g., prospect theory and framing) with uncertain outcomes, GPT has a pronounced tendency toward risk aversion and a preference for certainty. Conversely, GPT methodically searches for calculable solutions when faced with objective tasks but relies on heuristic reasoning if a formula is not readily available. As a result, it exhibits no bias in tasks like the cognitive reflection test (CRT) and base-rate neglect, where the solution "formula" exists but exhibits human-like biases in tasks like the conjunction fallacy and the confirmation bias tests, which are more logic-based tasks.

2. GPT shows an admirable level of consistency in decision making across contexts (that is, between the standard tests reported in the literature and hence, also present in the LLM's training data, and the new OM-specific tests that we created for the purpose of this study and thus, were unseen by the model). This response consistency provides additional evidence that GPT's behavior that we observed can be attributed to its systematic decision process. That GPT exhibits high consistency across diverse scenarios also suggests a necessary level of reliability and predictability that is crucial for managers implementing or encouraging the use of GPTs.

3. Furthermore, our comparative analysis between GPT-3.5 and GPT-4 models reveals critical insights into the evolution of GPT's decision-making capabilities. GPT-4 not only advances in accuracy but also displays increased decision biases in specific contexts. This dual-edged progression, with GPT-4 amplifying some biases while reducing others, indicates that adding more data and guardrails may prevent LLMs from improving performance across the board. This contrasts with the tremendous consistency in decision-making biases that we observe between the "early" and "late" versions of the same GPT-3.5 model; the minor updates that LLM vendors frequently release do not appear to impact decision biases of their models.

The observation that GPT exhibits certain biases while lacking others has multiple implications for behavioral operations in situations where LLMs act as advisors or delegates in operational tasks. For example, the presence of biases, such as overconfidence, risk aversion, and the hot-hand fallacy, could affirm existing insights within the field, whereas the absence of biases, like cognitive reflection, base-rate neglect, and the sunk cost fallacy, may challenge or even invalidate certain current understandings. Overall, our research systematically examines the specific behavioral biases manifested by GPT within operational contexts, laying the groundwork for developing a comprehensive understanding of LLM decision making and its implications for behavioral operations.

The rest of the article is organized as follows. In Section 2, we discuss the emerging literature on LLMs' decision-making processes. Section 3 outlines our experimental protocol for testing the 18 biases. In Section 4, we explore our primary results concerning GPT's performance against prototypical human behavior across various contexts and between model versions. We also identify key patterns in decision making that provide insights into the microfoundations underpinning GPT behavior. We conclude the article in Section 5, highlighting the implications of our findings.

For brevity, the main body only summarizes the results and their implications; all of the details are presented in the Online Appendix, which consists of three parts. Online Appendix A offers comprehensive methodological details on our experimental procedures, including the text of each vignette and code for running the experiments. Online Appendix B provides a breakdown of the individual biases and the statistical analysis supporting our results for the time 2 experiments. Finally, Online Appendix C presents the analyses of time 1 experiments.

2. Literature Review: LLMs and Decision Making

Since ChatGPT's release on November 30, 2022, there has been a surge in experimental research focusing on LLMs. This body of literature primarily addresses LLM decision making and is divided into two main branches. The first branch uses behavioral sciences to evaluate AI behavior, which Meng (2024) refers to as "AI Behavioral Sciences." A key goal of this area of research is assessing or enhancing AI's capability as a "silicon sample" that could potentially substitute for human behavior in experiments or market research (e.g., Li et al. 2024). The second branch looks into the impact of LLMs with the objective of enhancing decision-making processes or probing the broader implications of deploying these models (e.g., their effect on productivity) (Noy and Zhang 2023). Xu et al. (2024) categorizes these research branches as the "social science of AI" and "AI for social science," respectively. Our study is positioned as the "social science of AI" as we are investigating whether GPT exhibits human-like biases. However, we also intersect with "AI for social science" as our ultimate goal is using the understanding of GPT's biases to improve operational decision making. This approach aligns with the framework proposed by Davis et al. (2024), which argues that machine learning and behavioral sciences can work together to address OM problems.

Within the domain of "social science of AI," Binz and Schulz (2023) pioneered the exploration of GPT's capabilities by incorporating GPT models as participants in decision-making experiments. This seminal work focused on GPT-3, revealing its proficiency in decision making and deliberation and its adequacy in information search yet a notable deficiency in causal reasoning. Following Binz and Schulz (2023), a breadth of research has scrutinized GPT's efficacy across various fields, including psychology (Park et al. 2024), social science (Argyle et al. 2023), marketing (Brand et al. 2023), and economics (Horton 2023). A portion of this research focuses on identifying biases and novel effects unique to LLMs. Notably, Park et al. (2024) uncovered a "correct answer" effect, observing negligible variation in GPT-3.5's responses to some queries (e.g., regarding political orientation and economic preferences), which suggests GPT's presumption of definitive answers to such questions. Furthermore, applying action identification and construal-level theory, studies by Fennell (2023) and Kirshner (2024b) demonstrate GPT's tendency to abstractly describe behaviors. The majority of this burgeoning research stream examines GPT's decision making with regard to previously identified human biases, which is consistent with the objectives of our study.

Our investigation aligns with recent explorations into bounded rationality within GPT, echoing studies that have examined GPT's cognitive reflection and reasoning capabilities. Both Binz and Schulz (2023) and Hagendorff et al. (2023) investigated whether GPT exhibits cognitive reflection. Before ChatGPT's introduction, Binz and Schulz (2023) found that GPT-3 gave intuitive (but incorrect) responses to the cognitive reflection test. Additionally, both Dasgupta et al. (2022) and Binz and Schulz (2023) tested GPT-3's logical reasoning through its propensity to exhibit confirmation bias using the Wason selection task. We also employ this task (as well as its variants) in our analysis. Consistent with Dasgupta et al. (2022) and Binz and Schulz (2023), we find that GPT performs poorly at this task, exhibiting confirmation bias.

More recently, other research has expanded on the methodology of Binz and Schulz (2023) to investigate additional biases that we also consider in our study.[12] For instance, Ma et al. (2023) investigate decision-making frames using the disease outbreak problem and examines the gambler's fallacy behavior through a coin flipping experiment with GPT-3.5, an approach that mirrors ours. Although their findings on the gambler's fallacy align with ours, Ma et al. (2023) observed that GPT exhibits a preference for risk in losses and certainty in gains. Interestingly, these results mirror those from our time 1 experiments (see Online Appendix C for details). However, our subsequent findings show a shift toward a preference for certainty across both gains and losses in GPT-3.5 and GPT-4. Although this drastic change may seem unusual, Hagendorff et al. (2023) also observed behavioral changes across models (e.g., GPT's degree of cognitive reflection). Similar to our work, several articles have examined the conjunction fallacies using GPT-3.5 and GPT-4 models, including Macmillan-Scott and Musolesi (2024), Suri et al. (2024), and Wang et al. (2024). Consistent with our findings and Binz and Schulz (2023), who used GPT-3, the results consistently show that GPT models are prone to the conjunction fallacy regardless of the version. Lastly, the literature integrating LLMs into OM and supply chains has also been expanding (e.g., Wamba et al. 2023, Jackson et al. 2024), including the use of experimental methods based on Binz and Schulz (2023). For example, Su et al. (2023) explored GPT-4's capabilities with the classic newsvendor problem, and Kirshner (2024a) investigated GPT agents' decision making within the Management Science Replication Project (Davis et al. 2023).

