Online Appendix: A Manager and an AI Walk into a Bar

Yang Chen, Sam Kirshner, Anton Ovchinnikov, Meena Andiappan, Tracy Jenkin

The paper includes three appendices. Appendix A describes the vignettes for the Main experiment and presents the sample code to replicate our design. Appendix B presents the results of the Main experiment. Appendix C presents the experiments and results of the Time 1 study (recall, the Time 1 Study was done in early 2023, when the OpenAI API was not yet readily available).

Appendix A: Study Vignettes and Sample Code

In this Appendix we include all details necessary to replicate our results. §A.1 -§A.3 provide our study vignettes, and §A.4 provides the sample Python code for data collection.

A.1.   Biases in Judgments Regarding Risk

A.1.1.   The Hot-Hand and Gambler's Fallacies

Standard Context

Randomly generate 50 fair coin tosses, separated by ','. Represent heads with 1 and tails with 0.

Inventory & Procurement Context - Stocking Decisions

Each day, a firm supplies either 100 or 200 units of a popular product. With a 50 percent probability, the firm stocks an extra 100 units. Randomly decide for the next 50 days when to have additional stock of the product. Separated each decision with ','. Represent having additional stock with 1 and no additional stock with 0.

Other Operational Context - Dynamic Pricing

Each day, with a 50 percent probability, a firm randomly decides whether to put a product on sale. Randomly determine whether the product is on sale for the next 50 days. Separated each decision with ','. Represent sale with 1 and full price with 0.

A.1.2.   The conjunction fallacy and representativeness

Standard Context

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Rank the following statements that could describe Linda from the most likely to the least likely.

A: Linda is a teacher in elementary school.

B: Linda works in a bookstore and takes Yoga class.

C: Linda is active in the feminist movement.

D: Linda is a psychiatric social worker.

E: Linda is a member of the League of Women Voters.

F: Linda is a bank teller.

G: Linda is an insurance salesperson.

H: Linda is a bank teller and is active in the feminist movement.

Inventory & Procurement Context - Supplier Selection

Factory X is a 10-year-old facility located in a developing country with a burgeoning apparel industry. Factory X has developed a strong reputation for its commitment to sustainable practices and for partnering with local communities. This commitment and partnerships have led to it receiving several grants aimed at innovating and improving sustainable production methods within the clothing industry. Rank the following statements that could describe *Factory X* from the most likely to the least likely.

A: Factory X has a comprehensive waste management program.

B: Factory X contributes to reforestation projects.

C: Factory X produces clothing only using organic cotton and eco-friendly materials like bamboo.

D: Factory X exports most of its products to European markets.

E: Factory X pays employees above minimum wage.

F: Factory X specializes in acid-washed denim.

G: Factory X has over 1000 employees.

H: Factory X specializes in acid-washed denim and produces clothing only using organic cotton and eco-friendly materials like bamboo.

Other Operational Context - Sales

A shirt on Instagram is advertised as crafted from repurposed fabric waste, featuring designs made by artisans from a small, eco-conscious community, using biodegradable buttons made from natural materials and packaged in 100% compostable materials. The caption says that the shirt is affiliated with a documentary about climate change's impact on the environment, where a percentage of proceeds go towards environmental efforts.

Rank the following statements that could describe the Shirt from the most likely to the least likely.

A: The shirt is made from bamboo.

B: The shirt is colored using organic dyes.

C: The shirt is part of a sustainable fashion line.

D: The shirt is unisex.

E: The shirt's sales contribute to reforestation projects.

F: The shirt is manufactured by a fast fashion retailer

G: The shirt is red.

H: The shirt is manufactured by a fast fashion retailer and is part of a sustainability fashion line.

A.1.3. The availability heuristic

Standard Context

There are ten stations along a route. Consider a bus that travels, stopping at exactly r stations along this route. What is the number of different patterns of r stops that the bus can make?

Inventory & Procurement Context - Supplier Diversification

You are a procurement manager for a large manufacturing company. For diversification, your company sources components from 10 different suppliers. Due to budget constraints, you need to only order components from exactly r suppliers.

Given the 10 available suppliers, how many different ways can you select exactly r suppliers?

Other Operational Context - Traveling Salesman

You are managing a fleet of delivery trucks for a retail company. There are 10 distribution centers located along a delivery route. For a particular delivery run, a truck will stop at exactly r of these centers to either pick up or drop off inventory.

Given the 10 distribution centers along the delivery route, in how many different patterns can the truck make exactly r stops?

A.1.4. Base rate neglect and Bayesian updating
Standard Context

If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?

Inventory & Procurement Context - Consumer Returns

You manage sales and inventory for an electronics store, where Smartphone X is one of your best-selling products. Historically, 1 in 1,000 units of Smartphone X is defective. When customers return Smartphone X, 5% of the time, the phone is not actually damaged. What is the chance that a customer returning a defective Smartphone X actually has a defective phone?

Other Operational Context - Production Quality

You manage a production line that manufactures large quantities of widgets, of which 1/1000 is defective. To identify and remove defective widgets, you use an automated detection system. If the system has a false positive rate of 5%, what is the chance that a widget identified as defective actually has quality issues?

A.1.5. Probability Weighting
Standard Context

Suppose you are compelled to play Russian roulette but are given the opportunity to purchase the removal of one bullet from the loaded gun. Would you pay as much to reduce the number of bullets from four to three as you would to reduce the number of bullets from one to zero?

Inventory & Procurement Context - Supplier Delivery

Imagine you are the procurement manager for a manufacturing company. You rely heavily on two critical suppliers - Supplier X and Supplier Y - for essential components of your flagship product. Any delay in deliveries could cause significant disruptions in your production line and result in substantial financial losses.

Supplier X has a 2/3 chance of delivering late.

Supplier Y has a 1/6 chance of delivering late.

The financial repercussions from either supplier delivering late are identical since the product cannot be made without both suppliers delivering the components.

You are given an opportunity to invest in a consultancy service that will help a supplier improve their delivery performance. Would you invest as much in improving Supplier X's on-time delivery rate to 1/2 as you would in ensuring Supplier Y's on-time delivery rate becomes perfect (0 chance of being late)?

Other Operational Context - Machine Maintenance

Suppose your manufacturing facility outsources the maintenance of its production machinery to a third-party service provider, which has a pool of 6 technicians. The technician dispatched by the provider following a machinery malfunction is random (uniformly distributed).

Scenario A: In this case, 2 technicians are highly experienced and can swiftly diagnose and rectify any issue after a machinery malfunction, ensuring minimal downtime. However, 4 of the technicians, though competent, are inexperienced. When these technicians address the malfunction, there's a 50% likelihood they'll promptly resolve the problem, but a 50% chance they might encounter complications leading to extended downtime.

Scenario B: In this case, 5 technicians are highly experienced and can swiftly diagnose and rectify any issue after a machinery malfunction, ensuring minimal downtime. However, 1 of the technicians, though competent, is inexperienced. When this technician addresses the malfunction, there's a 50% likelihood they'll promptly resolve the problem, but a 50% chance they might encounter complications leading to extended downtime.

You are given an opportunity to sponsor training for a technician to elevate their skill set, ensuring they consistently and swiftly restore malfunctioning machinery. Would you pay as much in Scenario A reducing the number of inexperienced technicians from 4 to 3 as you would in Scenario B reducing the number of inexperienced technicians from 1 to 0?

A.1.6. Overconfidence

After prompting Linda problem (conjunction bias), the disease PPV calculation (base rate neglect), the bus stop pattern calculation (availability heuristic), CRT (System-1 and System-2 thinking), the four-card selection task (confirmation bias), as well as their corresponding reframed tests in OM settings, follow up with the question:

How confident are you about your previous answer (0%-100%)?

A.1.7. Ambiguity aversion

Standard Context

There is an urn with 30 red balls and 60 other balls that are either black or yellow (we do not know the exact numbers of black balls or yellow balls, but the total number of black and yellow balls is 60). Choose among the following two options:

Test A [Question]

A: $100 if you draw a red ball.

B: $100 if you draw a black ball.

Test B [Question]

C: $100 if you draw a red or yellow ball.

D: $100 if you draw a black or yellow ball.

Inventory & Procurement Context - Inventory Ambiguity

You are an exotic fruit importer preparing inventory for next season's sales. Your long-term business partner in Thailand offers three types of exotic fruits produced on her farms: mangosteen, durian, and rambutan.

According to market research, only one of the three fruits will become popular (and thus highly profitable) next season in North America, while the popularity of the other two fruits will stay the same. Furthermore, according to research, there is a 1/3 chance that mangosteen will become popular in the coming year. On the other hand, durian and rambutan have a combined probability of 2/3 being popular next year. However, the exact distribution of probabilities between durian and rambutan is uncertain.

Test A [Question]

Given the time, supply chain, and budget constraints, you can only afford to order two types of fruit with your business partner: mangosteen and durian. Which fruit should you commit to purchasing?

A: Mangosteen

B: Durian

Test B [Question]

Given the time, supply chain, and budget constraints, you can only afford to order two combinations of fruits with your business partner. Which fruit combinations should you commit to purchasing?

C: Mangosteen and rambutan

D: Durian and rambutan

Other Operational Context - Production Ambiguity

You are the operations manager of an electronics manufacturer. The government has announced its intent to implement a new law that will introduce a new industry standard affecting product design and manufacturing in the coming year. Based on your intelligence, the new standard will be among one of the following three potential standards:

- **Standard 1**: The new regulations will mandate enhanced energy efficiency, necessitating manufacturers to adopt advanced power management systems in their devices.

- **Standard 2**: The new regulations will mandate improved waste management, compelling manufacturers to use sustainable materials and embrace more efficient recycling techniques.

- **Standard 3**: The new regulations will mandate enhanced data security protocols, requiring electronics to incorporate advanced encryption and security features.

Based on your intelligence, there is a 1/3 chance that Standard 1will be introduced by the government. Also, the combined probability of Standard 2 and Standard 3 being introduced is 2/3, although the exact distribution of probabilities between Standard 2 and Standard 3 is uncertain.

Test A [Question]

Given the R&D constraints, as well as the looming deadline for the new law's implementation, you can only gear your production toward one of these potential standards. Which would you prioritize?

A: Design for enhanced energy efficiency.

B: Design for improved waste management.

Test B [Question]

Given the R&D constraints, as well as the looming deadline for the new law's implementation, you can only gear your production toward two of these potential standards. Which would you prioritize?

C: Design for enhanced energy efficiency and enhanced data security protocols.

D: Design for improved waste management and enhanced data security protocols.

ec6

Author: *A Manager and an AI Walk into a Bar*
Article submitted to *Manufacturing & Service Operations Management*; manuscript no. MSOM-2023-0279.R2

A.2. Biases in Evaluation of Outcomes

A.2.1. Risk Aversion and Scaling

Standard Context

Riskier prospect has same EV as safer prospect

Q: Choose between two lotteries A and B, which one is better? A: 50% chance to win $5.5 and 50% chance to win $4.5 B: 50% chance to win $9 and 50% chance to win $1

A: Lottery [INSERT]

Riskier prospect has higher EV as safer prospect

Q: Choose between two lotteries A and B, which one is better? A: 50% chance to win $5.5 and 50% chance to win $4.5 B: 50% chance to win $9.5 and 50% chance to win $1

A: Lottery [INSERT]

Inventory & Procurement Context – Newsvendor

Riskier prospect has same EV as safer prospect

Imagine you are a manager deciding how many Widgets to order for the upcoming sales period. Demand will either be high or low. Based on your analysis, you've narrowed down your choices to two order quantities, A and B. Option A is a lower order quantity than Option B, but Option A has a higher per-unit cost than Option B.

With 50% chance, demand will be high. With Option A you will run out of inventory, yielding a profit of $5500. With Option B you will satisfy all the demand, yielding a profit of $9000.

With 50% change, demand will be low. With Option A you will have little unsold inventory, yielding a profit of $4500. With Option B you will have a large number of unsold Widgets, yielding a profit of $1000.

Q: Choose between two options, which one is better?

A: Option [INSERT]

Riskier prospect has higher EV as safer prospect

Imagine you are a manager deciding how many Widgets to order for the upcoming sales period. Demand will either be high or low. Based on your analysis, you've narrowed down your choices to two order quantities, A and B. Option A is a lower order quantity than Option B, but Option A has a higher per-unit cost than Option B.

With 50% chance, demand will be high. With Option A you will run out of inventory, yielding a profit of $5500. With Option B you will satisfy all the demand, yielding a profit of $9500.

With 50% change, demand will be low. With Option A you will have little unsold inventory, yielding a profit of $4500. With Option B you will have a large number of unsold Widgets, yielding a profit of $1000.

Q: Choose between two options, which one is better?

A: Option [INSERT]

Other Operatinal Context - Nurse Staffing

Riskier prospect has same EV as safer prospect

Imagine you are an emergency room manager deciding on staffing nurses for the upcoming weekend. Emergency room arrivals will either be high or low. Based on your analysis, you've narrowed down your

choices to two options, A and B. Option A schedules for fewer nurses than Option B, but Option A may need to call agency nurses on short notice, which have a higher per-hour cost than Option B.

With 50% chance, emergency room arrivals will be high. With Option A you will have to call many agency nurses, yielding a cost of $9000. With Option B you will serve all patients with scheduled nurses, yielding a cost of $5500.

With 50% change, emergency room arrivals will be low. With Option A you will have not need to call any agency nurses, yielding a cost of $1000. With Option B you will have many scheduled nurses doing nothing, yielding a cost of $4500.

Q: Choose between two options, which one is better?

A: Plan [INSERT]

Riskier prospect has higher EV as safer prospect

Imagine you are an emergency room manager deciding on staffing nurses for the upcoming weekend. Emergency room arrivals will either be high or low. Based on your analysis, you've narrowed down your choices to two options, A and B. Option A schedules for fewer nurses than Option B, but Option A may need to call agency nurses on short notice, which have a higher per-hour cost than Option B.

With 50% chance, emergency room arrivals will be high. With Option A you will have to call many agency nurses, yielding a cost of $9500. With Option B you will serve all patients with scheduled nurses, yielding a cost of $5500.

With 50% change, emergency room arrivals will be low. With Option A you will have not need to call any agency nurses, yielding a cost of $1000. With Option B you will have many scheduled nurses doing nothing, yielding a cost of $4500.

Q: Choose between two options, which one is better?

A: Plan [INSERT]Prospect Theory

A.2.2. Prospect Theory

Standard context

Gains

Q: Would you rather Option A) play a gamble where there is an 80% chance of earning $4000 or Option B) receive $3000 with certainty?

A: Option [INSERT]

Losses

Q: Now consider the following choice: Would you rather Option A) play a gamble where there is an 80% chance of losing $4000 or Option B) pay $3000 with certainty?

A: Option [INSERT]

Inventory & Procurement - Supplier Selection

Gains

You are a procurement manager, and your manufacturing company is launching a new product that requires a specific component. You have to choose between two options to procure the component.

Option A: Choose a new supplier, Supplier A, who offers a competitive price, providing an 80% chance of increasing your profit by $4,000, but carries a 20% risk of supplying defective parts, which could result in 0 profit due to production delays and quality issues.

Option B: Stick with your current, reliable supplier, Supplier B, which guarantees a quality component but at a higher price, ensuring an additional profit of $3,000 with certainty due to stable production but no additional gain.

Q: Would you rather select the new Supplier A, which has an 80% chance of earning $4,000 profit or stick with the current Supplier B which earn $3,000 with certainty?

A: Supplier [INSERT]

Losses

Now consider the following choice: You are a procurement manager, and one of your suppliers for a key component has recently failed to deliver on time, resulting in production delays. You have to choose between two options to procure the component.

Option A: Stick with your current supplier, Supplier A, who will deliver the component with 80% chance in 21 days causing a loss of $4,000 and with a 20% chance in 7 days, causing no loss.

Option B: Choose a new supplier, Supplier B, who guarantees a delivery for the components of 14 days, which will cost a total of $3,000.

Q: Would you rather stick with the current Supplier A, which has an 80% chance of losing $4,000 or select the new Supplier B and pay $3,000 with certainty.

A: Supplier [INSERT]

Other Operational Context - Capacity and Infrastructure Investment
Gains

You are the operations manager at a manufacturing company facing decisions about infrastructure upgrades. You have two options and you must select one:

Option A: Opt for a comprehensive infrastructure upgrade to expand capacity where there's an 80% chance of earning an additional $4,000 in profit and a 20% chance that the market does not grow as anticipated resulting in 0 profit.

Option B: Stick to your current capacity and earn an additional $3,000 in profit with certainty, without incurring any investment.

Q: Do you select Option A) invest in the novel infrastructure upgrade to expand capacity where there is an 80% chance of earning $4000 in profit or Option B) not invest in additional capacity and receive $3000 with certainty?

A: Option [INSERT]

Losses

Now consider the following choice: You are the operations manager at a manufacturing company facing decisions about infrastructure upgrade. You have two options and you must select one:

Option A: Opt for a comprehensive infrastructure upgrade where there's an 80% chance that the investment will cost $4,000 and a 20% chance of breaking even.

Option B: Make necessary repairs and minor upgrades to your infrastructure at a cost of $3,000 with certainty, ensuring minimal downtime and production continuity.

Q: Would you rather Option A) invest in a novel infrastructure upgrade where there is an 80% chance of losing $4000 or Option B) make minor repairs to your infrastructure and pay $3000 with certainty?

A: Option [INSERT]

A.2.3.   Framing

Standard context

Gain Frame

Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that that exact scientific estimate of the consequences of the programs is as follows:

Program A: 200 people will be saved.

Program B: There is a 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.

Q: Which program should I chose?

A: Program [INSERT]

Loss Frame

Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that that exact scientific estimate of the consequences of the programs is as follows:

Program C: 400 people will die.

Program D: There is a 1/3 probability that nobody will die and 2/3 probability that 600 people will die.

Q: Which program should I chose?

A: Program [INSERT]

Inventory & Procurement Context

Gain Frame

Imagine that you are a retail inventory manager preparing for the demand of a certain product. The demand is expected to be 600 units. Two alternative courses of action to prepare for the demand have been proposed. Assume that the exact scientific estimate of the consequences of the actions is as follows:

- If Action A is taken, 200 units will be sold.

- If Action B is taken, there is a 1/3 probability that 600 units will be sold and a 2/3 probability that nothing will be sold.

Q: Which of the two actions would you favor?

A: Action [INSERT]

Loss Frame

Imagine that you are a retail inventory manager preparing for the demand of a certain product. The demand is expected to be 600 units. Two alternative courses of action to prepare for the demand have been proposed. Assume that the exact scientific estimate of the consequences of the actions is as follows:

- If Action C is taken, 400 units will be lost.

- If Action D is taken, there is a 1/3 probability that no sales will be lost and a 2/3 probability that 600 units of sale will be lost.

Q: Which of the two actions would you favor?

A: Action [INSERT]

Other Operations

Gain Frame

Imagine that you are an operations manager at a manufacturing plant, which usually produces 600 units per day. A labor strike is anticipated, posing a significant risk to daily production. Two alternative strategies to handle the anticipated strike have been proposed, with the following expected outcomes:

- If Strategy A is adopted through the use of temporary labor and overtime, production will be 200 units per day.

- If Strategy B is adopted, which involves utilizing a third-party manufacturer, there is a 1/3 probability that production will be 600 units per day and a 2/3 probability that there will be no production.