A distinctive aspect of our research is the comparative analysis of biases between models (GPT-3.5 versus GPT-4) and versions ("early" time 1 versus "late" time 2 GPT-3.5). Although other studies undertake somewhat similar comparisons, they typically concentrate on a limited number of biases (e.g., Suri et al. 2024, Wang et al. 2024). Furthermore, we tailor the examination of each bias to operational problems by introducing two novel (and as a result, untested and unseen) variations. By considering a breadth of biases and a depth of model and problem comparisons for each bias, we provide deeper insights into patterns of behavior that may emerge as firms and consumers integrate GPTs into their decision-making processes. Thus, our approach moves beyond the extant literature to facilitate establishing behavioral microfoundations underpinning GPT responses to decision problems relevant to OM.

3. Method and Experimental Protocol

To explore the behavioral decision biases of GPT, we followed *The Handbook of Behavioral Operations* chapter "Biases in individual decision-making" by Davis (2018), who introduced a comprehensive list of 18 well-established and prevalent behavioral biases most relevant to OM decision making; recall Figure 1. For each bias, we source experimental instructions from the original research studies (typically from the experimental economics or psychology literature) and perform the experiments with GPT closely following the original instructions. We compare the observed GPT outcomes in such *standard* contexts with the *prototypical* human behaviors reported in those studies.

We then rewrite the standard tests into vignettes relevant to OM. Specifically, we create descriptions of the problems pertaining to inventory (procurement, sourcing, etc.) and to other operational decision making (machine maintenance, etc.). We refer to these two sets of tests as *inventory* and *operations*. We try to keep the essence of the tests identical to the standard tests but frame them in distinctly operational contexts. Doing so provides dual benefits. First, we ensure that the language of the tests is "unseen" by the models to reveal their decision making under real-world scenarios, and second, we specifically evaluate GPT's behavior regarding OM decision making. The overarching goal of the OM vignette development is to produce results comparable with reference studies in humans. Accordingly, the new vignettes should be reflective of how the corresponding biases could manifest in OM and with model conditions that apply to "average users" of GPTs. Thus, we refrain from drastically changing the vignettes or applying complex prompt engineering techniques.

We performed the experiments on GPT-4 and two versions of GPT-3.5. First, the *time 1 study* was performed for the standard context using the January 30, 2023 version of ChatGPT—one of the earliest versions publicly released. The data were collected between January 31 and February 4, 2023 using the default settings of the web interface of ChatGPT as the API was not yet released. Because of the laborious nature of the manual data collection, only 10 responses were collected per bias. Nevertheless, the time 1 study gave us an important benchmark to understand the evolution of LLM decision making. It also provided insight into how to best design the larger-scale automated *main (time 2) experiment* that we performed next.

Specifically, in the time 1 study, we observed that ChatGPT tends to avoid definitive answers when asked to take a "best guess" without access to all necessary information. These conditions are, however, common in the economics and behavioral OM experiments. Thus, in circumstances where a preference is required, we make minor modifications to the original instructions (e.g., instead of asking "what is your preference," we ask "which option is better"). For task-based tests, such as those for risk aversion, framing, or regret, we perform a basic prompt engineering technique; instead of prompting "Q: Which is better?," we prompt "Q: Which is better? A: []" to persuade GPT to provide an immediate answer. Although the definitive answers could often be obtained by prompting a follow-up question, we opt for a structured prompt for more streamlined testing, analysis, and interpretation. Additionally, in tests with a "correct answer," we also follow up with the question "How confident are you about your previous answer (0%–100%)?" to obtain a calibration to examine GPT's level of overconfidence.

Please refer to Online Appendix A for the details of the standard, inventory, and operations test vignettes, and refer to Online Appendix C for the detailed information and results of the time 1 study.

3.1. Main (Time 2) Experiment: Data Collection

We prompt our test vignettes on two GPT models: GPT-3.5-turbo (referred to as GPT-3.5 in this paper) and GPT-4 on their June 11, 2023 versions of APIs. With three framings (standard, inventory, and operations) on both GPT models, we thus perform six sets of tests under each bias. A set of tests may further contain multiple conditions in accordance to the design of the reference studies. In each experimental condition, we perform 30 independent API calls with a temperature of one, OpenAI's default setting. This setting generates answers with a moderate level of variability, allowing us to observe the distribution of GPT's potential choices and preferences when queried repeatedly. We also use the default system prompt "I am a helpful assistant" in our API calls for it is not yet a prompt accessible to the web users. To maximize the generalizability of our results, we opt for default parameter values to meet the overarching goal of studying model responses that an average user would expect as opposed to what a sophisticated "power user" may obtain by prompting the model in very specific conditions. The API outputs are then cleaned by a human reader as GPT sometimes generates unexpected or unstructured outputs—especially GPT-3.5 because of its lower capacity to adhere to instructions. The majority of the API outputs in our study require limited processing before they can be analyzed. The code and API parameter settings for the data collection are available in Online Appendix A.

3.2. Data Analysis

We first test for the existence of the bias; that is, we compare GPT's responses and the unbiased responses as a reference. Then, we make four relative comparisons.

1. GPTs versus human prototypical behavior. See Section 4.1 for results. These results are based on the time 2 data. For the time 1 data, please see Online Appendix C for detailed results.
2. GPTs in the standard versus OM contexts. See Section 4.2 for results.
3. GPT-3.5 versus GPT-4 *models*. See Section 4.3 for results.
4. GPT-3.5 time 1 versus time 2 *versions*. See Section 4.4 for results.

We apply appropriate statistical tests for each bias and comparison to support our claims, including nonparametric tests and regression-based methods. R version 4.3.2 was used for data analysis. Conditions that generate rare responses are handled with simulation-based approximations. Sometimes, GPT generates responses that are ambiguous (e.g., in a choice between A and B, it chooses neither); we classify these responses into a "no preference" group in data cleaning and analyze them with the rest of the responses to stay unbiased. Note that this is essentially identical to how one would deal with humans who fail attention/manipulation checks. To reduce the number of hypothesis tests, we perform post hoc analyses with specific response groups if we need to identify the source of an effect and only if we observe an overall difference. The details of statistical methods for each of the 18 biases are provided in Online Appendix B.

To account for potential multiple comparison issues when performing multiple post hoc tests on the same data set, we implemented a correction for multiple comparisons without direct *p*-value adjustments. Rather, we strengthen the significance *p*-value threshold from the typical 0.05–0.01, and we also encourage readers to treat *p*-values as continuous measures reflecting the compatibility between the data and null hypothesis. This practice is in line with the American Statistical Association's guideline to *p*-values (Greenland et al. 2016, Wasserstein and Lazar 2016). We perform the hypothesis tests and report *p*-values as follows.

- We raise the requirement for a qualitatively significant result from 0.05 to 0.01. That is, we interpret only $p \leq 0.01$ as "significant." This is equivalent to a Bonferroni correction of five multiple testings throughout all of our results, ensuring that our positive findings are robust.
- We interpret *p*-values in (0.01, 0.05] as "borderline" and present them as continuous values. We would like to encourage the reader to adjudicate whether these *p*-values are sufficient evidence of incompatibility between the data and the null hypothesis. When a binary representation of the result is necessary, such as to determine the symbols in Tables 1–3, we present the direction of the associated effects but also note that they are borderline.
- We interpret *p*-values in (0.05, 1] as being compatible with the null and present them as continuous values.

4. Results

In this section, we provide an overview of the results of our analysis. For brevity, in the main body of the paper,

Table 1. Summary of Results for Judgments Regarding Risk

Bias	Prototypical behavior	Model	Standard	Inventory	Operations
Hot-hand fallacy	Exhibiting autocorrelations in randomization tasks	3.5	✓	✓	✓
		4	✓	✓	✓
Conjunction fallacy	Conjunction event being more likely than a component event	3.5	✓	★	✓
		4	✓	✓	✓
Availability heuristics	Overreliance on immediate or accessible examples	3.5	★	★	★
		4	★	★	★
Base-rate neglect	Ignoring relevant probabilities in favor of case-specific information	3.5	★	★	★
		4	★	★	★
Probability weighting	Overweighting low-probability events	3.5	✓	✗	✓[a]
		4	✓	✗	✗
Overconfidence	Overestimating performance on tasks	3.5	✓	✓	✓
		4	✓	✓	✓
Ambiguity aversion	Preferring choices with definite probabilities	3.5	★	★	✓[a]
		4	✓	✗	✓

Note. ✓, GPT exhibits human bias; ✗, GPT exhibits a different bias; ★, GPT acts rationally.
[a] Borderline.

Table 2. Summary of Results for Evaluations of Outcomes

Bias	Prototypical behavior	Model	Standard	Inventory	Operations
Risk aversion	Preferring lower-risk options over higher expected reward	3.5	✗	✗	✗
		4	✓	✓	✗
Prospect theory	Risk averse in gains, risk seeking in losses	3.5	✗	✗	★
		4	✗	✗	★
Framing	Risk averse in gain frame, risk seeking in loss frame	3.5	★	★	★
		4	★	★	★
Anticipated regret	Significant effect of the regret salience manipulation	3.5	★	★	★
		4	✓	★	✓
Mental accounting	Preference for joint over separate payoffs in mixed gains	3.5	★	★	★
		4	★	✓	✗
Reference dependence	Mental accounting preferences depend on frames	3.5	✓	★	✓
		4	✗	✓	★
Intertemporal choice	Discount factor decreases in time and size of payoffs	3.5	✗	✗	✗
		4	★	★	★
Endowment effect	Pronounced gap between willingness to accept and pay	3.5	★	✗[a]	✓
		4	✓	✓	✓
Sunk cost fallacy	Altering investment decisions given the presence of sunk costs	3.5	★	★	✓[a]
		4	★	★	★

Note. ✓, GPT exhibits human bias; ✗, GPT exhibits a different bias; ★, GPT acts rationally.
[a] Borderline.

we focus on the high-level summary of findings, and we refer the reader to Online Appendix B for the comprehensive analyses of each bias across the standard, inventory, and operations contexts. The results are summarized across three tables: Table 1 for biases relating to risk judgment, Table 2 for biases in outcome evaluation, and Table 3 for heuristics in decision making. Each table succinctly describes the biases and summarizes whether GPT-3.5 and GPT-4 responses align or deviate from human behavior across all three contexts. To classify behaviors, we evaluate each scenario against the prototypical human bias. For example, for the hot-hand fallacy, the task is to "randomly generate 50 fair coin tosses" in the standard context, and humans exhibit significant negative autocorrelation. Therefore, we test whether there is autocorrelation between subsequent coin toss outcomes in the GPT data. Rejecting the null because of a significant negative autocorrelation would constitute human-like behavior, which we indicate with a ✓ in Tables 1–3. If we reject the null because of strong evidence of a positive autocorrelation, this would constitute biased behavior that is different from humans, which we indicate with an ✗ in Tables 1–3. Failing to reject the null would lead to us finding evidence of rational decision making, which we indicate with a ★ in Tables 1–3. In most cases, a ★ in Tables 1–3 indicates a failure to reject the null.[13]

Evaluating the behaviors across contexts provides a level of consistency for each decision bias. If behavior is consistent in all three contexts (e.g., all are rational), then in our analysis, we classify the contexts as consistent.[14] If behavior is similar in two contexts (e.g., is biased like humans in two scenarios and rational in the third scenario), then we classify the bias as being somewhat consistent. If behavior is different in all three (e.g., has human-like bias in one scenario, is unbiased in another scenario, and exhibits a bias that is nonprototypical in the third scenario), then we classify the behavior as inconsistent.

In Sections 4.1–4.4, we report on the four comparisons as outlined above, and then, in Section 4.5, we review the recurring patterns that we observe in how GPT behaves across biases, versions, and models.

4.1. GPTs vs. Prototypical Human Behavior in Standard Contexts

Examining Tables 1–3 we observe that in 15 instances of 36 (18 biases for both GPT-3.5 and GPT-4), GPT mirrors human biases, whereas in 21 instances, it diverges

Table 3. Summary of Results for Heuristics in Decision Making

Bias	Prototypical behavior	Model	Standard	Inventory	Operations
Cognitive reflection	Propensity for relying on system 1 heuristics	3.5	★★★	★★✗	★✗✗
		4	★★★	★★★	★★★
Confirmation bias	Seek evidence supporting prior beliefs	3.5	✓	✗	✓
		4	✓	✓	✓

Note. ✓, GPT exhibits human bias; ✗, GPT exhibits a different bias; ★, GPT acts rationally.

from prototypical human responses. The direction of GPT's divergence largely depends on the decision-making category. In judgments regarding risk (Table 1) and heuristics in decision making (Table 3), which are primarily objective tasks,[15] GPT tends to make more rational decisions. For example, in base-rate neglect (Online Appendix B.1.4) or cognitive reflection (Online Appendix B.3.1), GPT produces results that are mostly bias free. In contrast, for evaluations of outcomes (Table 2), which are predominantly subjective preferences, the departure from human biases does not necessarily mean that GPT models are making decisions using rationality. Instead, we observe a different pattern of reasoning. For example, in prospect theory, humans are risk averse in gains and risk seeking in losses. In contrast, GPT-3.5 is risk averse in losses and risk seeking in gains, whereas GPT-4 is risk averse regardless of framing (Online Appendix B.2.2).

4.2. GPTs in the Standard vs. OM Contexts

Tables 1–3 document each GPT model's consistency of responses across the three contexts and reveal a notable level of stability. Most responses were either fully consistent (20 cases) or somewhat consistent (13 cases), whereas only two vignettes (mental accounting/reference dependence with GPT-4 and endowment effect with GPT-3.5) display inconsistency across all three contexts. In general, the degree of response consistency is stable across the three types of decision biases; the percentages of full consistency cases in risk judgment, outcome evaluation, and heuristics in decision making are 64%, 50%, and 50%, respectively

However, we observe two notable context-dependent effects.