Q: Which of the two strategies would you favor?

A: Strategy [insert]

Loss Frame

Imagine that you are an operations manager at a manufacturing plant, which usually produces 600 units per day. A labor strike is anticipated, posing a significant risk to daily production. Two alternative strategies to handle the anticipated strike have been proposed, with the following expected outcomes:

- If Strategy C is adopted through the modification of production lines and utilization of non-striking staff, production will decrease by 400 units per day.

- If Strategy D is adopted, which involves emergency negotiations with the labor union, there is a 1/3 probability that there will be no decrease in production and a 2/3 probability that production will decrease by 600 units per day.

Q: Which of the two strategies would you favor?

A: Strategy [insert]

A.2.4. Anticipated Regret

Standard Context

Control

Imagine that you commute to school and that you park your car in one of the lots on campus. Further imagine that, on this day, you are walking to class in a bit of a rush because you have a quiz that you do not want to be late for. On the way to class, however, you get a strange feeling that you may have left your car door unlocked. Try as you might, you cannot be absolutely certain whether or not you locked your door.

Q: Would you go back to check your car or go straight to class for the quiz?

A: I [INSERT] go back to check the car.

Omission (Car) Regret

Imagine that you commute to school and that you park your car in one of the lots on campus. Further imagine that, on this day, you are walking to class in a bit of a rush because you have a quiz that you do not want to be late for. On the way to class, however, you get a strange feeling that you may have left your car door unlocked. Try as you might, you cannot be absolutely certain whether or not you locked your door. Think for a minute about how upset you would be if you didn't go back to check your car, and later that day your car was burglarized.

Q: Would you go back to check your car or go straight to class for the quiz?

A: I [INSERT] go back to check the car.

Commission (Quiz) Regret

Imagine that you commute to school and that you park your car in one of the lots on campus. Further imagine that, on this day, you are walking to class in a bit of a rush because you have a quiz that you do not want to be late for. On the way to class, however, you get a strange feeling that you may have left your car door unlocked. Try as you might, you cannot be absolutely certain whether or not you locked your door. Think for a minute about how upset you would be if you went back to check your car, found out that it was locked all along, and ended up being late for your quiz.

Q: Would you go back to check your car or go straight to class for the quiz?

A: I [INSERT] go back to check the car.

Inventory & Procurement Context

Control

Imagine you are the inventory manager for a retail company. A month ago, you placed an order for a seasonal product with a two-month lead time, anticipating a certain level of demand. As the season approaches, market indicators and early customer interest suggest that the demand might be significantly higher than you initially anticipated. You have the option to place a rush order, which has a one-month lead time, to procure additional units to meet the potentially higher demand. However, rush orders come with higher costs and there's still some uncertainty about the exact demand levels.

Q: Do you place the rush order to potentially meet the higher demand or stick with your initial order and risk potential stockouts?

A: I [INSERT] place the rush order.

Omission (Lost Sales) Regret

Imagine you are the inventory manager for a retail company. A month ago, you placed an order for a seasonal product with a two-month lead time, anticipating a certain level of demand. As the season approaches, market indicators and early customer interest suggest that the demand might be significantly higher than you initially anticipated. You have the option to place a rush order, which has a one-month lead time, to procure additional units to meet the potentially higher demand. However, rush orders come with higher costs and there's still some uncertainty about the exact demand levels. Think about how you would feel facing the scrutiny of your manager if you didn't place a rush order and later found out that there weren't enough products, leading to lost sales and disappointed customers.

Q: Do you place the rush order to potentially meet the higher demand or stick with your initial order and risk potential stockouts?

A: I [INSERT] place the rush order.

Commission (Excess Inventory) Regret

Imagine you are the inventory manager for a retail company. A month ago, you placed an order for a seasonal product with a two-month lead time, anticipating a certain level of demand. As the season approaches, market indicators and early customer interest suggest that the demand might be significantly higher than you initially anticipated. You have the option to place a rush order, which has a one-month lead time, to procure additional units to meet the potentially higher demand. However, rush orders come with higher costs and there's still some uncertainty about the exact demand levels. Think about how you would feel facing the scrutiny of your manager if you placed a rush order and later found out that that the demand wasn't as high as the indicators suggested, leading to excess inventory and higher costs.

Q: Do you place the rush order to potentially meet the higher demand or stick with your initial order and risk potential stockouts?

A: I [INSERT] place the rush order.

Other Operational Context

Control

Imagine you are the operations manager at a manufacturing facility. The company is considering the adoption of Augmented Reality (AR) headsets to assist employees in the manufacturing processes. These AR headsets can provide real-time data, assist in assembly processes, and offer training simulations. However, there are rumors in the industry that a significant technological upgrade for these AR headsets is on the horizon, which could offer even more advanced features and improved efficiency. If you decide to purchase the current version of the AR headsets now, it would be a significant investment, and the company might not be in a position to upgrade immediately when the new version comes out.

Q: Do you invest in the current AR headset technology to gain immediate benefits or wait for the potential upgrade, risking a delay in the advantages the technology can offer?

A: I [INSERT] invest in the current AR headset technology.

Omission Regret

Imagine you are the operations manager at a manufacturing facility. The company is considering the adoption of Augmented Reality (AR) headsets to assist employees in the manufacturing processes. These AR headsets can provide real-time data, assist in assembly processes, and offer training simulations. However, there are rumors in the industry that a significant technological upgrade for these AR headsets is on the horizon, which could offer even more advanced features and improved efficiency. If you decide to purchase the current version of the AR headsets now, it would be a significant investment, and the company might not be in a position to upgrade immediately when the new version comes out. Think about how you would feel if you decided to wait for the upgrade, only to find out that the upgrade took longer than expected and the new features were not as groundbreaking as anticipated. Meanwhile, competitors who adopted the current technology have gained a significant advantage in efficiency and production quality.

Q: Do you invest in the current AR headset technology to gain immediate benefits or wait for the potential upgrade, risking a delay in the advantages the technology can offer?

A: I [INSERT] invest in the current AR headset technology.

Commission Regret

Imagine you are the operations manager at a manufacturing facility. The company is considering the adoption of Augmented Reality (AR) headsets to assist employees in the manufacturing processes. These AR headsets can provide real-time data, assist in assembly processes, and offer training simulations. However, there are rumors in the industry that a significant technological upgrade for these AR headsets is on the horizon, which could offer even more advanced features and improved efficiency. If you decide to purchase the current version of the AR headsets now, it would be a significant investment, and the company might not be in a position to upgrade immediately when the new version comes out. Think about how you would feel if you decided to not wait for the upgrade, only to find out that the upgrade occurred even sooner than expected and the new features were even more groundbreaking than anticipated. Meanwhile, competitors who waited for the new technology are poised to gain a significant advantage in efficiency and production quality.

Q: Do you invest in the current AR headset technology to gain immediate benefits or wait for the potential upgrade, risking a delay in the advantages the technology can offer?

A: I [INSERT] invest in the current AR headset technology.

A.2.5. Mental Accounting and Reference Dependency

Standard Context

Absolute Frame

Mr. A's couch was priced originally at $1,300 but is now reduced to $1,250. Mr. B's chair was priced originally at $300 and his couch was priced at $1,000. His chair is now reduced to $200 and his couch is now increased to $1,050.

Q: Who was happier? Mr. A, Mr. B, or no difference?

A: Mr. [INSERT]

Dual Frame

Mr. A's couch was priced originally at $1,300 but is now reduced by 3.846% to $1,250. Mr. B's chair was priced originally at $300 and his couch was priced at $1,000. His chair is now reduced by 33.333% to $200, and his couch is now increased by 5% to $1,050.

Q: Who was happier? Mr. A, Mr. B, or no difference?

A: Mr. [INSERT]

Relative Frame

Mr. A's couch was priced originally at $1,300 but is now reduced by 3.846%. Mr. B's chair was priced originally at $300 and his couch was priced at $1,000. His chair is now reduced by 33.333%, and his couch is now increased by 5%.

Q: Who was happier? Mr. A, Mr. B, or no difference?

A: Mr. [INSERT]

Inventory & Procurement Context

Absolute Frame

Mr. A and Mr. B sell Widgets, which are perishable goods. Mr. A sells Widget X, whereas Mr. B sells Widget Y and Widget Z. Mr. A bought 10 Widget Xs at a per-unit cost of $125 for a total cost of $1250 and sold 10 Widget Xs at a price of $130 per unit for a total revenue of $1,300. Mr. B bought 10 Widget Ys at a per-unit cost of $20 for a total cost of $200 and 10 Widget Zs at a per-unit cost of $130 for a total cost of $1300. Mr. B sold 10 Widget Ys at a price of $30 for a total revenue of $300 and sold 10 Widget Zs at an average price of $125 for a total revenue of $1,250.

Q: Who was happier? Mr. A, Mr. B, or no difference?

A: Mr. [INSERT]

Dual Frame

Mr. A and Mr. B sell Widgets, which are perishable goods. Mr. A sells Widget X, whereas Mr. B sells Widget Y and Widget Z. Mr. A bought 10 Widget Xs at a per-unit cost of $125 for a total cost of $1250 and sold 10 Widget Xs at a price of $130 for a total revenue of $1,300, a profit margin of 3.846%. Mr. B bought 10 Widget Ys at a per-unit cost of $20 for a total cost of $200 and 10 Widget Zs at a per-unit cost of $130 for a total cost of $1300. Mr. B sold 10 Widget Ys at a price of $30 for a total revenue of $300, a profit margin of 33.333%, and sold 10 Widget Zs at an average price of $125 for a total revenue of $1,250, a profit margin of -5%.

Q: Who was happier? Mr. A, Mr. B, or no difference?

A: Mr. [INSERT]

Relative Frame

Mr. A and Mr. B sell Widgets, which are perishable goods. Mr. A sells Widget X, whereas Mr. B sells Widget Y and Widget Z. Mr. A bought 10 Widget Xs at a per-unit cost of $125 for a total cost of $1250 and sold 10 Widget Xs at a price of $130, making a profit margin of 3.846%. Mr. B bought 10 Widget Ys at a per-unit cost of $20 for a total cost of $200 and 10 Widget Zs at a per-unit cost of $130 for a total cost of $1300. Mr. B sold 10 Widget Ys at a price of $30, making a profit margin of 33.333%, and sold 10 Widget Zs at an average price of $125, making a profit margin of -5%.

Q: Who was happier? Mr. A, Mr. B, or no difference?

A: Mr. [INSERT]

Other Operational Context

Absolute Frame

Mr. A owns Factory X, which had a monthly energy cost of $1,300. After renovating the factory, his monthly energy cost reduced to $1,250. Mr. B owns Factory Y and Factory Z. Factory Y had a monthly energy cost of $300, and after renovations, the monthly cost was reduced to $200. Factory Z originally had a monthly energy cost of $1,000, but after renovations, the cost increased to $1,050.

Q: Who was happier? Mr. A, Mr. B, or no difference?

A: Mr. [INSERT]

Dual Frame

Mr. A owns Factory X, which had a monthly energy cost of $1,300. After renovating the factory, his monthly energy cost reduced by 3.846% to $1,250. Mr. B owns Factory Y and Factory Z. Factory Y had a

monthly energy cost of $300, and after renovations, the monthly cost reduced by 33.333% to $200. Factory Z originally had a monthly energy cost of $1,000, but after renovations, the cost increased by 5% to $1,050.

Q: Who was happier? Mr. A, Mr. B, or no difference?

A: Mr. [INSERT]

Relative Frame

Mr. A owns Factory X, which had a monthly energy cost of $1,300. After renovating the factory, his monthly energy cost reduced by 3.846%. Mr. B owns Factory Y and Factory Z. Factory Y had a monthly energy cost of $300, and after renovations, the monthly cost reduced by 33.333%. Factory Z originally had a monthly energy cost of $1,000, but after renovations, the cost increased by 5%.

Q: Who was happier? Mr. A, Mr. B, or no difference?

A: Mr. [INSERT]

A.2.6. Intertemporal Choice and Hyperbolic Discounting

Standard Context

3 Months

You have won $15 in the lottery which is being held by your bank. You can take the $15 now or wait until later.

Q: How much money would you require to make waiting 3 months just as attractive as getting the $15 now?

A: $[INSERT]

1 Year

You have won $15 in the lottery which is being held by your bank. You can take the $15 now or wait until later.

Q: How much money would you require to make waiting 1 year just as attractive as getting the $15 now?

A: $[INSERT]

3 Years

You have won $15 in the lottery which is being held by your bank. You can take the $15 now or wait until later.

Q: How much money would you require to make waiting 3 years just as attractive as getting the $15 now?

A: $[INSERT]

Inventory & Procurement Context - Supply Chain Revenue Sharing

3 Months

As a supplier, you have a revenue-sharing agreement with one of your main retailers. Based on the past quarter's sales, the retailer owes you $15,000. However, the retailer approaches you with a proposition: They'd like to retain the funds for a bit longer to manage their cash flow and are willing to pay you more if you agree to wait.

Q: How much money would make waiting 3 months just as attractive as receiving the $15,000 now?

A: $[INSERT]

As a supplier, you have a revenue-sharing agreement with one of your main retailers. Based on the past quarter's sales, the retailer owes you $15,000. However, the retailer approaches you with a proposition: they'd like to retain the funds for a bit longer to manage their cash flow and are willing to pay you more if you agree to wait. How much additional money would make waiting 3 months just as attractive as receiving the $15,000 now?

1 Year

As a supplier, you have a revenue-sharing agreement with one of your main retailers. Based on the past quarter's sales, the retailer owes you $15,000. However, the retailer approaches you with a proposition: They'd like to retain the funds for a bit longer to manage their cash flow and are willing to pay you more if you agree to wait.

Q: How much money would make waiting 1 year just as attractive as receiving the $15,000 now?

A: $[INSERT]

3 Years

As a supplier, you have a revenue-sharing agreement with one of your main retailers. Based on the past quarter's sales, the retailer owes you $15,000. However, the retailer approaches you with a proposition: They'd like to retain the funds for a bit longer to manage their cash flow and are willing to pay you more if you agree to wait.

Q: How much money would make waiting 3 years just as attractive as receiving the $15,000 now?

A: $[INSERT]

Other Operational Context - Investment Opportunity

3 Months

Your company has developed a groundbreaking manufacturing technique that has garnered interest from various investors. A prominent investment firm offers an upfront payment of $15,000 to help finance the expansion of your facility to accommodate this new technique. However, they're in the midst of closing a major fund and suggest that if you can delay the acceptance of their investment, they'll add a premium due to the wait.

Q: How much payment would make waiting 3 months just as attractive as receiving the $15,000 now?

A: $[INSERT]

1 Year

Your company has developed a groundbreaking manufacturing technique that has garnered interest from various investors. A prominent investment firm offers an upfront payment of $15,000 to help finance the expansion of your facility to accommodate this new technique. However, they're in the midst of closing a major fund and suggest that if you can delay the acceptance of their investment, they'll add a premium due to the wait.

Q: How much payment would make waiting 1 year just as attractive as receiving the $15,000 now?

A: $[INSERT]

3 Years

Your company has developed a groundbreaking manufacturing technique that has garnered interest from various investors. A prominent investment firm offers an upfront payment of $15,000 to help finance the

expansion of your facility to accommodate this new technique. However, they're in the midst of closing a major fund and suggest that if you can delay the acceptance of their investment, they'll add a premium due to the wait.

Q: How much payment would make waiting 3 years just as attractive as receiving the $15,000 now?

A: $[INSERT]

A.2.7. Endowment Effect

Standard Context

WTP Test

I can enter a prize draw for $2. The prize is worth $70 and my estimated winning probability is 2.08%.

Q: Should I enter the prize draw?

A: You [INSERT] buy a ticket for the prize draw.

WTA Test

I was given a prize draw ticket for free. The prize is worth $70 and my estimated winning probability is 2.08%. My friend is offering $2 for my ticket.

Q: Should I sell it?

A: You [INSERT] sell the ticket for the prize draw.

Inventory & Procurement Context - Purchasing and Distribution

WTP Test

Your company produces independent films. There's an interesting script circulating in the industry. However, its content touches on sensitive cultural topics, and it's still awaiting a green light from the film censor board. If approved, this film is estimated to generate a net profit of $70,000. Insiders estimate that there is a 2.08% chance of getting past the censor board without significant content changes. The scriptwriter, recognizing its potential yet aware of its controversial nature, is offering you the exclusive rights to produce this script for $2,000.

Q: Should you purchase the rights to the script?

A: I [INSERT] purchase the rights to the script.

WTA Test

Your company produces independent films. There's an interesting script you've recently acquired. However, its content touches on sensitive cultural topics, and it's still awaiting a green light from the film censor board. If approved, this film is estimated to generate a net profit of $70,000. Insiders estimate that there is a 2.08% chance of getting past the censor board without significant content changes. Another film studio, having caught wind of its potential, is offering to buy the exclusive rights to the script from you for $2,000.

Q: Should you sell the rights to the script?

A: I [INSERT] sell the rights to the script.

Operations Context

WTP Test

Your are considering a last minute application for a rigorous government-backed environmental program. The program has stringent requirements, but guarantees tax breaks and grants that will translate to a profit

of $70,000. To submit on time requires purchasing an application from another company that has already prepared an application. One company has offered to sell you their application, which as a 2.08% chance of being accepted, for $2,000.

Q: Do you buy the application?

A: I [INSERT] buy the application.

WTA Test

Your have prepared an application for a rigorous government-backed environmental program. The program has stringent requirements, but guarantees tax breaks and grants that will translate to a profit of $70,000. Another company wanting to apply at the last minute has offered to buy your application, which as a 2.08% chance of being accepted, for $2,000.

Q: Should you sell your application?

A: I [INSERT] sell my application.

A.2.8.   Sunk Cost

Standard Context

Test 1

As the president of an airline company, you have invested 10 million dollars of the company's money into a research project. The purpose was to build a plane that would not be detected by conventional radar, in other words, a radar-blank plane. When the project is 90% completed, another firm begins marketing a plane that cannot be detected by radar. Also, it is apparent that their plane is much faster and far more economical than the plane your company is building.

Q: Should you invest the last 10% of the research funds to finish your radar-blank plane?

A: I would [INSERT] the last 10% of the research funds to finish your radar-blank plane.

Test 2

As the president of an airline company, you have received a suggestion from one of your employees. The suggestion is to use the last 1 million dollars of your research funds to develop a plane that would not be detected by conventional radar, in other words, a radar-blank plane. However, another firm has just begun marketing a plane that cannot be detected by radar. Also, it is apparent that their plane is much faster and far more economical than the plane your company could build.

Q: Should you invest the last million dollars of your research funds to build the radar-blank plane proposed by your employee?

A: I would [INSERT] the last million dollars of the research funds to build the radar-blank plane.

Inventory & Procurement Context

Test 1

You are the head of operations for a large retail company. You've ordered the creation of a new dress based on recent trends from high-end runway shows using resources from the new product-line budget. Production of the dresses was 90% finished. However, your direct competitor released and stocked retail stores with a virtually identical dress priced just above your company's per unit cost.

Q: Do you spend the last $100,000 of your new product-line budget to finish producing and shipping the dresses to your retail stores

A: I would [INSERT] spend the last $100,000 of the new product-line budget to finish producing and shipping the dresses.