4.2.1. Shift in Baseline Risk Tolerance. Although our OM vignettes were designed to mimic the bias of the standard context, there is a clear difference in the level of urgency and "seriousness" between the typical psychological contexts and those of a business decision. Consider probability weighting. The standard context is a Russian roulette game—a life-or-death decision. The OM context choices concern supplier delays and machine repairs, respectively; see Online Appendix A.1.5 for the full vignettes. The inherently different type of consequences between them may explain why both versions of GPT are more likely to overweight the lower probability in a Russian roulette game (eliminating the bullet completely) but are more likely to overweight the higher probability in OM settings (addressing processes more prone to errors). Although we do not observe framing effects in the main study regarding gains and losses specifically (see Online Appendix B.2.3), it is possible that the underlying "gravity" of the framing/context itself may trigger a shift in the model's risk tolerance. This is relevant if businesses try to apply GPT technologies to high-stakes industries, such as healthcare or defense, because the shift in baseline risk behavior may dominate other biases.

4.2.2. Training Data Effect. An additional benefit of reframing the vignettes into OM contexts is that unlike the classical (standard context) experiments, which are widely discussed in scientific literature and therefore, may be present in GPT's training data, the reframed vignettes are new to GPT. Testing GPT with only the standard context vignettes may thus result in GPT performing in a less biased manner than it otherwise would. We looked for cases in which GPT performs more rationally in the standard context but does not perform rationally in *both* OM contexts. For GPT-3.5, there is only one such bias, endowment effect, where GPT behaved rationally in the standard context but was classified as different (borderline) and human-like in the inventory and other operations contexts, respectively. In GPT-4, the model did not perform significantly better in the standard context than in *both* of the OM contexts in any of the 18 biases tested. These results suggest that whether the solution to a particular behavioral bias exists in the training data may not matter much for the GPT's biases.

4.3. Model Effects: GPT-3.5 vs. GPT-4

In this section, we discuss the observed similarities and differences between the biases of GPT-3.5 and GPT-4 models for each of the three groups of biases per Tables 1–3.

4.3.1. Biases in Judgments Regarding Risk. Table 1 shows that the two GPT models that we study often behave similarly to humans. In some cases, the human-like prototypical behaviors persist in GPT-3.5 and GPT-4 at similar levels. For example, in the standard context, these include conjunction fallacy (Fisher's exact test yields $p = 1$) (see Online Appendix B.1.2), probability weighting ($\chi^2 = 6.16$, $p = 0.1069$) (see Online Appendix B.1.5), and overconfidence (see Online Appendix B.1.6). However, in other cases of prototypical behaviors, GPT-4 may adopt the human-like biased behavior to a greater extent than GPT-3.5. For example, for the gambler's fallacy scenario, GPT-3.5 shows significant negative lag 1 autocorrelations in generating sequences of 50 random coin tosses (correlation coefficient = –0.1079, 95% confidence interval (CI): –0.1609, –0.0543) (see Online Appendix B.1.1 and Table 5 in the Online Appendix), implying a tendency toward the gambler's fallacy. However, GPT-4 displays even stronger negative lag 1 autocorrelations (correlation coefficient = –0.3388, 95% CI: –0.3857, –0.2902) (see Online Appendix B.1.1 and Table 5 in the Online Appendix), suggesting a greater adherence to this fallacy relative to GPT-3.5. Similarly, although GPT-3.5 struggles to process ambiguous information in our tests (thus, exhibiting no ambiguity aversion in the standard context, $\chi^2 = 0.11$, $p = 0.9464$) (see Online Appendix B.1.7), a strong ambiguity aversion emerges in GPT-4 ($\chi^2 = 39.5$, $p < 0.001$) (see Online Appendix B.1.7) as it becomes more capable in handling ambiguity.

When GPTs diverge from human biases, both models tend to be more rational than humans. Moreover, GPT-4 is decidedly more rational in its responses compared with GPT-3.5. For example, in the availability heuristic test, both models behave differently and with less bias compared with humans. However, GPT-4 shows higher accuracy than GPT-3.5 (odds ratio (OR) of GPT-4 making errors versus GPT-3.5: 0.13, $p < 0.0001$) (see Online Appendix B.1.3 and Table 6 in the Online Appendix), indicating an improvement in overcoming this bias. Similarly, in the base-rate neglect problem, GPT-3.5 demonstrates significantly less bias in its estimates compared with humans, often providing unbiased dominant responses (Fisher's exact test yields an OR of 27.41 with a $p < 0.0001$ between GPT-3.5 and a human's odds of providing the correct answer) (see Online Appendix B.1.4). GPT-4 further reduces this bias to the point that we do not observe a single incorrect answer in our experiment (Fisher's exact test against human performance yields $p < 0.0001$, OR is infinity) (see Online Appendix B.1.4). These results indicate that GPTs with greater capabilities are likely to generate more accurate responses for problems with factual answers as opposed to problems based on preferences or intuitions.

4.3.2. Biases in Evaluation of Outcomes.

Comparing responses between GPT-3.5 and GPT-4 in Table 2 reveals shifts in the model's decision making, with three of eight vignettes showing decision making that aligns more with the prototypical human bias in GPT-4.

- GPT-3.5 demonstrates a tendency toward risk-seeking behavior in the lottery tasks (e.g., risk aversion and prospect theory), often choosing the riskier option over the safer one, even when the expected values are identical. In contrast, GPT-4 demonstrates risk aversion, favoring less risky outcomes rather than basing choices strictly on expected values. For example, contrasting test 1 and test 2 for choosing between safe and risky lotteries in the risk aversion experiment in the standard context (see Online Appendix B.2.1), GPT-4's propensities to select a risky lottery decrease by 84% and 91%, respectively, as compared with GPT-3.5 ($p < 0.001$).
- When facing decisions involving regret, GPT-3.5 consistently opts to take action, irrespective of whether action or inaction regrets are made salient. GPT-4, however, makes decisions that clearly reflect the prompt's regret salience (Online Appendix B.2.4).
- Although GPT-3.5 did not show a clear bias toward the endowment effect—making similar decisions regardless of being a seller or a buyer—GPT-4 exhibits behavior consistent with the endowment effect (Online Appendix B.2.7).

Additionally, in one case (reference dependence), GPT-4 shifts away from the prototypical behavior observed in GPT-3.5. However, instead of shifting toward unbiased responses, GPT-4 exhibits the opposite behavior of what reference dependence would predict, essentially demonstrating the reference dependence of a different "flavor."

Taken together, it seems that for biases in outcome evaluations, the more capable GPT-4 shows no signs of shift toward unbiasedness and curiously, displays quite the opposite. This may be because of the nature of the tasks in this category, which tend to be more complex and preference based. Unlike CRT and similar tasks, outcome evaluation tests do not have "correct answers." This may create additional challenges for the models to learn the "rational" behavior. The relative lack of improvements in outcome evaluations suggests that companies may need to remain careful when employing GPTs for decision-making tasks involving preferences. Company-specific research and testing may be needed for businesses to "align" the preferences of its AIs with the desired behavior.

4.3.3. Heuristics in Decision Making.

The progression from GPT-3.5 to GPT-4 mirrors patterns that we observe both in judgments regarding risk (movement toward unbiased behaviors) and in evaluation of outcomes (movement toward prototypical human behaviors).