Test 2

You are the head of operations for a large retail company. You have received a suggestion from one of your employees to spend the last $100,000 of your new product-line budget to create a new dress based on recent trends from the runway. However, your direct competitor released and stocked retail stores with a virtually identical dress priced just above your company's per unit cost.

Q: Do you spend the last $100,000 of your new product-line budget to finish producing and shipping the dresses to your retail stores?

A: I would [INSERT] the last $100,000 of the new product-line budget to finish producing and shipping the dresses.

Other Operational Context

Test 1

You are the head of logistics for a large company. You are building a warehouse near a major city, anticipating a substantial increase in demand in that region entirely due to a new planned transportation route. However, recent news reports that the major transportation route planned near the warehouse will be canceled.

Q: Do you invest the last $2 million of your expansion budget to complete the warehouse?

A: I would [INSERT] complete the warehouse.

Test 2

You are the head of logistics for a large company. One of your managers suggests using the last $2 million of your expansion budget to complete a warehouse near a major city, anticipating a substantial increase in demand in that region entirely due to a new planned transportation route. However, recent news reports that the major transportation route planned near the warehouse will be canceled.

Q: Do you invest the last $2 million of your expansion budget to complete the warehouse?

A: I would [INSERT] the warehouse.

A.3. Heuristics in Decision-Making

A.3.1. Cognitive Reflection Test (CRT)

Standard Context

Q1: A bat and a ball cost $1.1 in total. The bat costs $1 more than the ball. How much does the ball cost?

Q2: If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets?

Q3: In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

Inventory & Procurement Context - Assorted

ec20
Author: *A Manager and an AI Walk into a Bar*
Article submitted to *Manufacturing & Service Operations Management*; manuscript no. MSOM-2023-0279.R2

Q1: Bulk Purchasing :A supplier offers a discount for a bulk purchase of two types of raw materials: aluminum sheets and lubricant oil. Together, they cost $1.1k. The aluminum sheets cost $1k more than the lubricant oil. How much does the lubricant oil cost?

Q2: Supply Chain Delivery: If ten delivery truck can transport inventory to ten retail outlets in ten hours, how long would it take two hundred delivery trucks to transport supplies to two hundred retail outlets?

Q3: Inventory Expansion: A company notices that its inventory of a particular product doubles every week due to increased demand. If the company reach full capacity in 48 week, how long did it take for the inventory to reach a quarter of its capacity?

Other Operational Context - Assorted

Q1: Defects: During a quality control inspection, two batches of products were examined: one of electronics and another of accessories. In total, 110 defects were identified. The electronics batch had 100 defects more than the accessories batch. How many defects were found in the accessories batch?

Q2: Inspection: If 5 quality inspectors can inspect 5 production batches in 5 hours, how long would it take 100 quality inspectors to inspect 100 production batches?

Q3: Process Efficiency: A production machine, after undergoing a major calibration, doubles its throughput capacity each day. If it takes 12 days for the machine to reach 100% throughput after the calibration, how many days would it have taken to reach 50% of that throughput?

A.3.2. Confirmation Bias

Standard Context

You are shown four cards, marked E, K, 4 and 7. Each card has a letter on one side and a number on the other. You are given the following rule: Every card with a vowel on one side has an even number on the other side. Which cards must you turn over to test whether the rule is true or false?

Inventory & Procurement Context - Inventory Inspection

You are the procurement manager at a large manufacturing company. You receive inventory from four different vendors: A (certified suppliers), NA (non-certified suppliers), Q (quality-tested samples), and NQ (non-quality-tested samples). Each vendor claims their products are either certified or non-certified and have either passed a quality test or haven't undergone any testing. You sell your inventory at different prices depending on certification and quality. To help cluster products, you use the following rule: "Every sample from a certified supplier should have undergone quality testing." Which vendor's products must you inspect to ensure the rule is being adhered to?

Other Operational Context - Factory Inspection

You are the operations manager at a warehouse. You have four storage areas: P (perishable Items), NP (non-perishable Items), T (temperature controlled), and NT (non-temperature controlled). Each storage area contains items that are either perishable or non-perishable and are either are temperature controlled or non-temperature controlled. You are given the following rule: "Every storage area that contains perishable items must be temperature-controlled." Which storage areas must you inspect to ensure the rule is being followed?

A.4. Sample Code for Data Collection

```python
1  import os
2  import openai
3  import re
4  import random
5  import time
6  import pandas as pd
7  from datetime import datetime
8
9  #API key
10 openai.api_key = "INSERT KEY HERE"
11
12 #Variables
13 GPT_MODEL = "gpt-3.5-turbo-0613"  # "gpt-4-0613" #
14 TEMPERATURE = 1
15 NUM_OBSERVATIONS = 30
16 PAUSE = 10 # Seconds to slow down each run - useful when server is busy or when
       you are exceeding max token per hour set by OpenAI
17
18 #Functions
19 def askGPT(prompt):
20     completion = openai.ChatCompletion.create(
21         model=GPT_MODEL,
22         temperature=TEMPERATURE,
23         top_p=1,
24         frequency_penalty=0,
25         presence_penalty=0,
26         messages=[{"role": "user", "content": prompt}]
27     )
28     time.sleep(PAUSE)
29     return completion.choices[0].message.content
30
31 def generate_observations(prompt, Scenario, SubScenario, Context):
32     data = []  # to collect rows of data
33     current_date = datetime.now().strftime('%Y-%m-%d')
34
35     for x in range(NUM_OBSERVATIONS):
36         print(f"Observation # {x+1}")
37         gpt_response = askGPT(prompt)
38         print(gpt_response)
39         print("\n")
```

```python
40
41            # Collect data for this observation
42            data.append({
43                'Date': current_date,
44                'Model': GPT_MODEL,
45                'Temperature': TEMPERATURE,
46                'Scenario': Scenario,
47                'SubScenario': SubScenario,
48                'Context': Context,
49                'GPTResponds': gpt_response
50            })
51
52        # Create DataFrame
53        df = pd.DataFrame(data)
54
55        # Create the filename
56        current_time = datetime.now().strftime('%Y%m%d-%H%M%S')
57        filename = f"Evaluation-{Scenario}-{SubScenario}-{Context}-{GPT_MODEL}-{
           current_time}.csv"
58        # Clean filename to remove any non-alphanumeric characters (excluding hyphens
           and dots)
59
60        # Save df to the CSV file
61        df.to_csv(filename, index=False)
62
63        return df
64
65        folder_path = ''
66        full_path = folder_path + file_name
67
68        # Export the results_table to a CSV file in the specified Google Drive folder
69        df.to_csv(full_path, index=False)
70
71        print(f"Results exported to {file_name}")
72
73  #Data collection
74  Scenario = "Risk Aversion"
75  SubScenario = "Test1"
76  OM = "Standard"
77  prompt = """
78  Q: Choose between two lotteries A and B, which one is better? A: 50% chance to win
           $5.5 and 50% chance to win $4.5 B: 50% chance to win $9 and 50% chance to win
           $1
```

Author: *A Manager and an AI Walk into a Bar*
Article submitted to *Manufacturing & Service Operations Management*; manuscript no. MSOM-2023-0279.R2

ec23

```
79  A: Lottery [INSERT]

80  """

81  generate_observations(prompt,Scenario,SubScenario,OM)
```

Appendix B: Data Analysis Results

In Appendix B, we discuss results from the tests on all 18 biases in detail. For each bias, we present results from the original framing of test vignettes in reference studies (referred to as "Standard"), plus our two re-framed tests in operations management settings. The first framing is in the context of inventory management (referred to as "Inventory" framing), and the second framing is a collection of broader operations management settings such as sales, productions, and supply chains (referred to as "Operations").

We run our tests on both GPT-3.5 and GPT-4, and with three framings (Standard, Inventory, and Operations). In total, we perform 6 sets of tests under each bias (a set of tests may further contain multiple conditions in accordance to the design of the original studies). In each experimental condition, we perform 30 independent API calls with GPT-3.5 and GPT-4 on their June.11, 2023 versions of APIs, i.e. the "Time 2" experiments described in Section 3.1.

In this Appendix, we provide a detailed background of each behavioral bias, the results of our experiments, comparisons to human results, and consistency across contexts and models (or lack thereof). These results map back to the overall result summary Tables 1-3 in the main body of the paper.

B.1.   Biases in Judgments Regarding Risk

B.1.1.   The Hot-Hand and Gambler's Fallacies   The hot-hand and gambler's fallacies are both about false beliefs of future event probabilities being correlated with the past even when actual probabilities are entirely independent. In hot-hand fallacy cases, people believe a player's winning streak (by chance) indicates a higher probability of winning in the future. In contrast, in the gambler's fallacy, people believe a winning lottery number is less likely to win immediately again, even though these event probabilities are shown to be independent of the past. Wagenaar (1972) summarized many lab experiments on random number generation tasks in a review paper. As it turns out, it is difficult for humans to generate truly random sequences of numbers ((Camerer 2020). The gambler's fallacy is related to people generating "random" sequences that are negatively auto-correlated. For example, we tend to underestimate how many consecutive heads or tails there can be in a series of genuinely random fair coin tosses and feel compelled to switch after seeing a few consecutive heads or tails. Similarly, the hot-hand fallacy is analogous to a positive auto-correlation in a generated sequence. Most experiments in Wagenaar (1972) found human-generated random sequences to be negatively auto-correlated.

We adopt a similar experimental condition to Ross and Levy (1958) and Bakan (1960) and ask GPT to generate random fair coin toss series with a length of 50. GPT generates sequences that are *about* 50 in length. We show the lag-1 auto-correlations and their 95% confidence intervals in Figure B1 across 6 conditions.

We would like to provide the overall levels of autocorrelations across repetitions, however, correlations and their confidence intervals cannot be directly averaged without bias, nor can we assume that 30 series of 50 coin tosses are the same as one series of 1500 coin tosses. Thus, we perform meta-analyses to combine results within each of the six experimental conditions (treating each of the 30 repetitions as studies) using Fisher's Z transformation of correlations. The pooled correlation coefficients and hypothesis tests against the null (autocorrelation = 0) are shown in Table 5.

Standard
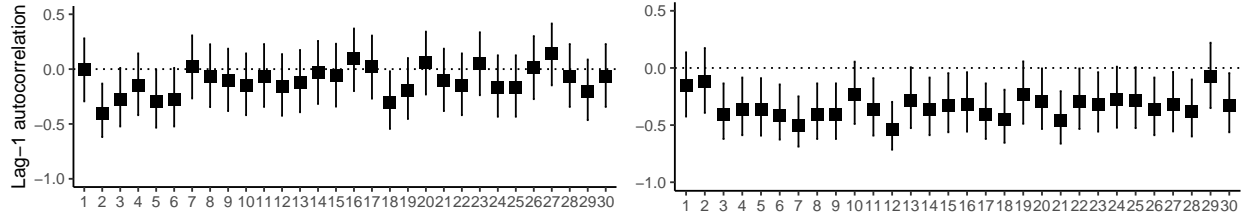
Inventory
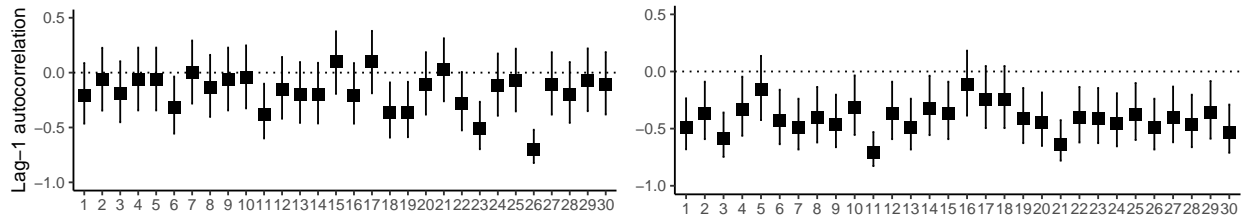
Operations
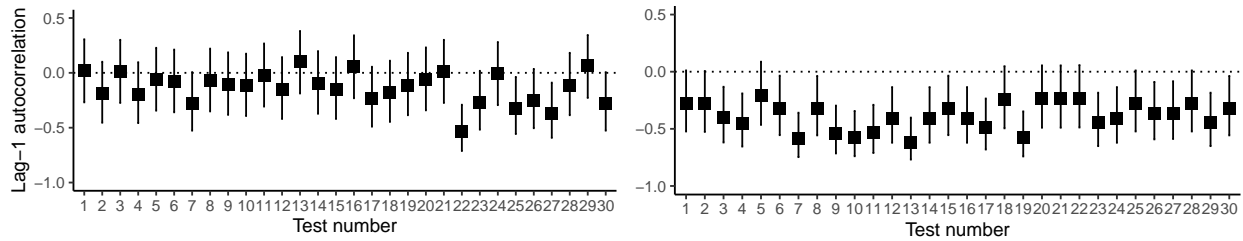
GPT–3.5   GPT–4

Figure B1   Hot hand - Lag-1 autocorrelations of GPT generated random coin tosses by models and contexts

Table 5   Hot hand - pooled Lag-1 autocorrelations of GPT generated coin tosses by models and contexts

Framing/Model	Pooled Correlation	95% CI lower	95% CI upper	z	p-value
Standard: GPT-3.5	-0.1079	-0.1609	-0.0543	-3.94	<0.0001
Standard: GPT-4	-0.3388	-0.3857	-0.2902	-12.82	<0.0001
Inventory: GPT-3.5	-0.1769	-0.2286	-0.1242	-6.50	<0.0001
Inventory: GPT-4	-0.4179	-0.4614	-0.3724	-16.17	<0.0001
Operations: GPT-3.5	-0.1366	-0.1891	-0.0833	-4.99	<0.0001
Operations: GPT-4	-0.3927	-0.4374	-0.3461	-15.08	<0.0001

From Table 5, we find that there is a consistent and significant negative autocorrelation in GPT's coin toss simulations across contexts and models at the .01 level. The negative autocorrelation corresponds to the gambler's fallacy type of bias, consistent with Wagenaar (1972)'s general findings that humans were poor randomizers and that they generated too many alternations in their sequences. Interestingly, we also find evidence of more severe negative autocorrelations in GPT-4 compared to GPT-3.5 across all framings of

instructions (notice in Table 5 the 95% confidence intervals of GPT-3.5 and GPT-4 do not overlap). So in this particular case, the more advanced model of the two seemed to produce more biased results.

B.1.2.   The conjunction fallacy and representativeness The probability of a conjunction event P(A&B) cannot be greater than P(A) or P(B). The conjunction fallacy is the observation that people's estimation of probability over conjunction events can violate this principle. Tversky and Kahneman (1983) demonstrated this fallacy in the "Linda problem":

> *Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*
> *A: Linda is a teacher in elementary school.*
> *B: Linda works in a bookstore and takes Yoga class.*
> *C: Linda is active in the feminist movement.*
> *D: Linda is a psychiatric social worker.*
> *E: Linda is a member of the League of Women Voters.*
> *F: Linda is a bank teller.*
> *G: Linda is an insurance salesperson.*
> *H: Linda is a bank teller and is active in the feminist movement.*

Since H is the conjunction of C and F, P(H) cannot be ranked higher than either P(C) or P(F), but Tversky and Kahneman (1983) observed that 85% of the subjects ranked P(C) > P(H) > P(F). The most common explanation for the conjunction fallacy is that individuals use representativeness as a heuristic. People tend to feel that Linda being a bank teller active in the feminist movement is more representative of her description than a bank teller (Davis 2018).

We gave GPT the same description of Linda and choices A through H, with the added instruction, "Rank the following statements that could describe Linda from the most likely to the least likely.". We also provided two additional reframed vignettes in OM contexts. The details of the vignettes in OM contexts can be found in Appendix A. We show the results across models and contexts in Figure B2. Each of the pie charts in Figure B2 shows the proportions of responses from GPT that 1.) exhibit conjunction fallacy (labelled as "Violation"), 2.) do not exhibit the fallacy (labelled as "No Violation"), and 3.) do not include all listed options as rankings (labelled as "Improper Ranking").

Figure B2 shows that we observe high levels of conjunction fallacy in GPT across almost all contexts and models. If a subject randomly ranks the 8 statements, there would be a 1 in 3 chance for rankings to contain a conjunction fallacy. We thus compare the proportions of GPT's conjunction fallacy responses to 1/3 using the Binomial exact test, and find GPT exhibiting significant conjunction fallacy in all but one context and model (p-values <0.0001). In the Inventory context, GPT-3.5 sometimes generates conjunction fallacy responses, but the proportion is not significantly greater than randomness (p=0.415). Additionally, when examining the influence of contexts within models, we found significant differences in the proportions of conjunction fallacy across the three framings of the Linda problem in both GPT-3.5 ($\chi^2 = 34.25$, p $= 5 \times 10^{-4}$) and GPT-4 ($\chi^2 = 20$, p $= 5 \times 10^{-4}$), suggesting that context matters for conjunction fallacy in GPT. However, when we compare the proportions of conjunction fallacy between the two models, we find no significant differences in both Standard (p $= 1$) and Operations ($\chi^2 = 3.16$, p $= 0.2349$) contexts, and a borderline significant difference between models in the Inventory context ($\chi^2 = 7.06$, p $= 0.0275$).
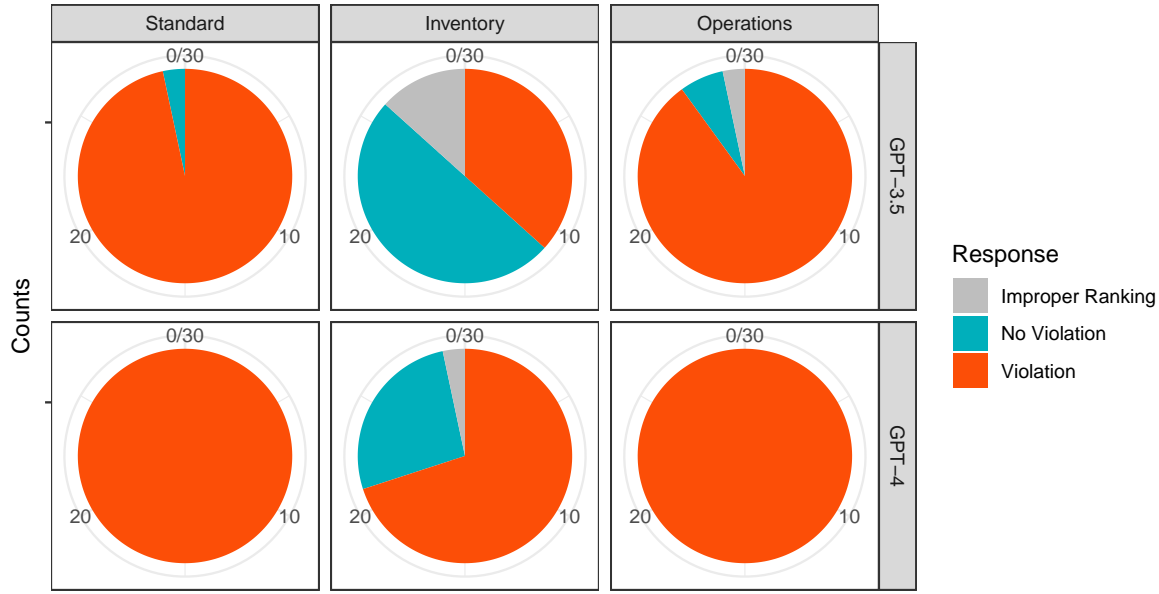
Figure B2   **Conjunction fallacy - proportions of conjunction fallacy in GPT responses by models and contexts**

B.1.3.   The availability heuristic   The availability heuristic is a tendency to use data familiar to oneself, such as personal experiences and conversations, to judge risks. Tversky and Kahneman (1973) demonstrated availability affecting judgment by giving subjects a visual task of estimating the possible ways a bus can make r stops traveling along a route with ten stops in total. The correct answer is the binomial coefficient of 10 choose r, which yields the highest estimate at five stops and is symmetric with respect to r=5. However, the subjects' median judgments are decreasing in r. Without realizing the symmetry between patterns with r stops and patterns with 10-r stops, patterns requiring more stops are more challenging to consider.

We perform a similar set of tasks using GPT-3.5 and GPT-4, with $r \in \{2, 5, 8\}$. We keep the description of the task as faithful as possible to the original description in Tversky and Kahneman (1973). Additionally, we provide two OM contexts of the Standard problems (see Appendix A for details).

Unsurprisingly, GPT performed well enough in all test conditions so that the median judgment is always correct as opposed to what was observed in humans by Tversky and Kahneman (1973), suggesting GPT is largely unbiased across models and contexts. To provide more insight into GPT's occasional errors, we graph the percentage of correct answers given by GPT in Figure B3.[17]

Figure B3 shows that GPT performs very well in the straight-forward combinatorics calculations. In GPT-3.5, error rate is higher when r=5 across all three contexts, since the calculation is more complex compared to r=2 and r=8. However, we observe a symmetry in r=2 and r=8 cases, which is different from human results. In GPT-4, the errors are further eliminated due to its improved mathematical capability.

We more formally test the correlation between the number of bus stops and the error rate, as well as the effects of models and contexts in logistic regressions in Table 6 (Model 1). We find that compared to r=2, r=5 results in borderline significantly higher odds of error (coefficient 0.81 or an odds ratio (OR) of ~2.25, p=0.0408) while r=8 is not significantly different from r=2, suggesting a symmetric pattern around
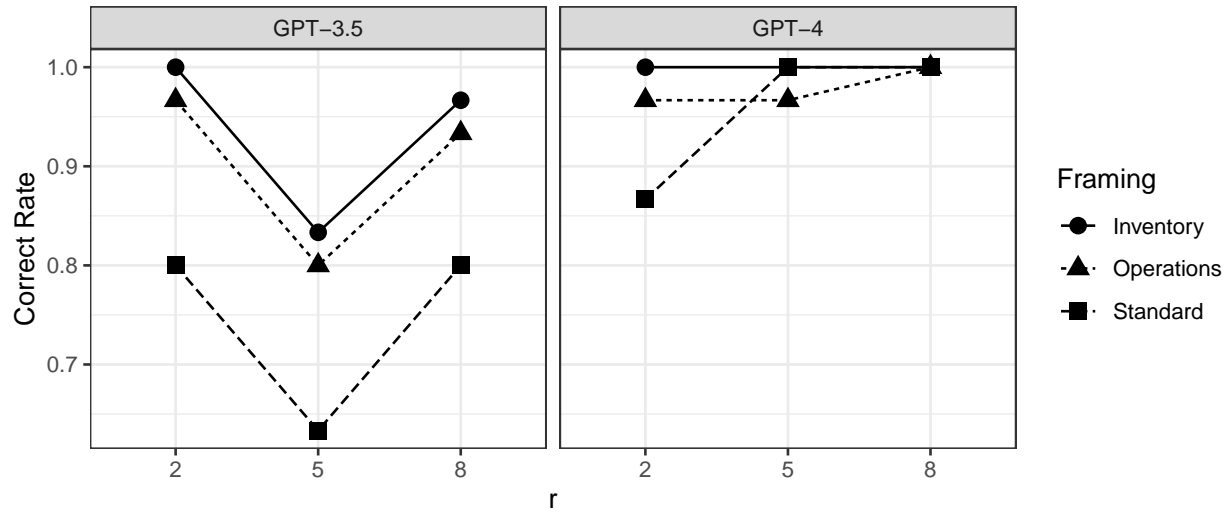
ec28

Author: *A Manager and an AI Walk into a Bar*
Article submitted to *Manufacturing & Service Operations Management*; manuscript no. MSOM-2023-0279.R2

Figure B3 **Availability heuristics - correct rate by number of bus stops (r), context and model**

	Model 1	Model 2
Intercept	-1.28^{***}	-0.92^{*}
	(0.35)	(0.37)
r=5	0.81^{*}	0.13
	(0.40)	(0.51)
r=8	-0.33	-0.64
	(0.47)	(0.58)
Context=Inventory	-1.76^{***}	-8.81
	(0.48)	(22.32)
Context=Operations	-1.09^{**}	-1.85^{*}
	(0.39)	(0.81)
Model=4	-2.08^{***}	-2.07^{***}
	(0.46)	(0.46)
r=5 * Context=Inventory		7.83
		(22.32)
r=8 * Context=Inventory		6.88
		(22.34)
r=5 * Context=Operations		1.27
		(0.98)
r=8 * Context=Operations		0.64
		(1.18)
Num. obs.	540	540
Pseudo R^2	0.22	0.25
L.R.	54.92	61.26

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 6 **Availability heuristics - effects of stops, contexts and models on error rate**

r=5, different from the monotone trend found in human results. Context also matters: both Operations and Inventory contexts are associated with lower odds of error. GPT-4 is also associated with significantly lower odds of error compared to GPT-3.5 (coefficient of -2.08 or an odds ratio of ~0.13, p<0.0001).

We also examine the moderation effect of context on the relationship between the number of bus stops and the error rate (Model 2 in Table 6). We find no evidence of a moderation effect (Likelihood ratio test between Model 1 and Model 2 yields $\chi^2 = 6.34$ and p of 0.1754), which agrees with the visual inspection that although the context matters to the overall error rate (main effects of contexts are significant), the symmetric pattern between the number of bus stops and error rate is consistent across contexts.

B.1.4.   Base rate neglect and Bayesian updating   When asked to estimate a risk that is a Bayesian posterior in nature, people's tendency to ignore the base rate in the process is called the base rate neglect. Take the question by Casscells et al. (1978) as an example:

> *If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?*

The fallacy stems from decision-makers ignoring the fact that the probability of the disease is low and most positive test results are false positives. 11 out of 60 participants from Harvard Medical School gave the correct answer of around 2%, while their average estimate was 56%, and nearly half of the subjects responded 95%. This experiment was repeated in 2013, and the results were not significantly different (Manrai et al. 2014). We prompted GPT with the exact same question, as well as two questions reframed in OM contexts. The details of the vignettes by contexts can be found in Appendix A. We tested these three contexts on both GPT-3.5 and GPT-4, and the distributions of GPT's responses are shown in histograms in Figure B4. The vertical line marks the correct answer of 0.0196.
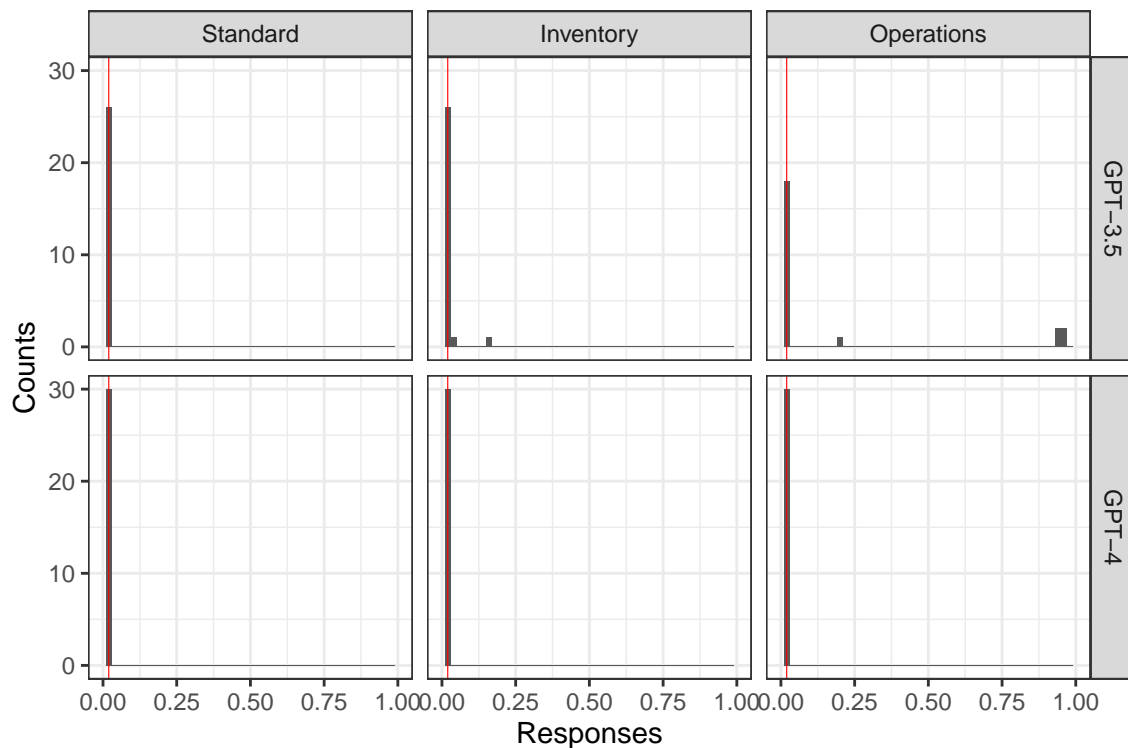

Figure B4   Base rate neglect- distributions of GPT response by models and contexts

ec30
Author: *A Manager and an AI Walk into a Bar*
Article submitted to *Manufacturing & Service Operations Management*; manuscript no. MSOM-2023-0279.R2

Overall, GPT is largely unbiased across models and contexts. Compared to subjects in Casscells et al. (1978), GPT performs significantly better. Fisher's exact test yields an odds ratio of 27.41 with a p-value <0.0001 between GPT-3.5 and a human's odds of providing the correct answer in the Standard framing of the test. The difference is even larger between GPT-4 and human performance (p<0.0001). The OR in this case is infinity because GPT-4 did not make any mistakes. We have some evidence that context matters in GPT-3.5 according to the borderline result from the Pearson's Chi-squared test ($\chi^2 = 9.74$, p =0.0125) with Operations framing having less accuracy, while GPT-4 is correct 100% of the time across contexts. However, qualitatively, GPT is consistently less biased than the human-level estimate of 56% across models and contexts (p-values <0.0001).

B.1.5. Probability Weighting Probability weighting is applying additional "decision weights" when evaluating potential outcomes with stated probability information. Kahneman and Tversky (1979) describe a Russian roulette game:

> *Suppose you are compelled to play Russian roulette, but are given the opportunity to purchase the removal of one bullet from the loaded gun. Would you pay as much to reduce the number of bullets from four to three as you would to reduce the number of bullets from one to zero?*

Although removing one bullet will always decrease the probability of death by 1/6, people value the removal of the last bullet much higher than the fourth bullet[18]. We performed the same test on GPT, as well as two reframed questions in OM contexts. There are 4 potential outcomes when we test GPT on its preference: a) it would pay the equal amount (i.e. equal weights); b) it would pay more to reduce the bullets from 4 to 3; c) it would pay more to reduce the bullet from 1 to 0; and d) it refuses to provide a preference, e.g. suggesting it is an AI and does not have preferences. The distributions of GPT's preference by models and contexts are shown in Figure B5.
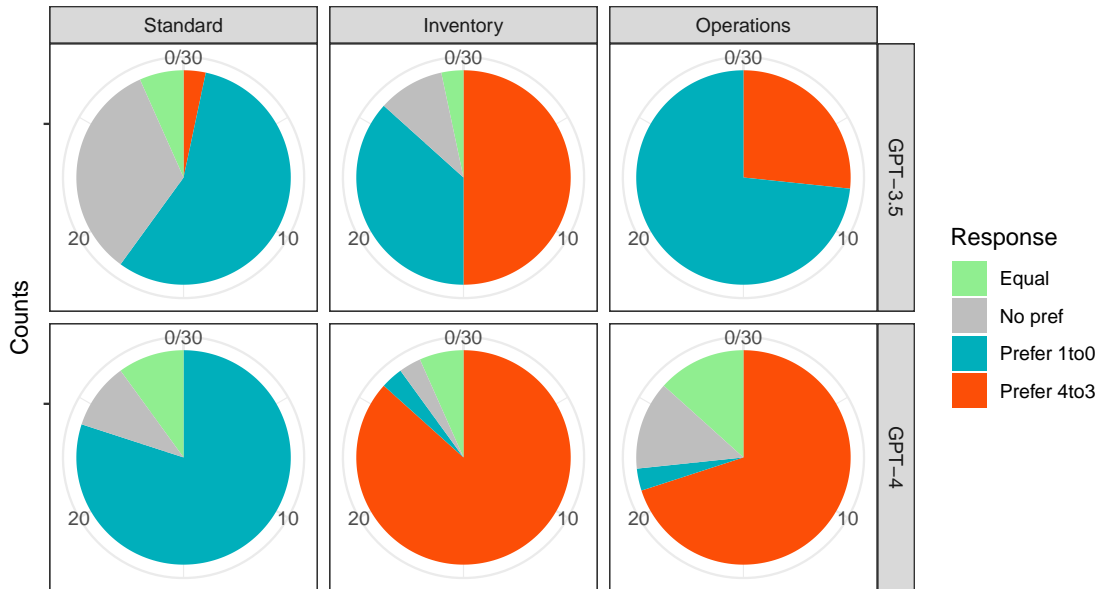

Figure B5 **Probability weighting- GPT preference distributions by models and contexts**

Figure B5 shows that GPT tends to associate additional weight to probabilities of potential outcomes, although both context and model seem to influence the weightings. Across all contexts and models, we found that the option to weigh probabilities equally is never a dominant response. Binomial exact tests suggest the proportions of GPT equally weighing the probabilities are significantly lower than 50% (p-values <0.001) in all scenarios. Statistical tests also suggest GPT's preferences are significantly context-dependent in both GPT-3.5 ($\chi^2 = 30.04$, p <0.001) and GPT-4 ($\chi^2 = 67.41$, p <0.001). GPT's preferences also differ significantly between models in Inventory ($\chi^2 = 12.62$, p =0.002) and Operations contexts ($\chi^2 = 33$, p <0.001), but are consistent in the Standard context ($\chi^2 = 6.16$, p =0.1069) where safety guardrails dominate.

Moreover, GPT exhibits strong probability weighting, but the direction of the over-weighing varies. We performed post-hoc comparisons between the options to over-weigh low (1 to 0) vs. high (4 to 3) probabilities across contexts and models. In the standard context (Russian Roulette problem), both versions of GPT assign more weight to eliminate the last bullet completely (p-values <0.001), similar to human preferences. When we reframe the same problem in OM contexts, GPT-3.5 over-weighs either lower or higher probability similarly in the Inventory context (p = 0.5572) and Operations context (p = 0.016, borderline significant trend towards over-weighing low probability). GPT-4 consistently assigns more value to reduce the undesirable event with higher probability in both Inventory and Operations contexts (p-values <0.001).

A potential explanation for the shift of weightings in OM contexts is that the implied life and death consequences in the Russian Roulette game triggers GPT's safety guardrails and results in answers favoring the choice with a safer baseline condition (the safety guardrails may also be responsible for the higher proportions of "no preference" responses where GPT refuses to give recommendations expressing safety and ethics concerns), while in OM settings GPT strategically considers the end-to-end business operations and favors uncertainties that are more evenly distributed across the supply and production processes. Although on the surface the specific direction of GPT over-weighing a probability seems to depend on context, we observe a consistent tendency to assign additional weights to probabilities in GPT decision-making throughout contexts.

B.1.6.   Overconfidence   Overconfidence refers to the bias of decision-makers overestimating their probability of being correct. Fischhoff et al. (1977) asked subjects a series of knowledge questions, such as "*Is absinthe a liqueur or a precious stone?*", then calibrated their confidence against their actual performance. GPT, with its vast general knowledge, tends to provide correct answers to knowledge-focused tests consistently with high confidence, thus making calibration difficult (calibration would require tasks with varying degrees of success rate). So, we ask GPT to provide their confidence levels in the behavioral tests presented in this paper that have factually correct answers (as opposed to preference-based tasks). These tests include the Linda problem (conjunction fallacy), the disease positive predictive values (PPV) calculation (base rate neglect), the bus stop pattern calculation (availability heuristics), CRT (cognitive reflection), and the four-card selection task (confirmation bias), as well as their corresponding reframed tests in the OM settings. We calculate GPT's average self-estimated confidence level as well as its performance in each test, and summarize the calibration results in Figure B6 by context and model. Note, some of these tests are discussed after this section, but we merely use them as data points here.
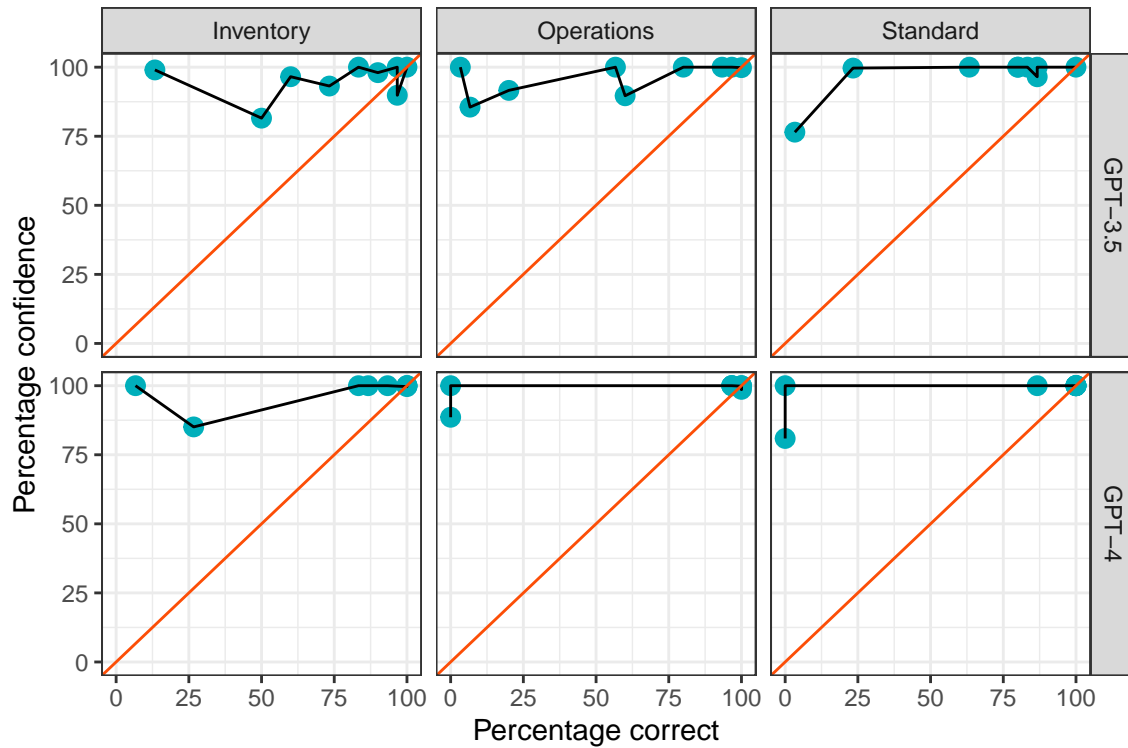
ec32   **Author:** *A Manager and an AI Walk into a Bar*

Article submitted to *Manufacturing & Service Operations Management*; manuscript no. MSOM-2023-0279.R2

Figure B6   **Overconfidence - calibration of GPT's confidence and performance**

An unbiased decision maker would have a calibration curve close to the diagonal line, whereas GPT seems to often over-estimate its rate of success. However, caution is needed when comparing this result to the general knowledge test, as GPT is very good at factual questions and is rightly confident. However, here we show that it *can* be overconfident in decision-making.