For the cognitive reflection tests, although GPT-3.5 predominantly provides answers with system 2 (i.e., correct) responses (average correct items from standard CRT are 2.7 of 3, 95% CI: 2.53, 2.87) (see Online Appendix B.3.1), GPT-4 delivers system 2 responses across all instances (we do not observe mistakes in standard CRT in our sample) (see Online Appendix B.3.1). This is consistent with results in availability heuristics and base-rate neglect experiments, which have correct answers requiring mathematical/statistical calculations. GPT-4, with its superior model capabilities, pushes results further from prototypical human biases toward unbiased behavior.

In scenarios testing for confirmation bias, GPT-3.5 produces responses that align with the prototypical bias, whereas GPT-4 intensifies this pattern, generating responses that always exhibit confirmation bias (χ^2 test on proportions of correct responses in the standard context test, four-card problem, between the two models yields $p < 0.0001$) (see Online Appendix B.3.2). This is consistent with results in risk aversion, anticipated regret, and endowment effect, where GPT-4's responses move further toward prototypical behavior. Interestingly, the four-card problem used for examining confirmation bias also has a correct answer, although it requires logic and reasoning instead of calculations.

Similar to the Linda problem used to examine the conjunction fallacy, GPT fails to recognize the mathematical nature of the problem and appears to apply heuristics, resulting in a strong prototypical bias.

4.4. Version Effects: "Early" vs. "Late" GPT-3.5

LLM vendors frequently update their models. However, they generally omit details on exactly how models differ and what constitutes a minor update (e.g., GPT-3.5-turbo-0613 versus GPT-3.5-turbo-1106) versus a major update (e.g., GPT-3.5 versus GPT-4). We refer to the former as "version" updates and to the latter as the "model" updates. Although the previous section showed that model updates have a substantial impact on GPT's decision-making tendencies, it is unclear whether we would expect to see similar effects with version updates or whether the GPT decision making is relatively stable across different versions of the same model.

To test for this, we performed the standard context tests on two versions of GPT-3.5 at two time points, which we introduced earlier as time 1 and time 2. Time 1 data collection was conducted in January and February 2023 using the January 30, 2023 version of GPT-3.5 (one of the earliest versions) through the web interface because the API was not available then. Time 2 data collection was conducted from October to December 2023 using the June 11, 2023 release of GPT-3.5-turbo API—one of the last available versions of GPT-3.5.

Table 4 summarizes the results of these tests. Note that unlike the rest of the paper, we keep the significance level, alpha, at 0.05 for the initial experiments conducted at time 1. Recall that those experiments were performed on a single context and model with a sample size of 10 because of the manual data collection (recall that the API was not yet released then), making alpha inflation unlikely.

Overall, we find remarkable consistency in GPT-3.5's behavior over time. In 14 of 18 biases, the results are qualitatively identical. In 3 of 18 biases (hot-hand fallacy, risk aversion, and prospect theory[16]), the differences in GPT's decisions can be attributed to the operationalization of data collection, such as statistical power, prompt structure, web versus API, etc. Only in 1 of 18 cases, namely anticipated regret, do we find some evidence of a shift in the model's preferences. We found that in both experiments, the default recommendation from GPT was to take an action. However, in the time 1 experiment, the increased salience of action regret would push the model from recommending action to not making any recommendation (no preference), showing that the model was sensitive to regret salience ($p < 0.001$) (Online Appendix C.2.4). In contrast, in the time 2 experiment, GPT always recommended taking the action, regardless of the action or inaction regret salience (Online Appendix B.2.4).

4.5. Decision Patterns

Although the literature on evaluating decision biases in LLMs is quickly emerging, much of the literature focuses on a small selection of biases, preventing researchers from drawing more general conclusions about the behavioral patterns. By examining 18 biases, we uncover commonalities in GPT's decision-making processes across biases. Although each set of vignettes tests distinct biases, several common and overlapping features (e.g., risk, information salience, and probability) manifest across multiple vignettes. The patterns in GPT's responses can, therefore, help us establish microfoundations of GPT's behavior and decisions. In the assessment below, we focus on GPT-4, which has clearer patterns of divergence or convergence to prototypical behaviors depending on the task type.

4.5.1. Risk and Certainty. GPT-4's decision making often reflects a clear preference for certainty, influencing whether its responses deviate or replicate human behavior. GPT-4's strong inclination toward certainty leads GPT-4's approach to differ from humans in the context of framing effects and prospect theory. It consistently chooses certainty over gambles across situations that vary in their framing, despite the gambles

Table 4. Summary of Results for Two Versions of the GPT-3.5 Model Comparing Time 1 and Time 2

Judgment regarding risk			Evaluations of outcomes			Heuristics		
Bias	T1	T2	Bias	T1	T2	Bias	T1	T2
Hot-hand fallacy	★	✓	Risk aversion	★	✗	Cognitive reflection	★	★
Conjunction fallacy	✓	✓	Prospect theory	✗	✗	Confirmation bias	✓	✓
Availability heuristics	★	★	Framing	★	★			
Base-rate neglect	★	★	Anticipated regret	✗	★			
Probability weighting	✓	✓	Mental accounting	★	★			
Overconfidence	✓	✓	Reference dependence	✓	✓			
Ambiguity aversion	★	★	Intertemporal choice	✗	✗			
			Endowment effect	★	★			
			Sunk cost fallacy	★	★			

Note. T1, time 1; T2, time 2; ✓, GPT exhibits human bias; ✗, GPT exhibits a different bias; ★, GPT acts rationally.

having similar expected payoff. GPT-4 also shows a preference for guaranteed outcomes under domains of gains and losses, diverging from the propensity to gamble under losses as suggested by prospect theory. In other cases, the preference for certainty leads to the behavior consistent with humans. For example, GPT-4's avoidance of ambiguity reflects a preference for known (and certain) over unknown risks. As Davis (2018) describes it, "the devil you know is better than the devil you don't" (Davis 2018, p. 168).

4.5.2. Information Salience. We observe that whether information salience biases GPT-4's decision making depends on whether the scenarios involve a readily accessible calculable solution. For example, the anticipated regret and the conjunction fallacy scenarios do not have concrete values to apply formulaically. In these cases, GPT-4 weights the salient information as a basis for its decision making, even if not relevant. Conversely, in situations where a precise formula is applicable, such as with the availability heuristic and base-rate neglect, GPT-4 can identify and apply these formulas, sidelining less relevant information. However, this tendency is not absolute. The endowment effect is a notable exception, indicating that how information is framed can still significantly impact GPT-4's decision making, which overrides its capacity to disregard irrelevant details.

4.5.3. Probability and Statistics. GPT-4 exhibits a mixed but generally superior performance compared with the typical human behavior in probability tasks; however, its effectiveness varies. For example, it displays a tendency toward the gambler's fallacy and the conjunction fallacy, areas where calculations are not immediate. In contrast, GPT-4 does not fall for the base-rate neglect, where humans fail to engage in calculations with a counterintuitive result. Relatedly, GPT-4 is less prone to the sunk cost fallacy, showing an aptitude for disregarding previously invested resources when making decisions about future actions as the objective, quantitative reasoning would dictate.