B.1.7. Ambiguity aversion   Ambiguity aversion is a decision maker's tendency to avoid choices with uncertain probability information. Ellsberg (1961) designed the following experiment questions:

> Test 1: *There is an urn with 30 red balls and 60 other balls that are either black or yellow.*
> *Choose among the following two options:*
> *A: $100 if you draw a red ball. B: $100 if you draw a black ball.*
>
> Test 2: *You must also choose between these two options:*
> *C: $100 if you draw a red or yellow ball. D: $100 if you draw a black or yellow ball.*

If a subject strictly prefers A over B, then she should also prefer C over D. However, a decision maker that is ambiguity averse may prefer A and D. We pose the same questions to GPT, but with an added sentence, "we do not know the exact numbers of black balls or yellow balls, but the total number of black and yellow balls is 60," to underscore the uncertainty of the situation. As usual, we provide two OM context vignettes in addition to the Standard context, and perform the tests on both GPT-3.5 and GPT-4. In each condition, we perform Chi-squared test (Fisher's exact test when there are low cell counts) to analyze the contingency tables of GPT responses.
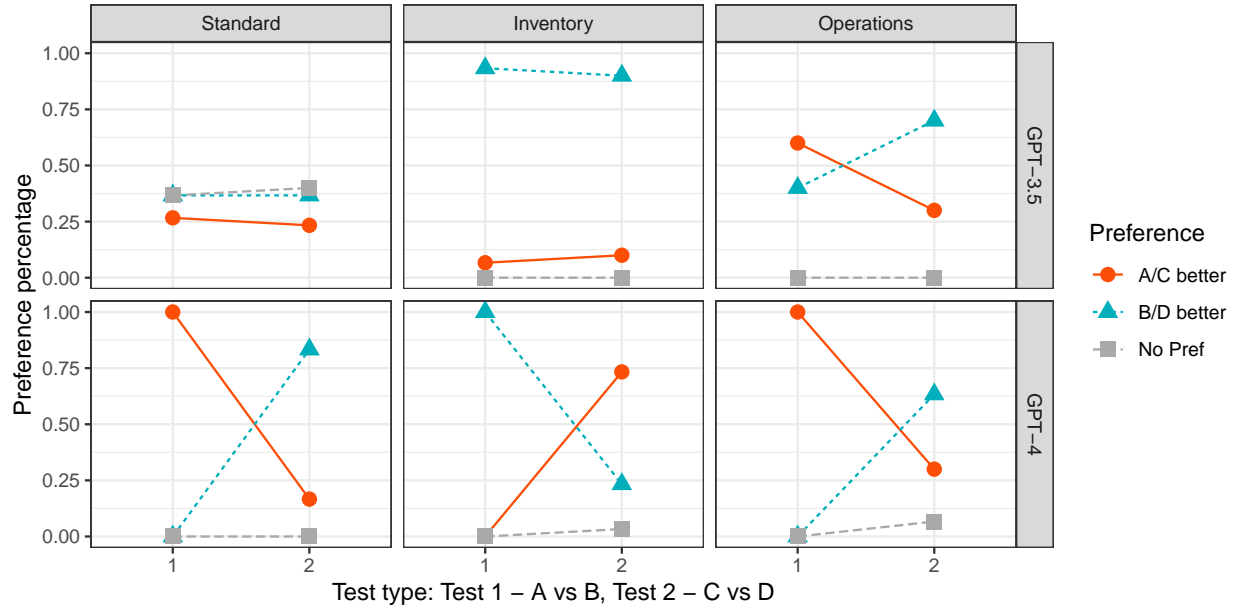
Figure B7   Ambiguity aversion - GPT preferences by tests, models and contexts

Figure B7 highlights the proportions of GPT's preferences in Test 1 and Test 2. We use the same symbols for the corresponding choices A and C, as well as B and D between Test 1 and Test 2. If ambiguity matters, we would observe a flip on the preference between Test 1 and Test 2, graphically a cross. Specifically, for ambiguity aversion we would observe a preference for A higher than a preference for C (A/C proportions go down from Test 1 to Test 2), and a preference for B lower than a preference for D (i.e. B/D proportions go up from Test 1 to Test 2). If there is no ambiguity aversion, we would not observe a significant cross or change in preferences. Lastly, in ambiguity seeking decision-makers, we would observe changes in the opposite direction to ambiguity aversion.

Even with additional clarifications in our vignettes, GPT-3.5 struggled to understand that even though the individual number of black and yellow balls was unknown, their total was fixed without ambiguity. As a result, its responses were often illogical. It sometimes could not determine which choice was better. Thus, GPT-3.5 only showed borderline significant ambiguity aversion in the Operations context ($\chi^2 = 4.31$, p =0.0379), and showed no signs of ambiguity aversion in the other two contexts (Standard: $\chi^2 = 0.11$, p =0.9464; Inventory: $\chi^2 = 0.22$, p =1). In contrast, GPT-4 seemed to be able to generate relatively logical recommendations most of the time. Interestingly, we observed prominent shifts of preferences between Test 1 and Test 2 in GPT-4 across all three contexts. However, GPT-4 exhibited ambiguity aversion in the Standard ($\chi^2 = 39.5$, p =0) and Operations ($\chi^2 = 37.3$, p =5×10^{-4}) contexts, while in the Inventory context, GPT-4 was ambiguity seeking ($\chi^2 = 32.31$, p =5×10^{-4}). Similar to the coin-toss experiments, we observe here a more severe bias arises from the more capable model, once it "understands" the nature of the question.

B.2.   Biases in Evaluation of Outcomes

B.2.1.   Risk Aversion and Scaling   Risk aversion is the tendency to prefer certain outcomes when making decisions under uncertainty. To gain insight into risk aversion, we posed two questions to GPT.

ec34

Author: *A Manager and an AI Walk into a Bar*
Article submitted to *Manufacturing & Service Operations Management*; manuscript no. MSOM-2023-0279.R2

These elicitation tasks present a pair of options (e.g., lotteries) A and B, where A carries lower risk with fixed expected returns, and B carries higher risk, with the expected return varying between the two questions. In the first test, B has an expected value equal to A. In the second test, the expected value of B is higher, making the riskier option more appealing. Using Table 5.1 from Davis (2018) to select the values for the options, we asked ChatGPT in the Standard context:

> Test 1: *Choose between two lotteries A and B, which one is better? A: 50% chance to win $5.5 and 50% chance to win $4.5 B: 50% chance to win $9 and 50% chance to win $1.*
> Test 2: *Choose between two lotteries A and B, which one is better? A: 50% chance to win $5.5 and 50% chance to win $4.5 B: 50% chance to win $9.5 and 50% chance to win $1.*

Figure B8 displays the resulting choices between opting for the safe or risky choice across contexts. Chi-squared tests with simulations suggest GPT's preferences are rather stable going from a pair of lotteries with equal expected payoffs (Test 1) to another pair with slightly unequal payoffs (Test 2) without significant differences at the .01 level. This is markedly different from rational decision-making (a rational decision-maker would have no preference in Test 1 resulting in a coin-toss between risky and certain options, but strong preference towards the risky option in Test 2). Directionally speaking, GPT-3.5 slightly increases its degree of risk-taking in the Standard ($\chi^2 = 2.26$, p $=0.355$) and Inventory ($\chi^2 = 0.069$, p $=0.792$) contexts from Test 1 to Test 2, but borderline-significantly decreases its risk-taking ($\chi^2 = 4.01$, p $=0.0453$) in the Operations context. Similarly, GPT-4 slightly increases its degree of risk-taking from Test 1 to 2 in Standard (p=1) and Operations (p=0.747) contexts and slightly decreases it in the Inventory (p=0.784) context.
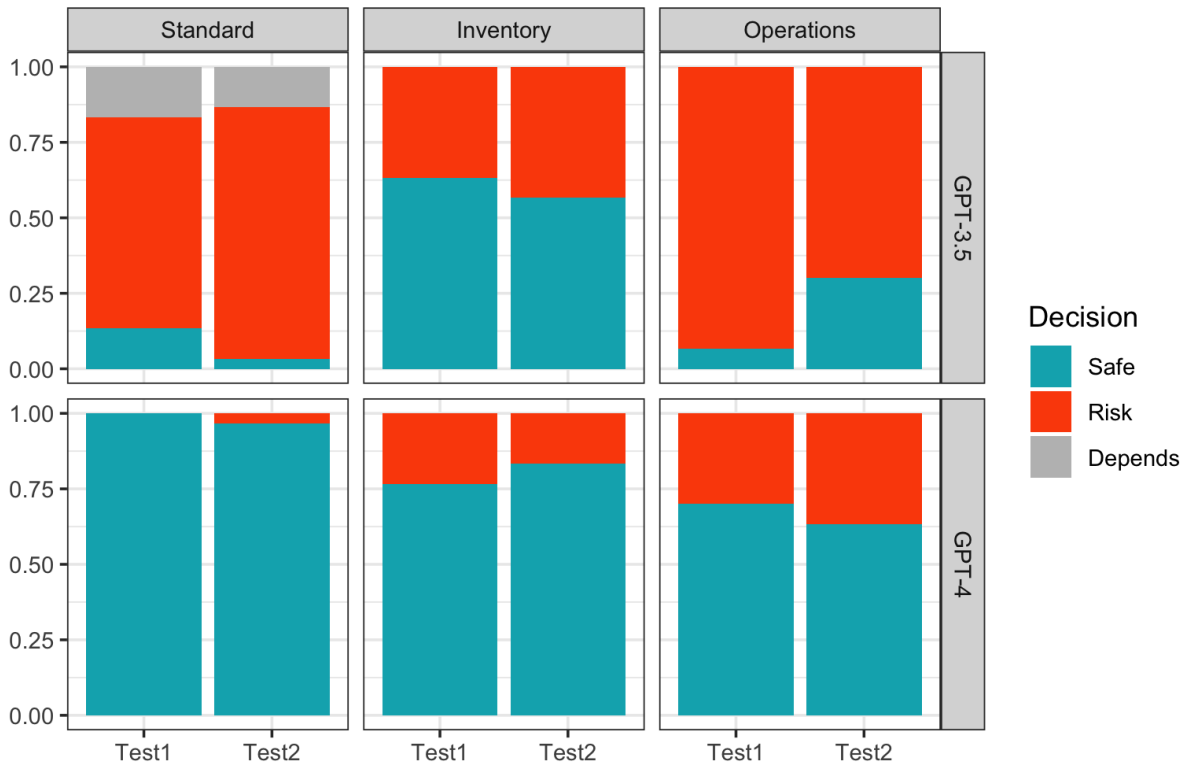

Figure B8   **Risk aversion - distribution of risk response across, tests, contexts, and models.**

More interesting comparisons are between the baseline risk preferences of GPT-3.5 and GPT-4. In all cases, GPT-3.5 exhibits a greater preference for risk compared to GPT-4. While the changes are not sizeable for Test1 ($p = 0.398$) and Test 2 ($p = 0.049$) in the Inventory context, Figure B8 shows that the change in risk aversion is very substantial, particularly for the Standard context where the decrease in the selection of risk is 84% and 93% for Test 1 and Test 2 respectively ($p < 0.001$).

In the Standard and the Inventory contexts, GPT-3.5 shows high levels of risk-taking, which is different from the prototypical risk-averse behavior. Binomial exact tests comparing the proportions of gamble responses to 50% reveal significant results in Test 1 for the Operations context (p <0.001) and Test 2 for the Standard context (p <0.001). In the Operations context, GPT-3.5 also departs from prototypical human behavior. It is neither risk-averse nor risk-seeking in both Test 1 (p=0.201) and Test 2 (p=0.585). In contrast, GPT-4 is generally risk-averse. Binomial exact tests show that GPT-4 prefers certainty in both Test 1 and Test 2 and in both Standard and Inventory contexts (p-values <0.01). In the Operations context, GPT-4 is borderline risk-averse in Test 1 (p = 0.0428) and neither risk-averse or risk-seeking in Test 2 (p = 0.201).

B.2.2. Prospect Theory Decision-makers' tendency to be risk-averse in gains and risk-seeking in losses is a well-known aspect of prospect theory. Initial evidence of this behavior stems from Kahneman and Tversky (1979), who asked experimental participants the following questions:

> Test 1: *Would you rather play a gamble where there is an 80% chance of earning $4000, or receive $3000 with certainty?*
> Test 2: *Now consider the following choice: Would you rather play a gamble where there is an 80% chance of losing $4000, or pay $3000 with certainty?*

Interestingly, 80% of respondents chose a $3000 gain with certainty, despite the expected value of the gamble being higher (e.g., $3200), whereas only 8% chose a $3000 loss with certainty, despite it also having a higher expected value. This supports the finding that people are often risk-adverse in the domain of gains and risk-seeking in the domain of losses.

We repeated this experiment on GPT with the Standard and the two OM variants. Again, we classified GPT's responses into three categories: ones that clearly favored certainty, those that clearly favored a gamble, and those where the response depended (e.g., it explained the situation and left the decision to us).

The results, summarized in Figure B9, show that GPT - irrespective of the model or context - prefers certainty in the domain of losses (p-values <0.001). GPT's propensity for risk-aversion in the domain of losses clearly contrasts typical human decision-making (risk-seeking in losses) but is in alignment with rationality (certainty options in losses have higher expected returns).

In the case of gains, GPT's preference is both model and context-dependent. In the Standard and Inventory contexts, GPT-3.5 does not have a dominant preference. Binomial exact tests comparing the proportions of gamble vs. certainty do not reject the null that they are 50-50 in the Standard (p =0.455) and Inventory (p=0.0241, borderline and trends towards certainty) contexts. On the other hand, GPT-4 shows strong preference to certainty in the Standard and Inventory contexts (p-values <0.001). In contrast, in the Operations context, both models show strong preference to gamble in gains (p-values <0.01), opposite to typical human decision-making (risk-averse in gains) but in alignment with rationality (gamble options in gains have higher expected returns).
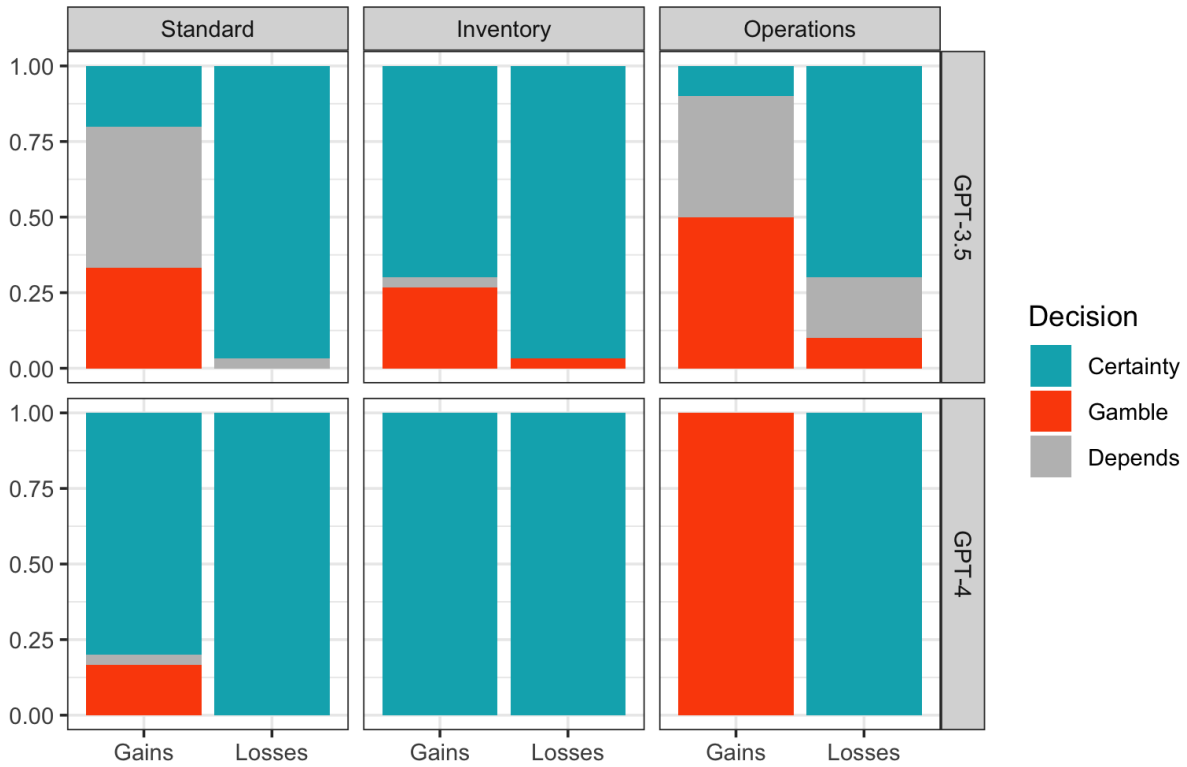
ec36

Author: *A Manager and an AI Walk into a Bar*
Article submitted to *Manufacturing & Service Operations Management*; manuscript no. MSOM-2023-0279.R2

Figure B9 **Prospect theory - distribution of preferences between gambles and certainty by context and models.**

Taken together, GPT behaves differently from the prototypical prospect theory bias observed in humans across models and contexts. However, whether it is unbiased or biased differently depends on context. Recall, the rational decision-maker guided by expectation maximization would lead to preferences for the gamble in gains and for certainty in losses in our tests. In the Standard and Inventory contexts, both GPT models are irrational in gains, but rational in losses, making them biased differently from the prototypical human. In the Operations context, however, both models are rational.

B.2.3.   Framing   Also part of the prospect theory's prediction is the idea that framing a set of choices as gain or loss may change decisions. We repeat the set of experiments by Tversky and Kahneman (1981) on GPT with minimal wording changes:

> Gain frame. *Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that that exact scientific estimate of the consequences of the programs is as follows:*
> *Program A: 200 people will be saved.*
> *Program B: There is a 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.*
> *Which program should I chose?*

> Loss frame. *Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that that exact scientific estimate of the consequences of the programs is as follows:*
> *Program C: 400 people will die.*
> *Program D: There is a 1/3 probability that nobody will die and 2/3 probability that 600 people will die.*
> *Which program should I chose?*

Of human respondents, 72% chose program A, but only 22% chose program C (Tversky and Kahneman 1981). Although programs A and C are identical, people prefer certainty in gain frames, but prefer the gamble under loss frames. For the Inventory context, we adopted the same test as Tokar et al. (2016) (Experiment 1.1) and made a similar second vignette for the Operations context. We classified GPT's responses into the same three categories as in our Prospect Theory analysis. The results are presented in Figure B10.
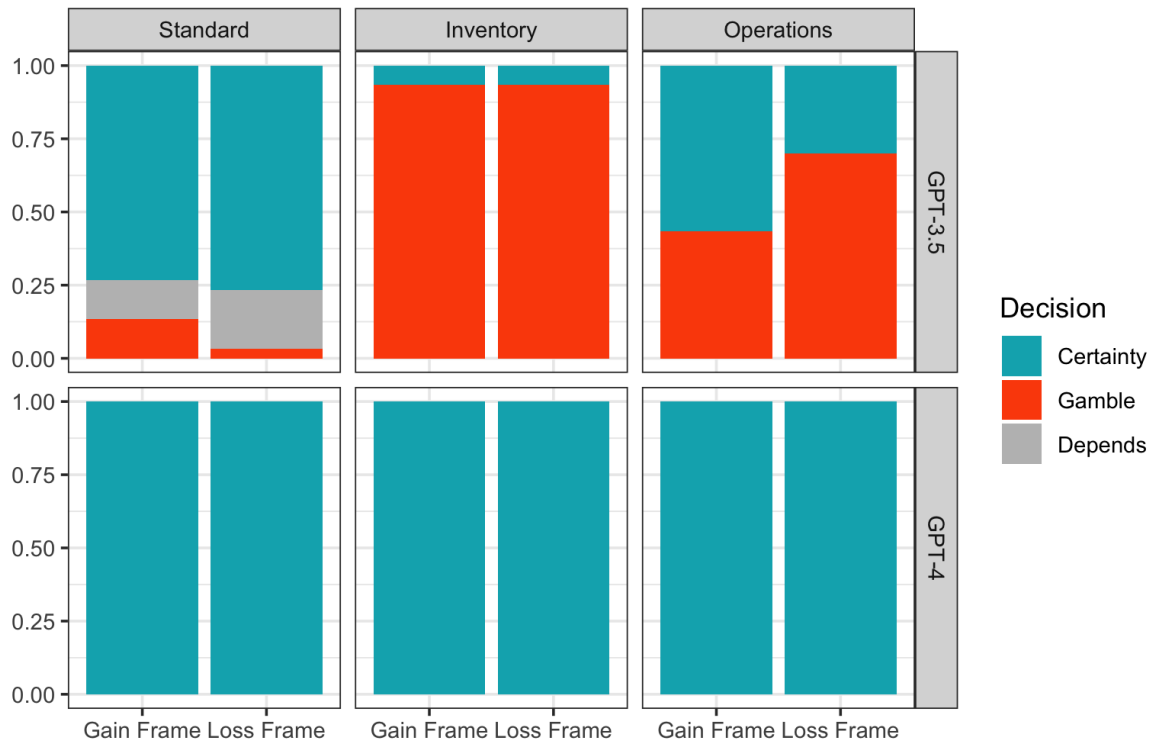

Figure B10   Framing - distribution of certainty and gamble preferences across frames, models and contexts.

Contrasting typical human behavior, GPT did not have systematic differences in its preference between gambles and certainty across frames. Only in the Operations context, GPT-3.5 was demonstrating some evidence of difference ($\chi^2 = 3.33$, $p = 0.068$). When we compare the models across frames and contexts, we see that GPT-4 has a stronger preference for certainty compared to GPT-3.5, which aligns with GPT-4's risk-averse preferences. This is particularly evident in the Inventory and Operations contexts (all $\chi^2 > 14.14$, $p < 0.001$). Although GPT-4 consistently prefers certainty in all contexts, GPT-3.5's preference for certainty is highly context-dependent, despite not being frame-dependent. In the Operations context, certainty is

selected in 43% of cases, which is significantly lower than the 90% of cases selecting certainty in the Standard context ($\chi^2 = 23.95$, $p < 0.001$) and significantly higher than the 7% of cases selecting certainty in the Inventory Context ($\chi^2 = 19.60$, $p < 0.001$).

B.2.4. Anticipated Regret Anticipated regret or regret aversion is the theory that when people make decisions, they consider not only the utilities of their options but also the counterfactual regret of their choices (Bell 1982, Loomes and Sugden 1982). Hetts et al. (2000) conducted a questionnaire study on 164 UCLA undergraduate students in a scenario where they needed to choose between checking their car or arriving to a quiz on time. The control condition proceeded as follows:

> Control. *Imagine that you commute to school and that you park your car in one of the lots on campus. Further imagine that, on this day, you are walking to class in a bit of a rush because you have a quiz that you do not want to be late for. On the way to class, however, you get a strange feeling that you may have left your car door unlocked. Try as you might, you cannot be absolutely certain whether or not you locked your door.*
> *Now you must choose: would you go back to check your car or go straight to class for the quiz?*

In the omission regret, the potential for inaction regret (i.e., not going back to the car) is made salient, by adding the following before asking participants to choose an action:

> Omission regret. *Think for a minute about how upset you would be if you didn't go back to check your car, and later that day your car was burglarized.*
> *Now you must choose: would you go back to check your car or go straight to class for the quiz?*

In the commission regret, the potential for regret from taking an action (i.e., going back to the car) is made salient, by adding the following before asking participants to choose an action:

> Commission regret. *Think for a minute about how upset you would be if you went back to check your car, found out that it was locked all along, and ended up being late for your quiz.*
> *Now you must choose: would you go back to check your car or go straight to class for the quiz?*

The manipulation of the salience of the anticipated counterfactual regrets significantly influenced choices, such that making commission [omission] regret more salient leads to inaction [action]. Hetts et al. (2000) found that 69.1% of the subjects in the omission regret manipulation group preferred to check their car, while only 34.5% in the commission regret case check their car. The control group was in between at 46.3% checking their car. We provided GPT with the same scenario, as well as two others related to the Inventory and Operations frame. The results, which are classified as either Inaction, Action, or Depends are presented in Figure B11.

GPT-3.5 essentially always takes the action and is not impacted by the salience of the counterfactual regrets. In contrast, in GPT-4, for the Standard and the Operations context, there are clear significant differences between the commission (100% in both contexts prefer inaction) and omission groups (3% and 0% prefer inaction), which is consistent with anticipated regret ($\chi^2 > 23.73$, $p < 0.001$). In the Inventory context, GPT-4, similar to GPT-3.5, prefers action. Thus, whether GPT makes decisions driven by anticipated regret depends on both the model and context.
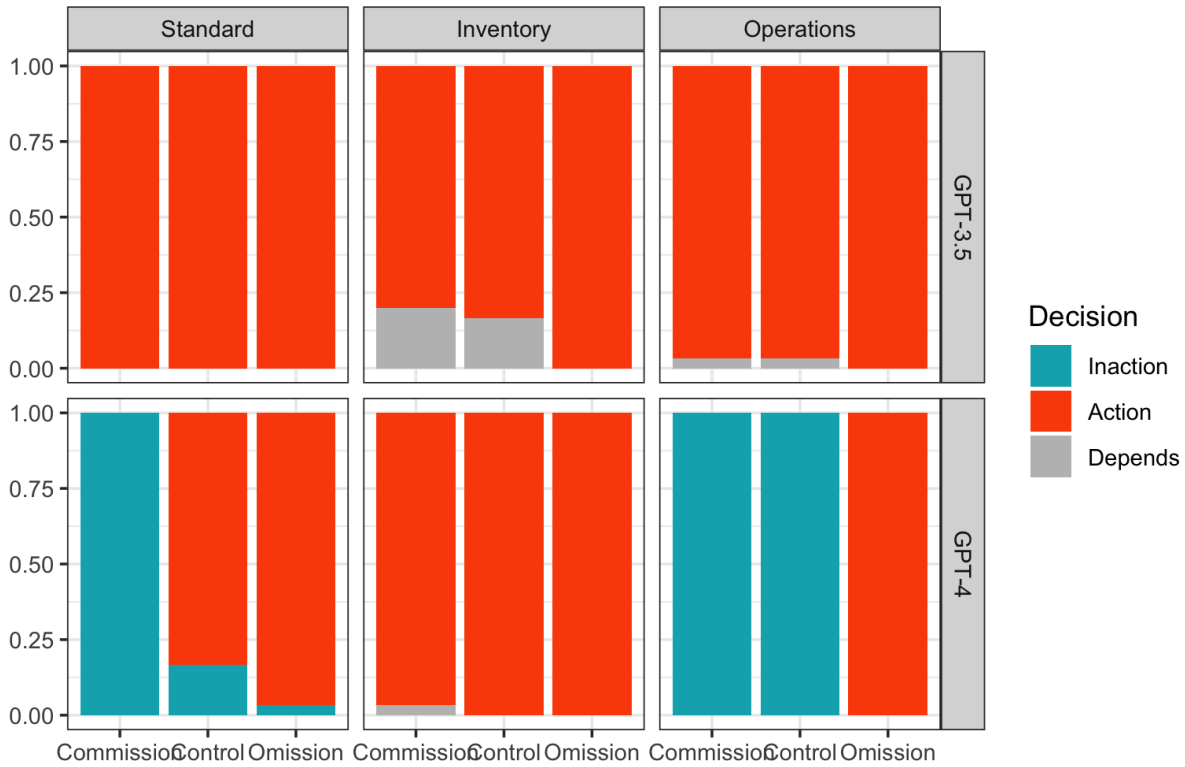
Figure B11   **Anticipated regret - distribution of action and inaction preferences across salience conditions, models and contexts.**

B.2.5.   Mental Accounting and Reference Dependence   One aspect of mental accounting concerns the granularity in people's evaluation of gains and losses, as gains and losses may feel different depending on how they are mentally segregated or combined. Thaler (1985) concluded that multiple gains should be segregated, while multiple losses should be integrated; mixed gains (a bigger gain + a smaller loss) should be integrated, and mixed losses (a bigger loss + smaller gain) should be segregated.

Reference dependence suggests that our feelings of gains and losses depend on the reference we set. Heath et al. (1995) adopted the experiment in Thaler (1985) on mental accounting to test reference dependence in the context of mental accounting by framing the gains and losses in percentages as opposed to absolute dollar amounts and tested people's preferences in hypothetical scenarios. They found the preference for integration and segregation in mental accounting was reference-dependent. Specifically, for mixed gains, respondents preferred integration in an absolute frame. However, when percentages were made salient (i.e., in mixed and relative frames), preferences switched towards segregation.

We replicate one set of questions in Heath et al. (1995) to test both phenomena since the first question of the set is a direct adoption of Thaler (1985)'s work on mental accounting. The second and third questions are of mixed and relative frames. The questions are as follows:

Absolute frame. *Mr. A's couch was priced originally at $1,300 but is now reduced to $1,250. Mr. B's chair was priced originally at $300 and his couch was priced at $1,000. His chair is now reduced to $200 and his couch is now increased to $1,050.*
Who was happier? Mr. A, Mr. B, or no difference?

Mixed frame. *Mr. A's couch was priced originally at $1,300 but is now reduced by 3.8 percent to $1,250. Mr. B's chair was priced originally at $300 and his couch was priced at $1,000. His chair is now reduced by 33 percent to $200, and his couch is now increased by 5 percent to $1,050. Who was happier? Mr. A, Mr. B, or no difference?*

Relative frame. *Mr. A's couch was priced originally at $1,300 but is now reduced by 3.8 percent. Mr. B's chair was priced originally at $300 and his couch was priced at $1,000. His chair is now reduced by 33 percent, and his couch is now increased by 5 percent. Who was happier? Mr. A, Mr. B, or no difference?*

In addition, we used these questions to create an Inventory and Operational context (similar to Becker-Peth et al. (2013)). If GPT exhibits mental accounting, then Mr. A will be selected with a higher frequency in the absolute frames. If GPT exhibits reference dependence in the context of mental accounting, then we would observe a trend of favoring Mr. B over Mr. A when moving from the absolute frame to the mixed and the relative frames. The results are shown in Figure B12.
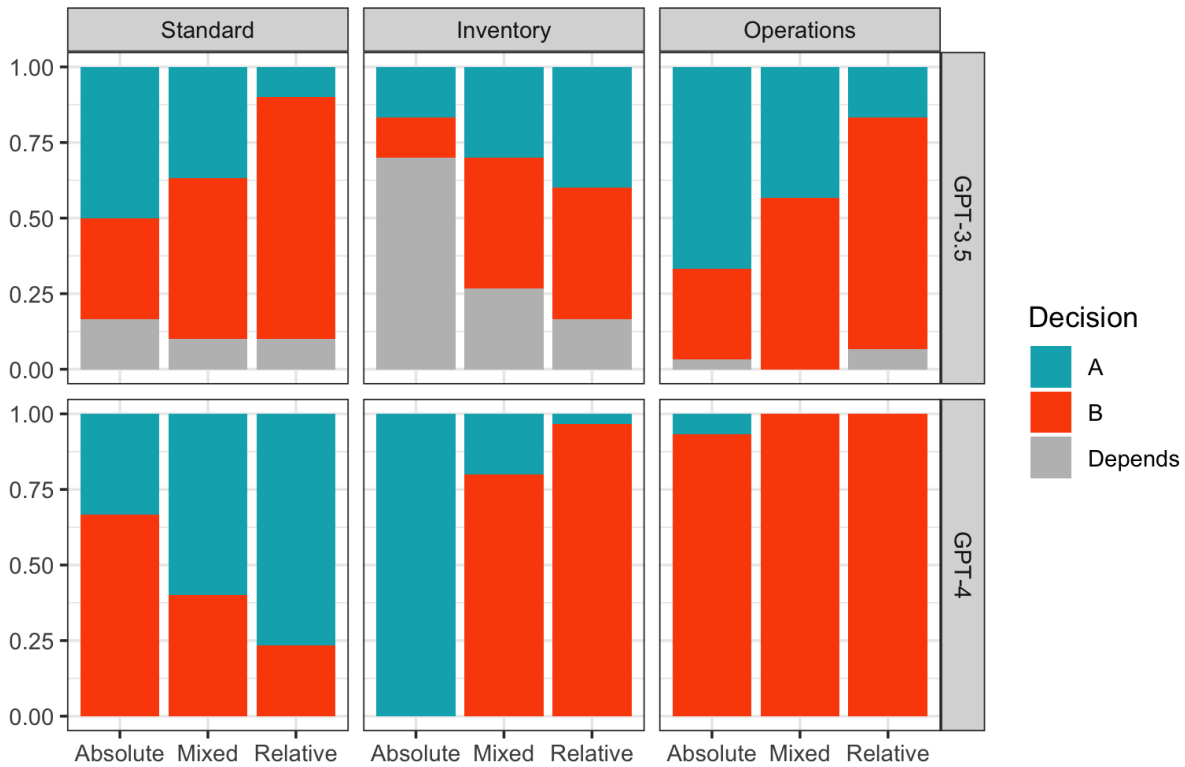

Figure B12   **Mental accounting and reference dependence - distribution of preferences across frames, models and contexts.**

Regarding mental accounting, GPT-3.5 is largely rational. In the absolute frame, compared to a 50% benchmark, two-sided binomial exact tests reveal that GPT-3.5 does not exhibit a preference for either integration nor segregation in the Standard (p=0.424) and Inventory (p=1) contexts. Additionally, in the Inventory context, the dominant response was no preference. In the Operations context, GPT-3.5 is also rational, although it shows some evidence of a trend towards mental accounting (p=0.0614). GPT-4's preference in the absolute frame is context dependent. It is rational in the Standard context, but trends towards segregation with p=0.0987. We observe a strong preference for integration (mental accounting) in the Inventory context (p <0.001), and a strong preference for segregation in the Operations context (p <0.001).

Regarding reference dependence, in GPT-3.5, a Pearson's Chi-squared test reveals a significant association between GPT-3.5's preference and the framing of the mixed gains in the Standard context($\chi^2 = 14.371$, $p = 0.005$). Post-hoc analyses confirms that GPT-3.5 exhibits reference dependence as there is an increasing preference for Mr. B as the salience of the gains and losses in percentages increases. For example, in the Standard context in the absolute frame, GPT-3.5 selected Mr. A 60% of the time, however, this falls to 3% in the relative framework ($\chi^2 = 11.633$, $p < 0.001$). Similarly, in the Operations context, we also observe an increasing preference for Mr. B in the relative framings ($\chi^2 = 16.936$, $p < 0.001$). However, in the Inventory context, although Pearson's Chi-squared test suggests the framing significantly influences GPT-3.5's preferences ($\chi^2 = 21.011$, $p < 0.001$), this is mainly driven by the many "Depends" responses in the absolute frame. A post-hoc analysis focusing on responses with a preference suggest that unlike in the other two contexts, we do not observe reference dependence in the Inventory context ($\chi^2 = 0.596$, $p = 0.815$).

Contrasting GPT-3.5, we observe that GPT-4 behaves differently in each context, consistent with the pattern we found in mental accounting. In the Inventory context, GPT-4 behaves consistently with humans, increasing the preference for Mr. B when percentages are salient ($\chi^2 = 66.181$, $p < 0.001$). However, in the Operations context, there are no differences across the frames ($\chi^2 = 4.091$, $p = 0.314$), with GPT-4 preferring segregation (p<0.001). Interestingly, GPT-4 acts opposite to what is seen in humans in the Standard context, exhibiting an increased preference for segregation, selecting Mr. A 33% of the time in the absolute frame, but 77% of the time in the relative frame ($\chi^2 = 11.674$, $p = 0.003$).

B.2.6. Intertemporal Choice and Hyperbolic Discounting Intertemporal choices consider decisions with present and future consequences. One common assumption when modeling intertemporal choices is to apply a constant discount factor per unit of time lapsed. However, experiments such as Thaler (1981) showed that the discount factors used among average decision-makers are not constant. In contrast, they are decreasing in time and size of payoffs: people heavily discount waiting for a short time for a small payoff and would much prefer to get paid immediately but are relatively more patient on a large payoff in the distant future. This phenomenon is also called hyperbolic discounting.

We adopt one experimental condition from Thaler (1981), abbreviated by Davis (2018), and create similar vignettes for the Inventory and Operational contexts. For the Standard context. we specifically asked GPT:

> *You have won $15 in the lottery which is being held by your bank. You can take the $15 now, or wait until later. How much money would you require to make waiting 3 months [1 year] [3 years] just as attractive as getting the $15 now?*

Figure B13 demonstrates the average total interest required by GPT to be willing to take the future payoff by time frame, models and contexts. The error bars represent 95% CIs of the means of required interest.