5. Discussion

Organizations are integrating LLMs to enhance operations, supply chain management, and customer interactions because of potential gains in efficiency, accuracy, and scalability. Concurrently, consumers are also adopting LLM tools, like ChatGPT, for advice and recommendations to support decision making across routine activities. Given the increasing reliance on ChatGPT by both organizations and consumers, we examine whether it adheres to principles of rational decision making or if it exhibits the behavioral biases often found in humans. Understanding when ChatGPT mimics or diverges from human decision making is essential for effectively leveraging its capabilities.

It is likewise critical to realize that the "human-like" behavioral biases of LLMs could also be very different from those of prediction models (i.e., "traditional" AI). An example from Agrawal et al. (2022) distinctly highlights the differences in prediction strategies between humans and traditional AI. In a classical psychology experiment that they reference, humans tasked with predicting the next element in a sequence like OXXOXOXOXOXXOOXXOXOXXXOXX tend to randomize between X and O, with a slight preference for X given its higher (60%) appearance rate. This approach results in a prediction accuracy of 52%, just over the chance of coin flipping. Agrawal et al. (2022) pointedly note that if you want to maximize your chances of a correct prediction, you would always choose X. "What such experiments tell us is that humans are poor statisticians … No prediction machine [AI] would make an error like this" (Agrawal et al. 2022, p. 68), which emphasizes the gap between human intuition and statistical optimization of a typical machine learning algorithm. Yet, our findings with ChatGPT suggest that it approaches decisions in a manner similar to humans: for example, regarding the hot-hand and gambler's fallacies. When prompted to predict the next three letters in a sequence, ChatGPT's prediction reflects a human-like approach: "[A] cautious prediction for the next three letters, considering the desire to maintain some level of alternation and the recent prevalence of 'x's, could be: 'oxo' or 'oxx.'" This indicates a preference for alternation and a slight bias toward "x's," akin to human reasoning patterns, and diverges from the purely statistical approach expected of traditional AI.

This example illustrates that LLMs, like ChatGPT, adopt a decision-making process that significantly diverges from traditional AI models analyzed in the economics and management literature. Moreover, the similarity to human reasoning patterns requires a systematic examination of biases as it marks a pivotal shift in our understanding of AI's capabilities and limitations in replicating human decision making. Without understanding when GPT exhibits biases like humans and when it behaves closer to traditional AI, organizations risk misinterpreting LLMs capabilities in complex decision-making environments, potentially leading to inefficiencies in operational and strategic decisions. Thus, understanding GPT's decision making is crucial for integrating it into OM problems as it will enable better decision making where ChatGPT is unbiased and can identify when to use prompt engineering to counteract biases to enhance performance.

To contextualize our understanding of GPT's decision-making biases within the context of OM, we examined 18 relevant biases documented by Davis (2018). Analyzing these biases allows us to uncover the microfoundations of GPT's decision making, offering insights into its logic and potential biases, which have several key benefits. First, it helps identify the extent to which GPT models replicate or diverge from human reasoning patterns, which can be both an advantage and a limitation depending on the context. Second, it sheds light on the potential for GPT to overcome human biases—or alternatively, to exhibit new forms of biases inherent in its training data or algorithms. Third, uncovering where GPT exhibits biases can help better predict its behavior in complex decision-making environments, offering valuable insights for developers, users, and policymakers. Although our focus is on OM, analyzing these biases contributes to the broader discourse of LLMs' decision making as these biases underpin decision making across organizational and consumer contexts. Together, our results have the following implications for managers.

1. Our results highlight that GPT's decision making exhibits a relatively consistent pattern. When faced with subjective decisions, GPT tends to favor lower-risk outcomes and strongly prefers certainty. In objective scenarios, GPT's approach is methodical. It first determines if a calculable solution is readily available for the problem at hand. If it can identify the formula, GPT applies it to derive an answer. However, when faced with objective questions that lack a straightforward formulaic answer, GPT resorts to heuristic-based reasoning, mimicking "system 1" thinking akin to human instinctual responses. In many ways, this approach resembles how humans behave except that GPT's ability to recognize applicable formulas and apply them accurately surpasses humans. These results indicate that organizations will reap the greatest benefits by deploying GPT within more objective workflows that align with established formulas, particularly those in which humans may struggle because of limited cognitive capacity.

2. Despite the high degree of sensitivity to prompt context reported in other studies (e.g., Dou 2023), where even small wording changes can significantly alter GPT's responses, in our study GPT displayed remarkable response consistency across the standard, inventory, and operational contexts. On the surface, this result is surprising, especially considering the significant variations in the vignettes' contexts and lengths. However, the fact that GPT had a systematic approach to solving problems can explain why its decisions were largely consistent across contexts. For managers, these results provide optimism that LLMs can offer reliable support even when decision and problem contexts change.

3. Selecting between models, such as GPT-3.5 and GPT-4, represents a trade-off in cost and performance, which managers must consider when integrating GPT into their organizations. At the time of conducting our experiments, the free version of ChatGPT was based on the GPT-3.5 model, whereas the premium version used GPT-4. Our results demonstrate that GPT-4's performance improves over GPT-3.5, particularly for solving problems with objective solutions, suggesting that managers should investigate the trade-offs between the model performance and cost implications based on the nature of the tasks.

Given that the behaviors of LLMs are black boxes, our findings also establish microfoundations for future research in behavioral OM involving AI decision making at the manager, value-chain partner, and consumer levels as the two examples below illustrate:

Newsvendor problem. There is substantial evidence that newsvendor decision making is influenced by a myriad of biases. For example, human newsvendors exhibit gambler's fallacy behavior (e.g., Bolton and Katok 2008), risk aversion (e.g., Becker-Peth et al. 2018), and overprecision (a form of overconfidence) (Ren and Croson 2013), leading to the so-called "pull-to-center" effect. Long and Nasiry (2015) showed that prospect theory can explain the pull-to-center effect observed in newsvendor experiments by setting a reference point that reflects a newsvendor's most salient payoffs. Taken together, these and other related studies built a theory of newsvendor behavior based on the microfoundations of more generic human biases. Importantly, such a theory allowed researchers to extrapolate what behaviors might emerge in other contexts, such as contract design (e.g., Becker-Peth et al. 2013) and newsvendor competition (e.g., Ovchinnikov et al. 2015), which in turn, provided firms with strategies to mitigate the impact of biases or even capitalize on them.

We show that GPT is also risk averse and overconfident, often to an even larger degree than humans, which suggests that the aforementioned strategies may continue to work well should buyers outsource ordering decisions to GPT agents. Not surprisingly, recent GPT experiments on the newsvendor problem (e.g., Su et al. 2023, Kirshner 2024a) show that GPT is impacted by biases, like risk aversion and demand chasing. Our results provide evidence that this is not driven by the problem but instead, is driven by the microfoundations of GPT's behavior.

Wait-or-buy and wait-and-buy problems. Baucells et al. (2017) found that incorporating consumers' behavioral anomalies into markdown optimization may increase revenue over the standard models that assume that consumers are rational. Among other biases, they considered hyperbolic discounting. Our study, on the other hand, shows that GPT does not exhibit the hyperbolic discounting bias. If consumers sought purchasing advice from ChatGPT, the documented revenue increase would be smaller. However, as consumers would make more rational decisions, firms may also benefit, leading to a Pareto improvement. At the same time, Özer and

Zheng (2016) found that firms can increase revenues by incorporating consumers' anticipated regret into their markdown optimization strategies. Because GPT is sensitive to anticipated regret, seeking advice from GPT may not change the overall implications of their findings. Thus, by exploring biases of GPT, our research can help generate predictions on how consumers leveraging ChatGPT will react to operational pricing problems. In a related problem studied by Kremer and Debo (2016), consumers decide whether to wait to buy a product, receiving cues on quality from the queue's length. Kirshner (2024a) finds that ChatGPT does not recommend waiting in line when the queue length is ambiguous. Our research provides an explanation for this result because our results show that GPT exhibits ambiguity aversion.