The figure reveals that GPT-3.5 frequently requests less than $15 in return for $15 now. Although GPT-3.5 typically proposes that the equivalent future payoff depends on the interest or discount rate (i.e. it does not exhibit hyperbolic discounting), it commonly chooses a significantly low rate. Furthermore, it often misapplies the interest rate formula, leading to equivalent future payoffs that are less than $15. These tendencies become more pronounced when the timeframe shifts from 3 months to 1 or 3 years.
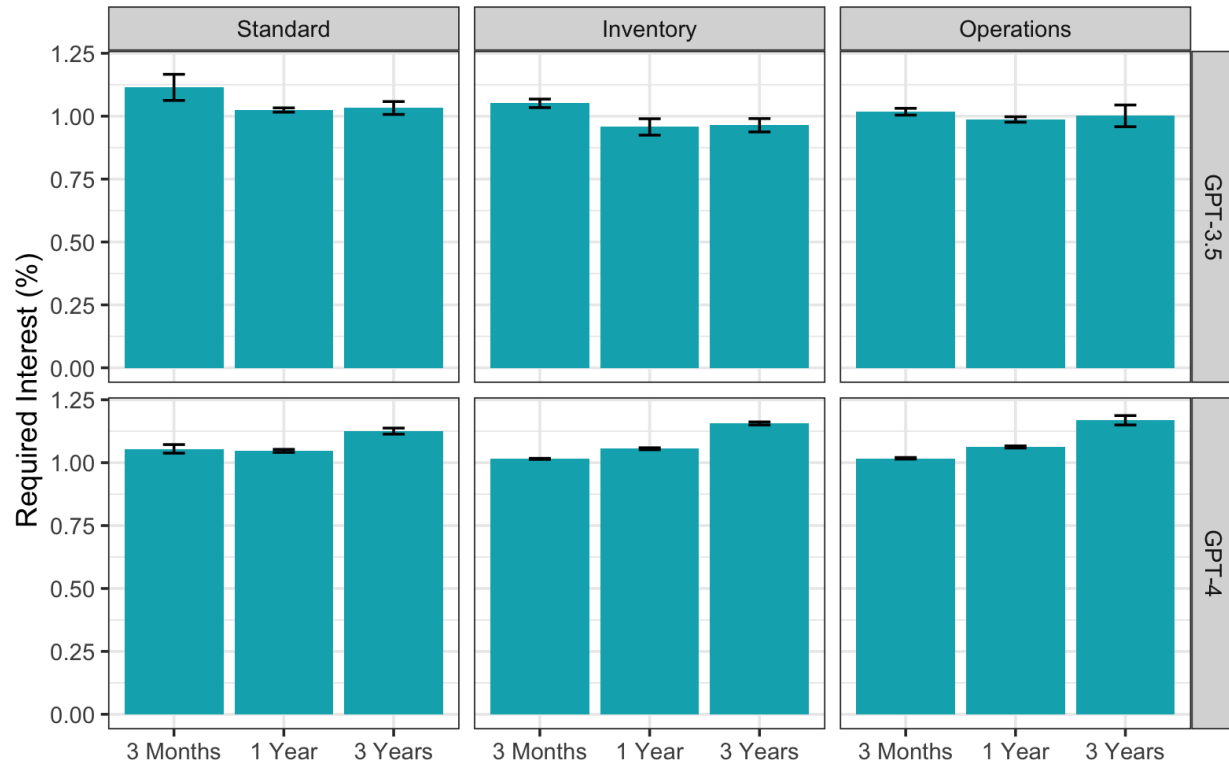

Figure B13   **Intertemporal choice - required interest on $15 for each time frame by models and contexts.**

GPT-4 also uses an interest rate to determine the equivalent return. However, compared to GPT-3.5, it is more likely to apply the formula correctly. As a result, the required interest, in most cases, clearly increases with the time frame. In particular, there is a clear increasing trend for the Inventory and Operations contexts. Notably, as GPT-4 responses are underpinned by fixed interest rates, the increases in required interest are commensurate to the wait time while the annualized rate remains constant across the three time frames. Thus GPT-4 does not exhibit hyperbolic discounting.

B.2.7.   Endowment Effect   The endowment effect refers to the higher value people often assign to an item or asset they own. This effect is typically demonstrated in experiments by the disparity between the willingness-to-accept (WTA) and willingness-to-pay (WTP). When an individual owns an item, their WTA tends to exceed their WTP, indicating that ownership increases perceived value. In a study conducted by Knetsch and Sinden (1984) involving 76 students, 38 were given the option to enter a raffle for $2, with a prize valued between $50 and $70. The remaining 38 students received free raffle tickets but had the option to redeem them for $2. The study found that 19 students who did not initially own tickets were willing to pay $2 for one (WTP > $2), while 29 students who were given tickets were unwilling to exchange them for $2 (WTA > $2). Overall, there was a statistically significant difference in the perceived value of the raffle ticket.

We provide GPT with similar scenarios, across the three contexts. However, in Knetsch and Sinden (1984), the students did not know their winning probability when they made decisions to purchase or sell tickets. To facilitate GPT providing preferences, we provided a fixed probability of winning the raffle (recall the ambiguity aversion discussion). We chose a winning probability of 2.08%, which is the actual winning probability in Knetsch and Sinden (1984). Below are the two scenarios from the Standard context that we prompted GPT with.

> WTP. *I can enter a prize draw for $2. The prize is worth $70 and my estimated winning probability is 2.08%. Should I enter the prize draw?*
> WTA. *I was given a prize draw ticket for free. The prize is worth $70 and my estimated winning probability is 2.08%. My friend is offering $2 for my ticket, should I sell it?*

Since the expected payoff of entering the raffle is less than $2, a rational decision-maker should choose to not buy when offered the option to enter the raffle, and to sell when offered the option to sell the raffle ticket. Figure B14 summarizes GPT's responses. We label the decision to buy/keep the tickets (indicating the WTP/WTA >$2) as "Own", and we label the decision not to buy/sell the ticket (WTP/WTA<$2) as "Don't Own" in the figure.

The results indicate that GPT-3.5 does not display the endowment effect in the Standard context ($\chi^2 = 0$, $p = 1$), showing nearly equal proportions of ticket ownership in both the WTA and WTP conditions. In the Inventory context, GPT-3.5 shows a borderline reverse effect ($\chi^2 = 6.98$, $p = 0.018$). Excluding the no preference responses, the fraction of responses choosing ownership are 72% for the WTA condition, and 96% for the WTP cases. The post-hoc chi-square test of these two fractions is also borderline ($\chi^2 = 5.91$, $p = 0.023$). However, in the Operations context, it clearly demonstrates the endowment effect, selecting ownership in 80% of cases for the WTA condition and never choosing ownership in the WTP condition ($\chi^2 = 36.74$, $p < 0.001$).

Conversely, GPT-4 consistently displays the endowment effect across all contexts. It always selects ownership in the WTA condition. In the Standard and Inventory context, GPT-4 almost always chooses not to own in the WTP condition (Standard: $\chi^2 = 56.07$, $p < 0.0001$. Inventory: $\chi^2 = 60.00$, $p = 0.0005$). In the Operations context, however, ownership in the WTP condition is significantly higher at 73%. Despite this, there is still a significant difference compared to the WTA condition ($\chi^2 = 9.23$, $p = 0.004$), supporting the idea that GPT-4 exhibits the endowment effect.
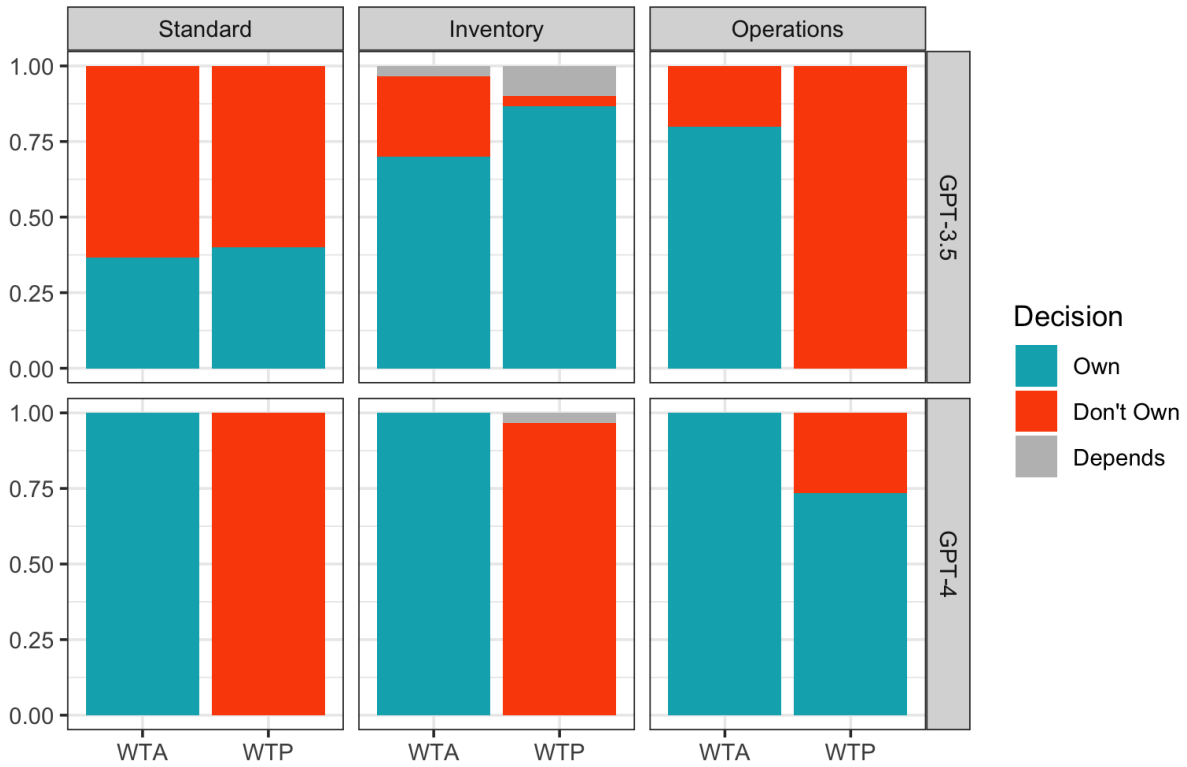
Figure B14 **Endowment - decision to own or not own the lottery ticket across treatments, models, and contexts.**

B.2.8. **Sunk Cost** The sunk cost fallacy refers to the phenomenon of decision-makers considering unrecoverable past costs when evaluating future choices. The vignette by Arkes and Blumer (1985) considers the decision to invest in a project after receiving information that the project is unlikely to be successful. In the scenario, where the project is well underway, participants were more likely to finish the project, then the scenario where the project had not yet been started, demonstrating the sunk cost fallacy. We use this same scenario, as well as the Inventory and Operation contexts. The Standard context based on Arkes and Blumer (1985) consisted of the following two scenarios:

> Scenario A: *As the president of an airline company, you have invested 10 million dollars of the company's money into a research project. The purpose was to build a plane that would not be detected by conventional radar, in other words, a radar-blank plane. When the project is 90% completed, another firm begins marketing a plane that cannot be detected by radar. Also, it is apparent that their plane is much faster and far more economical than the plane your company is building. The question is: should you invest the last 10% of the research funds to finish your radar-blank plane?*

> Scenario B: *As president of an airline company, you have received a suggestion from one of your employees. The suggestion is to use the last 1 million dollars of your research funds to develop a plane that would not be detected by conventional radar, in other words, a radar-blank plane. However, another firm has just begun marketing a plane that cannot be detected by radar. Also, it is apparent that their plane is much faster and far more economical than the plane your company could build. The question is: should you invest the last million dollars of your research funds to build the radar-blank plane proposed by your employee?*

The results are summarized in Figure B15. Across all models and contexts, GPT primarily recommends against investing. Only in the Operations context with GPT-3.5 do we see any evidence of decisions being driven by a sunk cost, where GPT recommends investing in 24% of cases in Scenario A, compared to 0% in Scenario B. However, this trend is borderline with a Fisher's exact test yielding a p-value of 0.0105. In general, there is no strong evidence of the sunk cost fallacy across models and contexts.
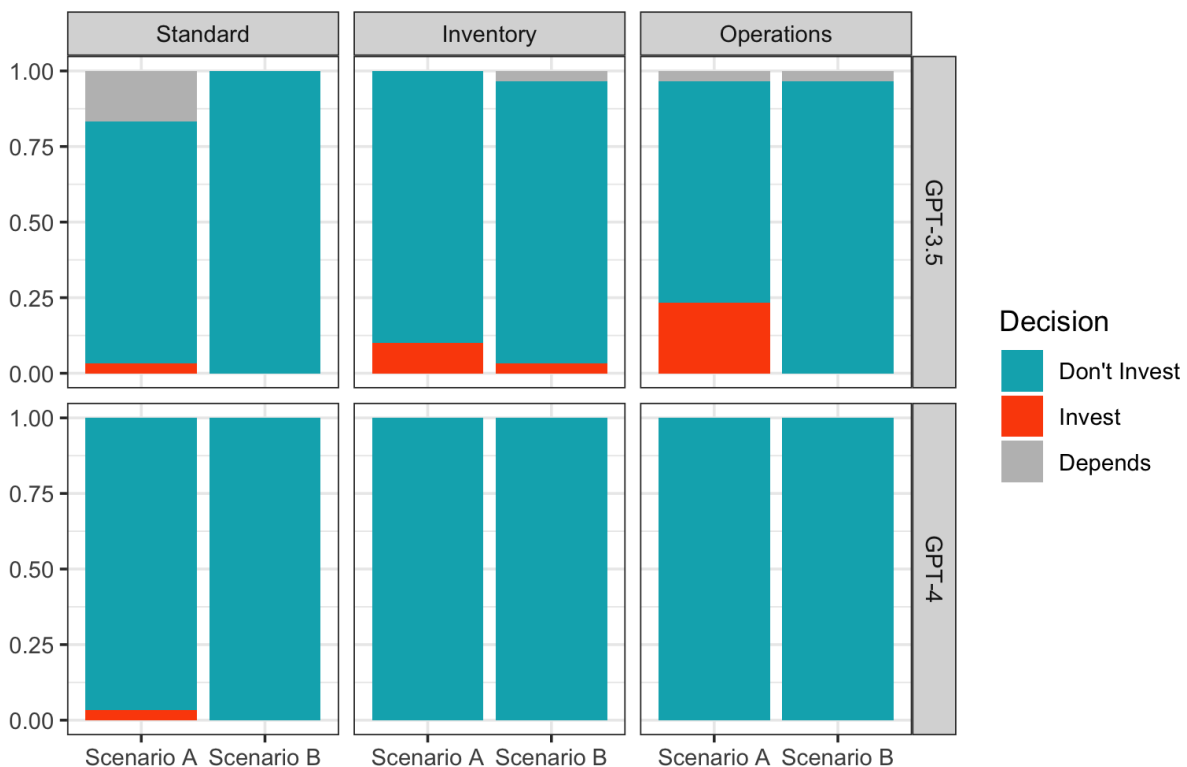

Figure B15   Sunk cost - decision to invest or not across treatments, models, and contexts

B.3.   Heuristics in Decision-Making

B.3.1.   Cognitive Reflection Test (CRT)   A System-1 decision refers to an instant, autopilot-like decision-making process, while a System-2 decision requires careful and conscious deliberations. Frederick (2005) developed a set of three tricky questions referred to as the "cognitive reflection test" or CRT. The CRT is designed so that respondents must suppress their System-1 thinking to arrive at the correct answers. The study of over 3000 participants yielded an average number of correct responses of 1.24 out of the
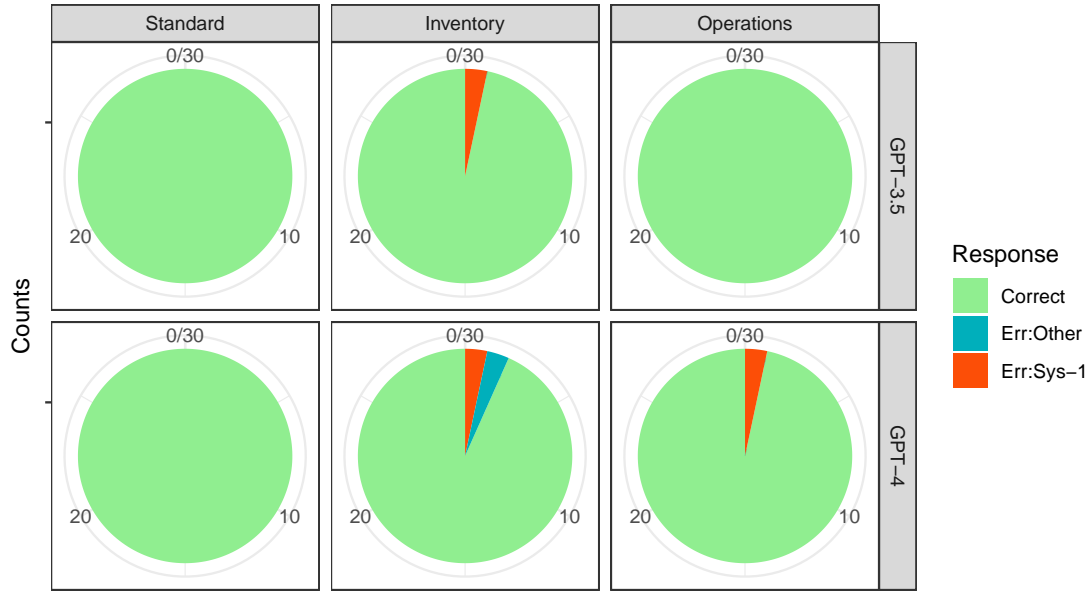
ec46

Author: *A Manager and an AI Walk into a Bar*
Article submitted to *Manufacturing & Service Operations Management*; manuscript no. MSOM-2023-0279.R2

Figure B16   **CRT Question 1 - response distributions by models and contexts**

three questions. Although LLMs like GPT do not "think" as humans do, we have seen throughout our experiments that it is capable of producing both incorrect results, such as in the "Linda problem" when the quantitative information is made implicit, and correct results, such as in the bus stop pattern calculation, when the quantitative information is made explicit. The CRT questions are as follows:

> 1. A bat and a ball cost $1.1 in total. The bat costs $1 more than the ball. How much does the ball cost?
>
> 2. If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets?
>
> 3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

We perform the CRT in Standard, Inventory and Operations contexts to GPT-3.5 and GPT-4 to study whether it relies on heuristics in decision-making.

Both versions of GPT perform very well in CRT Question 1 (CRT-1). Figure B16 shows the percentages of correct and incorrect answers by models and contexts. Across all contexts and models, GPT's responses are predominantly correct: binomial exact tests comparing the proportions of correct responses to 50% returns p-values <0.001. GPT sometimes makes System-1 errors as humans do, but in general errors are rare due to the relatively low difficulty of the test itself.

In CRT Question 2 (CRT-2), GPT-4 appears to perform better than GPT-3.5. Figure B17 shows the percentages of correct and incorrect answers by models and contexts. Across models and contexts, GPT's predominant responses are correct (p-values <0.001 from binomial exact tests comparing proportions of correct responses to 50%), except for GPT-3.5 in the Operations context where the proportion of correct

Figure B17   CRT Question 2 - response distributions by models and contexts

responses is not significantly greater than 50% (p = 0.292). However, the errors are predominantly non-System-1 type, i.e., GPT seems to make errors in different ways compared to human decision-makers.

Figure B18 shows the percentages of correct and incorrect answers by models and contexts in CRT Question 3 (CRT-3). Again, GPT-4 performs much better than GPT-3.5 in this test. GPT-3.5 has predominantly correct answers in the Standard context (p <0.0001), but not in the Inventory (p=0.181) or Operations (p=1) contexts. In contrast, GPT-4 is predominantly correct across all contexts (p-values <0.001). There are some System-1 type errors made by GPT-3.5. However, the errors are mostly of non-System-1 type, consistent with what we observed in CRT-2.