As these two examples illustrate, there is a distinct benefit for researchers to examine the microfoundations of behaviors with respect to each "individual" bias in addition to the "multibias" behaviors that reveal themselves at the operational problem level, like in the newsvendor or wait-or-buy problems. This is not much different from similar research with humans, with the exception of the chronological divide. For humans, the psychology and behavioral economics literature established the individual biases first, and the behavioral operations literature exhibited the problem-wide behaviors second. With GPT, both seem to progress in parallel.

Similarly, the detailed comparison between GPT-3.5 and GPT-4 also offers researchers insight into the evolution of GPT's ability to handle decision biases. Although GPT-4 has become more accurate overall, its behavioral tendencies have also been amplified. For example, GPT-4 has improved in certain areas of heuristics and intuitive statistics (e.g., it did not exhibit base-rate neglect). Yet, it also increased its propensity toward confirmation bias and the conjunction fallacy, which are also areas of heuristics and intuitive statistics. In addition, our analysis reveals that GPT-4 has a greater aversion to risk compared with GPT-3.5, which indicates that OpenAI developers may be implementing specific guardrails impacting how GPT evaluates uncertainty. Extrapolating the results of increased guardrails suggests that as organizations increasingly integrate LLMs into their decision making, they may lean toward taking less, potentially insufficient risk. Lastly, we have also pinpointed areas where GPT exhibits decision making consistent with human biases. This opens avenues for researchers to explore developing interventions through prompt engineering to enhance GPT's performance by mitigating these specific biases but not others.

Although we consider a substantial number of biases, our work is not meant to be an exhaustive study of biases in LLMs or their implications for behavioral operations. Rather, this research serves as a starting point. Beyond examining other biases, follow-up studies can also investigate the biases considered in our study with other LLMs and with other experiments testing these biases for robustness. Similarly, our experiments were conducted in English; however, recent research indicates that behavioral biases can also depend on language (e.g., Leng 2024). In addition, future research can test how prompt engineering may be used to mitigate scenarios in which GPT exhibits biases, first within vignette settings and then within more complex, multiround, and multiplayer decisions, going beyond the zero-shot single decision-making settings that we analyzed. Finally, we note that verifying that GPT truly "understands" a question is complex, and knowing the source of biased responses is challenging. Nevertheless, knowing when an AI (acting as an assistant or a decision maker) *appears* biased is important, even if the root causes for such biases may be fundamentally different from those driving biases in human behavior.

Acknowledgments

The title of this paper was created with the assistance from ChatGPT as follows. The authors asked ChatGPT to suggest a humorous and catchy title based on the paper's introduction. It initially offered "A Manager and a Bias Walk into a Bar." The authors adapted this to "A Manager and an AI … " to better reflect the parallel structure, and then, they added a more descriptive subtitle (" … : Does Chat GPT Make Biased Decisions Like We Do?").

Endnotes

[1] See https://www.similarweb.com/blog/insights/ai-news/chatgpt-notebooklm/ (accessed November 16, 2024).

[2] See https://www.statista.com/outlook/tmo/artificial-intelligence/generative-ai/worldwide\#users.

[3] See https://www.cnbc.com/2023/04/08/chatgpt-is-being-used-for-coding-and-to-write-job-descriptions.html.

[4] See https://www.forbes.com/sites/bernardmarr/2023/03/21/revolutionizing-retail-how-chatgpt-is-changing-the-shopping-experience/.

[5] See https://www.gartner.com/en/webinar/464445/1096048.

[6] See https://www.fool.com/money/research/chatgpt-credit-card-recommendations/.

[7] See https://www.pwc.com/gx/en/industries/consumer-markets/consumer-insights-survey.html.

[8] See https://www.pwc.com/us/en/industries/industrial-products/library/industrial-products-trends.html.

[9] See https://www.businessinsider.com/walmart-using-ai-to-negotiate-deals-with-some-equipment-suppliers-2023-4.

[10] See https://openai.com/blog/how-should-ai-systems-behave/ (accessed February 20, 2023).

[11] See https://help.openai.com/en/articles/6783457-chatgpt-general-faq (accessed February 16, 2023).

[12] Research on decision making with LLMs has also expanded beyond studying biases to include economic scenarios. For example, Horton (2023) conducted four behavioral economics experiments with GPT-3, exploring areas such as social preferences, fairness, status quo bias, and minimum wage perceptions. Chen et al. (2023)

analyzed how GPT makes budgetary decisions involving risk, time, social, and food preferences, discovering that GPT generally makes choices that align with utility maximization. The economic-focused literature on GPT decision making mainly explores how GPT performs in economic games, such as the prisoner's dilemma (e.g., Akata et al. 2023), the ultimatum game (e.g., Phelps and Russell 2023), and the dictator game (e.g., Brookins and DeBacker 2023), or it examines its approach to social aspects, like public goods and trust (e.g., Mei et al. 2024). For instance, Leng and Yuan (2024) looked into GPT's social behaviors, including its approach to distributional choices, reciprocity preferences, and its reaction to group identity signals, whereas Xie et al. (2024) observed that LLMs demonstrate trust in the classical trust game setting.

[13] The statistical tests for conjunction fallacy and confirmation bias are with respect to chance selections. If the null is rejected, post hoc tests can determine whether this is because of rational, human-like, or different behavior. The null in the cognitive reflection tests is human behavior. As many of the tests rely on multiple responses (e.g., across two or three frames), whereas the null of no changes across frames corresponds to the rational behavior, rejecting the null results in post hoc tests to determine the behavioral category. We refer the reader to Online Appendix B for specific details for each decision bias.

[14] A bias that is classified as consistent can still have statistical differences across contexts. For example, GPT's average degree of accuracy in the base-rate neglect vignettes differs statistically across contexts (and models). However, on volume, GPT-3.5 does not exhibit the bias in each context. So, we classify the behavior as consistent across contexts.

[15] All of the biases explored in Tables 1 and 3 have objective solutions, with the exception of ambiguity aversion and probability weighting. Although ambiguity aversion and probability weighting have "rational" solutions, the bias represents a preference, similar to the biases in Table 2. In other words, "rational" response could emerge even in situations that do not have "correct" solutions and are purely preference based.

[16] In prospect theory experiments, the two GPT versions both exhibit biases different from humans but toward different directions.

References

Agrawal A, Gans J, Goldfarb A (2022) *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence* (Harvard Business Press, Cambridge, MA).

Akata E, Schulz L, Coda-Forno J, Oh SJ, Bethge M, Schulz E (2023) Playing repeated games with large language models. Preprint, submitted May 26, https://arxiv.org/abs/2305.16867.

Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D (2023) Out of one, many: Using language models to simulate human samples. *Political Anal.* 31(3):337–351.

Baucells M, Osadchiy N, Ovchinnikov A (2017) Behavioral anomalies in consumer wait-or-buy decisions and their implications for markdown management. *Oper. Res.* 65(2):357–378.

Becker-Peth M, Katok E, Thonemann UW (2013) Designing buyback contracts for irrational but predictable newsvendors. *Management Sci.* 59(8):1800–1816.

Becker-Peth M, Thonemann UW, Gully T (2018) A note on the risk aversion of informed newsvendors. *J. Oper. Res. Soc.* 69(7):1135–1145.

Binz M, Schulz E (2023) Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. USA* 120(6):e2218523120.

Bolton GE, Katok E (2008) Learning by doing in the newsvendor problem: A laboratory investigation of the role of experience and feedback. *Manufacturing Service Oper. Management.* 10(3):519–538.

Brand J, Israeli A, Ngwe D (2023) Using LLMs for market research. Preprint, submitted March 30, http://dx.doi.org/10.2139/ssrn.4395751.

Brookins P, DeBacker JM (2023) Playing games with GPT: What can we learn about a large language model from canonical strategic games? Preprint, submitted July 10, http://dx.doi.org/10.2139/ssrn.4493398.

Chen Y, Liu TX, Shan Y, Zhong S (2023) The emergence of economic rationality of GPT. Preprint, submitted May 22, https://arxiv.org/abs/2305.12763.

Dasgupta I, Lampinen AK, Chan SC, Creswell A, Kumaran D, McClelland JL, Hill F (2022) Language models show human-like content effects on reasoning tasks. Preprint, submitted July 14, https://arxiv.org/abs/2207.07051.

Davis AM (2018) Biases in individual decision-making. Donohue K, Katok E, Leider S, eds. *The Handbook of Behavioral Operations* (Wiley, Hoboken, NJ), 149–198.

Davis AM, Mankad S, Corbett CJ, Katok E (2024) OM Forum—The best of both worlds: Machine learning and behavioral science in operations management. *Manufacturing Service Oper. Management.* 26(5):1605–1621.

Davis AM, Flicker B, Hyndman K, Katok E, Keppler S, Leider S, Long X, Tong JD (2023) A replication study of operations management experiments in management science. *Management Sci.* 69(9):4977–4991.

Dou Z (2023) Exploring GPT-3 model's capability in passing the Sally-Anne test a preliminary study in two languages. Preprint, submitted February 9, https://doi.org/10.31219/osf.io/8r3ma.

Fennell E (2023) Action identification characteristics and priming effects in ChatGPT. Preprint, submitted May 12, https://doi.org/10.31234/osf.io/aqbvk.

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG (2016) Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *Eur. J. Epidemiology* 31(4):337–350.

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

Horton JJ (2023) Large language models as simulated economic agents: What can we learn from *Homo silicus*? Preprint, submitted January 18, https://arxiv.org/abs/2301.07543.

Jackson I, Ivanov D, Dolgui A, Namdar J (2024) Generative artificial intelligence in supply chain and operations management: A capability-based framework for analysis and implementation. *Internat. J. Production Res.* 62(17):6120–6145.

Kirshner SN (2024a) Artificial agents and operations management decision-making. Preprint, submitted March 13, http://dx.doi.org/10.2139/ssrn.4726933.

Kirshner SN (2024b) GPT and CLT: The impact of ChatGPT's level of abstraction on consumer recommendations. *J. Retailing Consumer Services* 76:103580.

Kremer M, Debo L (2016) Inferring quality from wait time. *Management Sci.* 62(10):3023–3038.

Leng Y (2024) Can LLMs mimic human-like mental accounting and behavioral biases? Preprint, submitted February 13, http://dx.doi.org/10.2139/ssrn.4705130.

Leng Y, Yuan Y (2024) Do LLM agents exhibit social behavior? Preprint, submitted December 23, https://arxiv.org/abs/2312.15198.

Li P, Castelo N, Katona Z, Sarvary M (2024) Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Sci.* 43(2):254–266.

Long X, Nasiry J (2015) Prospect theory explains newsvendor behavior: The role of reference points. *Management Sci.* 61(12):3009–3012.

Ma D, Zhang T, Saunders M (2023) Is ChatGPT humanly irrational? Preprint, submitted September 22, https://doi.org/10.21203/rs.3.rs-3220513/v1.

Macmillan-Scott O, Musolesi M (2024) (Ir)rationality and cognitive biases in large language models. Preprint, submitted February 14, https://arxiv.org/abs/2402.09193.

Mei Q, Xie Y, Yuan W, Jackson MO (2024) A Turing test of whether AI chatbots are behaviorally similar to humans. *Proc. Natl. Acad. Sci. USA* 121(9):e2313925121.

Meng J (2024) AI emerges as the frontier in behavioral science. *Proc. Natl. Acad. Sci. USA* 121(10):e2401336121.

Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654):187–192.

Ovchinnikov A, Moritz B, Quiroga BF (2015) How to compete against a behavioral newsvendor. *Production Oper. Management* 24(11):1783–1793.

Özer Ö, Zheng Y (2016) Markdown or everyday low price? The role of behavioral motives. *Management Sci.* 62(2):326–346.

Park PS, Schoenegger P, Zhu C (2024) Diminished diversity-of-thought in a standard large language model. *Behav. Res. Methods* 56(6):5754–5770.

Phelps S, Russell YI (2023) The Machine Psychology of Cooperation: Can GPT models operationalise prompts for altruism, cooperation, competitiveness and selfishness in economic games? Preprint, submitted May 13, https://arxiv.org/abs/2305.07970.

Ren Y, Croson R (2013) Overconfidence in newsvendor orders: An experimental study. *Management Sci.* 59(11):2502–2517.

Su J, Lang Y, Chen KY (2023) Can AI solve newsvendor problem without making biased decisions? A behavioral experimental study. Preprint, submitted September 14, http://dx.doi.org/10.2139/ssrn.4567157.

Suri G, Slater LR, Ziaee A, Nguyen M (2024) Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *J. Experiment. Psych. General* 153(4):1066–1075.

Terwiesch C (2023) Would ChatGPT3 get a Wharton MBA? A prediction based on its performance in the operations management course. Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, Philadelphia.

Wamba SF, Queiroz MM, Jabbour CJC, Shi CV (2023) Are both generative AI and ChatGPT game changers for 21st-century operations and supply chain excellence? *Internat. J. Production Econom.* 265:109015.

Wang P, Xiao Z, Chen H, Oswald FL (2024) Will the real Linda please stand up … to large language models? Examining the representativeness heuristic in LLMs. Preprint, submitted April 1, https://arxiv.org/abs/2404.01461.

Wasserstein RL, Lazar NA (2016) The ASA statement on *p*-values: Context, process, and purpose. *Amer. Statistician* 70(2):129–133.

Xie C, Chen C, Jia F, Ye Z, Lai S, Shu K, Gu J, et al. (2024) Can large language model agents simulate human trust behaviors? Preprint, submitted February 7, https://arxiv.org/abs/2402.04559.

Xu R, Sun Y, Ren M, Guo S, Pan R, Lin H, Sun L, Han X (2024) AI for social science and social science of AI: A survey. *Inform. Processing Management* 61(3):103665.