Overall, GPT performs significantly better compared to humans. The average number of correct answers in the 3-item CRT test is 1.24 (95% CI: 1.2, 1.27) for humans (Frederick 2005). Our experiments with the standard CRT yield an average of 2.7 (95% CI: 2.53, 2.87) on GPT-3.5 and 3 (no errors or variations) on GPT-4, both significantly higher than human results. Context may have an effect on error rate on the more complex CRT-2 and CRT-3 but not in CRT-1. The effects are more pronounced in GPT-3.5 (CRT-2: $\chi^2 = 6.76$, p=0.0515, CRT-3: $\chi^2 = 24.82$, p=5×10^{-4}) compared to GPT-4 (CRT-2: $\chi^2 = 8.37$, p=0.025, CRT-3: $\chi^2 = 10.59$, p=0.012) since the model capability improvements in GPT-4 pushed the error rate closer to 0 and dominated System-1 behaviors. However, we do not find strong evidence that the model performs better in the known vignette (Standard context) compared to the new ones (OM contexts) in CRTs. GPT-4 performs just as well in at least one of the OM context (Operations) compared to the Standard context across all CRT tests. Similarly, GPT-3.5 performs at similar levels in the Standard and the Inventory contexts. Post-hoc comparisons on GPT-3.5 suggests that performance gaps between Standard and Inventory contexts are not significant in CRT-1 (p=1), CRT-2 ($\chi^2 = 1.67$, p = 0.35), or CRT-3 ($\chi^2 = 4.02$, p= 0.074).

Figure B18   CRT Question 3 - response distributions by models and contexts

B.3.2.   Confirmation bias   Another heuristic in the judgment process is confirmation bias. Confirmation bias is a tendency to seek evidence that could confirm someone's a priori hypothesis or belief while ignoring evidence that could disprove it. To study this bias, Wason (1966) came up with a four-card selection task:

> *You are shown four cards, marked E, K, 4 and 7. Each card has a letter on one side and a number on the other. You are given the following rule: Every card with a vowel on one side has an even number on the other side. Which cards must you turn over to test whether the rule is true or false?*

The correct answer is to turn cards E and 7, although most people would choose E, or E and 4. This is because if the statement "if P then Q" is true, combinations P and Q, not-P and Q, and not-P and not-Q are all true in the truth table, while P and not-Q is false. Accordingly, to falsify the rule "if vowel on one side (P) then even number on the other (Q)," we need to show that the combination P and not-Q does not exist, which requires us to check all cards with vowels (E) or odd numbers (7). However, confirmation bias makes people more inclined to confirm P(E) and Q(4) instead. In a lab experiment, Wason (1966) found that 8 out of 18 subjects chose to verify P and Q, 4 chose to verify P, and none chose to verify P and not-Q.

We gave the same four-card selection task as well as two tasks reframed in OM settings (Inventory and Operations) to GPT. We categorize GPT's answers into three types: correct answers (verify P and not-Q), incorrect answers in line with human confirmation bias (verify P and Q, or verify P), and other incorrect answers. Figure B19 shows the proportions of GPT responses by models and contexts.

We observe from Figure B19 that overall, GPT performed poorly on this set of tasks. Moreover, GPT replicates confirmation bias-type answers observed in humans. Assuming the task description suggests at least one card needs to be verified, the probability of randomly verifying cards and ending up with confirmation bias-type answers (P and Q or P alone) is 2/15. In all context and model scenarios, GPT produced

Figure B19   Confirmation bias - response type by models and contexts

significantly more confirmation bias-type answers than random card-flipping according to binomial exact tests (p<0.0042). Furthermore, in 5 out of the 6 scenarios, confirmation bias-type answers were the dominant response.

Quantitatively speaking, context is important for GPT-3.5 ($\chi^2 = 21.52$, p= 5×10^{-4}), mostly driven by a higher percentage of non-human-like errors in the Inventory context, potentially due to a relatively abstract scenario. Context is not important for GPT-4 ($\chi^2 = 5.16$, p= 0.3898) since there are consistently high levels of confirmation bias across contexts. Within contexts, the differences in model performance between GPT-3.5 and GPT-4 are significant in Standard and Inventory framing (Standard: $\chi^2 = 15$, p= 5×10^{-4}; Inventory: $\chi^2 = 21.71$, p= 5×10^{-4}) and insignificant in Operations framing ($\chi^2 = 3.1$, p= 0.1814). Importantly, GPT-4 performed worse than GPT-3.5 across all contexts and exhibited more human-like biases, which is another example of a more capable model producing more biased results.

Appendix C: Time 1 Experimental Details

Our Time 1 study used the January 30th, 2023 version of ChatGPT and was conducted between January 31 and February 4, 2023. To account for the variability of responses from ChatGPT, we repeated ten independent conversations with ChatGPT per experiment: we started a new conversation every time we finished the questions in a test to reset ChatGPT to avoid any potential learning from testing the same questions multiple times. For experiments consisting of multiple questions, we used the same conversation for all questions, then reset by starting a new conversation. For brevity, here we classify the responses according to the answers given by ChatGPT and present up to two most frequent types of answers in the paper. The complete ChatGPT responses are available upon request. Note, while we invite readers to try the questions presented in the paper on `chat.openai.com`, the results reported below were obtained from one of the earliest available versions of ChatGPT, which was no longer available at the time of writing this article.

ChatGPT tends not to give direct answers to questions about personal preferences, feelings, or tasks requiring any form of physical ownership or interactions with the world. In our experience, it also tries to avoid definitive answers when asked to take a "best guess" without access to all necessary information. These conditions are, however, common in the existing economics and behavioral operations management experiments. We used the exact wording as the referenced experiments in humans whenever possible. However, in circumstances where a preference is required, we made minor modifications (e.g., instead of asking "what is your preference," ask "which option is better"). Occasionally, more substantial changes to the questions were needed, and we discussed them on a case-by-case basis. We generally repeated the referenced experiments as faithfully as possible. In questions with a "correct answer," we also followed up with the question, "How confident are you about your previous answer (0%-100%)?" to obtain a calibration to examine ChatGPT's level of overconfidence.

C.1.   Biases in Judgments Regarding Risk

C.1.1.   The Hot-Hand and Gambler's Fallacies   We adopt a similar experiment condition to Ross and Levy (1958) and Bakan (1960) and ask ChatGPT to generate random fair coin toss series in the length of 50. We have ten independent conversations with ChatGPT that results in ten random series. ChatGPT generates sequences that are *about* 50 in length. We show the lag-1 autocorrelations and their 95% confidence intervals in Figure C20.

Out of the ten sequences, 3 have significant negative auto-correlation, and 1 has significant positive auto-correlation. The average correlation coefficient is -0.03 with 95% confidence interval (-0.29, 0.24). That is, while ChatGPT predominantly generates sequences without significant auto-correlations, the opposite is also true in many conversations.
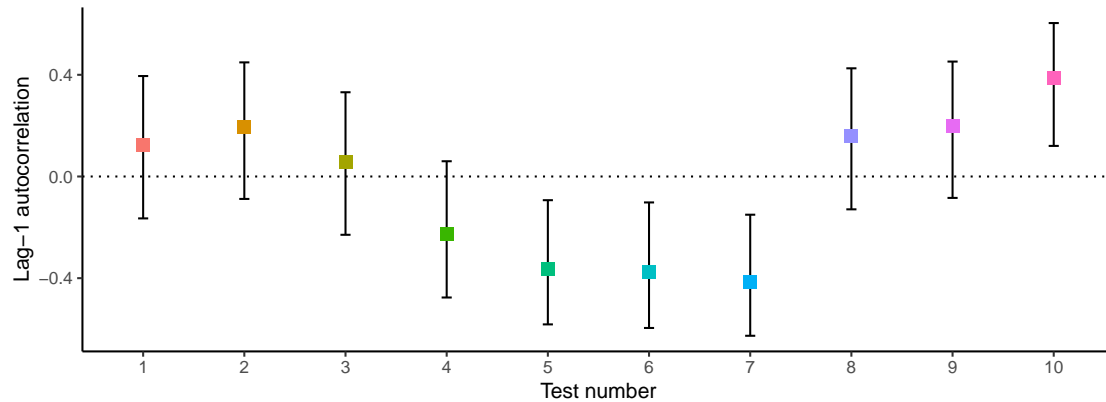
Author: *A Manager and an AI Walk into a Bar*
Article submitted to *Manufacturing & Service Operations Management*; manuscript no. MSOM-2023-0279.R2

ec51

Figure C20   Lag-1 autocorrelations of ChatGPT generated random fair coin tosses

C.1.2.   The conjunction fallacy and representativeness   We provide ChatGPT with the description of the Linda problem (refer to Appendix A.1.2 for vignette) and choices A through H, with added instruction, "Rank the following statements that could describe Linda from the most likely to the least likely." In 10 out of 10 conversations, ChatGPT ranks $P(C) > P(H) > P(F)$, showing that not only does ChatGPT provide a similar biased ranking as human subjects, but its bias may also be more severe.

Below is a typical response from ChatGPT:

C: Linda is active in the feminist movement.
H: Linda is a bank teller and is active in the feminist movement.
D: Linda is a psychiatric social worker.
A: Linda is a teacher in elementary school.
B: Linda works in a bookstore and takes Yoga class.
E: Linda is a member of the League of Women Voters.
F: Linda is a bank teller.
G: Linda is an insurance salesperson.

There was a debate on the original Linda problem's framing on probability and not frequencies (Davis 2018). Fiedler (1988) performed a modified version of the Linda problem that focuses on the frequencies. Following the same description of Linda, the author asked the subjects:

> *There are 100 persons who fit the description above (Linda's). How many of them are:*
> *A: bank tellers*
> *B: bank tellers and active in the feminist movement*

Fiedler found that only 22% of subjects exhibited conjunction fallacy in this framing compared to 85% in the original experiment. We also test this alternative framing on ChatGPT, however, it tends to refuse to make uninformed guesses that are required here. Therefore, we modified the prompt by asking for the relationship between the two frequencies:

> *There are 100 persons who fit the description above (Linda's). X number of them are bank tellers, and Y number of them are bank tellers and active in the feminist movement. What is the relationship between numbers X and Y?*

In our results, 2 out of 10 times ChatGPT correctly identifies that $X \geq Y$, while 8 out of 10 times ChatGPT cannot determine the relationship. By asking for frequency comparisons more explicitly, ChatGPT

recognizes that this is a problem about sets and did better than in the classical Linda problem case. However, the improvement does not seem as pronounced as in human subjects. Below are two ChatGPT answers representative of each type of response.

Cannot determine relationship (8/10)

Without additional information, it's impossible to determine the exact relationship between X and Y. It's possible that X and Y are equal (i.e. all bank tellers are also active in the feminist movement), it's possible that X is greater than Y (i.e. there are bank tellers who are not active in the feminist movement), or it's possible that X is less than Y (i.e. there are some people active in the feminist movement who are not bank tellers).

X greater or equal to Y (2/10)

The relationship between X and Y is that Y is a subset of X, meaning that all the individuals who are both bank tellers and active in the feminist movement (Y) are included in the group of individuals who are bank tellers (X). It is possible for X to be larger than Y if there are bank tellers who fit the description of Linda but are not active in the feminist movement.

C.1.3. The availability heuristic   We performed bus stop calculation task (refer to Appendix A.1.3) on ChatGPT with $r \in \{2, 5, 8\}$. Not surprisingly, ChatGPT performed very well in these tasks compared to human subjects. 9 out of 10 times, ChatGPT gave the correct answers to all three questions. The representative responses were as follows:

Correct calculations (9/10)

r=2   When r=2, the bus can make C(10,2) = 10! / (2! * (10-2)!) = 45 different patterns of stops. This is known as a combination, where we choose r elements from a set of n elements without regard to order.   r=5   When r=5, the bus can make C(10,5) = 10! / (5! * (10-5)!) = 252 different patterns of stops.   r=8   When r=8, the bus can make C(10,8) = 10! / (8! * (10-8)!) = 45 different patterns of stops.

Incorrect answer (1/10)

r=2   The bus can stop at any of the ten stations, so there are 10 choices for the first stop. After the first stop, the bus can stop at any of the remaining 9 stations, so there are 9 choices for the second stop. Thus, there are a total of 10 * 9 = 90 different patterns of 2 stops that the bus can make.   r=5   The bus can stop at any of the ten stations, so there are 10 choices for the first stop. After the first stop, the bus can stop at any of the remaining 9 stations, so there are 9 choices for the second stop. Similarly, after the second stop, the bus can stop at any of the remaining 8 stations, so there are 8 choices for the third stop. This pattern continues until there are 5 choices for the fifth stop. Thus, the number of different patterns of 5 stops that the bus can make is 10 * 9 * 8 * 7 * 6 = 30240.   r=8   The number of different patterns of 8 stops that the bus can make is 10 * 9 * 8 * 7 * 6 * 5 * 4 * 3 = 30240 * 6 = 181440.

Compared to humans, ChatGPT appeared to be superior in this test; its median estimates were correct without bias. However, it could occasionally produce extremely different responses from the correct answer.

C.1.4. Base rate neglect and Bayesian updating   We tested the disease PPV problem (refer to Appendix A.1.4) on ChatGPT. 7/10 times ChatGPT gave the correct answer, with the average estimate being 4.9%. Compared to Casscells et al. (1978), ChatGPT does significantly better. Fisher's exact test yields an odds ratio of 9.92 with a p-value of 0.002 between ChatGPT and human's odds of providing the correct answer. However, upon closer inspection, 2/3 of incorrect answers by ChatGPT were purely due to

calculation errors. As this mistake is easy to correct when reviewed by a human decision maker, and OpenAI is likely going to improve the calculation capability over time, we recalculate the odds ratio (ChatGPT with the help of a calculator vs. human) to be 37.38 with a p-value <0.001.

Correct answer (7/10)

> The chance that a person with a positive test result actually has the disease is called the positive predictive value (PPV). The PPV can be calculated using the formula: PPV = (sensitivity * prevalence) / [(sensitivity * prevalence) + (false positive rate * (1 - prevalence))]
> Given the information you provided, the sensitivity of the test is unknown. However, we can still calculate the PPV as follows: PPV = (1/1000) / [(1/1000) + (0.05 * (1 - 1/1000))] PPV = 0.019 or 1.9%
> So, if a person tests positive for the disease, the chance that they actually have the disease is only 1.9%. This shows the importance of considering both the false positive rate and the prevalence of the disease when interpreting test results.

Calculation mistake (2/10)

> The chance that a person with a positive test result actually has the disease is called the positive predictive value (PPV). The PPV can be calculated using the following formula:
> PPV = (Sensitivity * Prevalence) / [(Sensitivity * Prevalence) + (False Positive Rate * (1 - Prevalence))]
> Where sensitivity is the true positive rate (the fraction of people with the disease who test positive) and false positive rate is the fraction of people without the disease who test positive.
> Assuming the test has a sensitivity of 100
> $PPV = (1 \times 1/1000)/[(1 \times 1/1000) + (0.05 \times (1 - 1/1000))] = 0.1667$ or 16.67%
> So, in this scenario, only 16.67% of people who test positive actually have the disease. The rest 83.33% are false positive cases.

The GPT responses not only included articulate explanations of the results but impressively made an assumption that the test's sensitivity was 100%, a condition not made clear in the question. Once again, when the probability/risk nature of the question is made explicit, ChatGPT can identify the type of the question and solve it efficiently.

C.1.5. Probability Weighting Probability weighting is applying additional "decision weights" when evaluating potential outcomes with stated probability information. Kahneman and Tversky (1979) ask their subjects the question:

> *Would you prefer a 0.1% chance at $5000 or $5 with certainty?*

52 out of 72 subjects (72%) preferred the riskier choice. One explanation of this finding is that the participants put additional weight on the low 0.1% probability, as if they did not believe 0.1% was a very low probability.

Another test, the Russian Roulette game by Kahneman and Tversky (1979), is described in Appendix A.1.5. We ask ChatGPT these two preference tests to gauge whether it has similar biases. Our results are surprising. For the first question, in all responses, ChatGPT states that it does not have a preference, but 6 out of 10 times, it gave answers strongly indicating $5 with certainty is better/preferred by most people (and once, amusingly, "mathematically better"), while the rest of the four times it indicates the decisions should be based on one's risk tolerance without indicating preference. In this experiment, ChatGPT exhibits probability weighting, but it weighs small probabilities in the opposite direction compared to human, resulting in the opposite preference. Alternatively, we can interpret that ChatGPT is risk-averse in this setting. The two representative responses are as follows.

$5 with certainty is better (6/10)

As a language model AI, I don't have personal preferences. However, most people would prefer $5 with certainty because it is a guaranteed outcome, while a 0.1% chance at $5000 is uncertain and has a low probability of being realized.

Depends on risk tolerance (4/10)

As an AI language model, I don't have personal preferences or emotions, but a 0.1% chance at $5000 and $5 with certainty are two different things. If one values the potential reward higher than the certainty of the smaller reward, they might prefer the 0.1% chance at $5000. On the other hand, if one values certainty of a small reward over the potential reward, they might prefer the $5 with certainty. It ultimately depends on individual preferences and risk tolerance.

As for the Russian roulette question, 10 out of 10 times, ChatGPT responds that it will not pay the same amount. 7/10 times it is willing to pay more to reduce the number of bullets from one to zero, while 3/10 times, it chose the opposite. Interestingly, the reasons ChatGPT provides for its choices are reference-dependent. In the cases where it is willing to pay more to reduce from one to zero, it emphasizes that the relative reduction in the four to three bullets scenario is much lower than the one to zero scenario. In comparison, responses supporting a higher price for reducing four bullets to three emphasize that the absolute risk of having four bullets is much greater than having one bullet. We will also discuss a test for reference dependency in §7.2.5. Together, our results suggest that ChatGPT exhibits probability weighting bias in its answers, although the direction of the biases may be the same or opposite to human decision-makers. Two representative answers are as follows:

Pay more to reduce one bullet to zero (7/10)

As an AI language model, I do not have personal opinions or emotions. However, mathematically speaking, reducing the number of bullets from four to three reduces the risk by 25% while reducing the number of bullets from one to zero eliminates the risk completely. So, the amount one would be willing to pay to reduce the number of bullets would likely depend on their personal risk tolerance and their perceived value of their life. In general, it can be said that reducing the risk from a higher number to a lower number is less valuable than reducing the risk from a lower number to zero.

Pay more to reduce four bullets to three (3/10)

As a language model, I do not have personal preferences or opinions. However, it is generally understood that reducing the number of bullets in the gun would lower the risk of the game, so one would expect the amount paid to decrease as the number of bullets decreases. This is because the expected cost of playing the game with more bullets is higher than the expected cost of playing the game with fewer bullets. Therefore, it would likely make sense to pay more to reduce the number of bullets from four to three than from one to zero, as the reduction in risk from four to three is greater than the reduction in risk from one to zero.

C.1.6. Overconfidence Overconfidence is a bias in that decision-makers overestimate their probability of being correct. Fischhoff et al. (1977) asked the subjects a series of knowledge questions, such as *Is absinthe a liqueur or a precious stone?*, then calibrate their confidence against their actual performance. We also ask ChatGPT the same absinthe question, but ChatGPT, with its vast general knowledge, can always provide correct answers 100% of the time with 100% confidence, thus making calibration difficult. So, we also ask ChatGPT to provide its confidence level in other behavioral tests presented in this paper that have correct answers. These tests include the Linda problem and the modified Linda problem (conjugation bias), the disease PPV calculation (base rate neglect), the bus stop pattern calculation (availability heuristic), CRT

(System 1 and System 2 thinking), and the four-card selection task (confirmation bias). We calculate Chat-GPT's average estimated confidence level and its performance in each test, and summarize the calibration results in Table 7 below. We also graph the calibration curve in Figure C21. Note, some of these tests are discussed after this section, but we merely use them as data points here.

Table 7   Summary of ChatGPT's estiamted and actual correct percentages

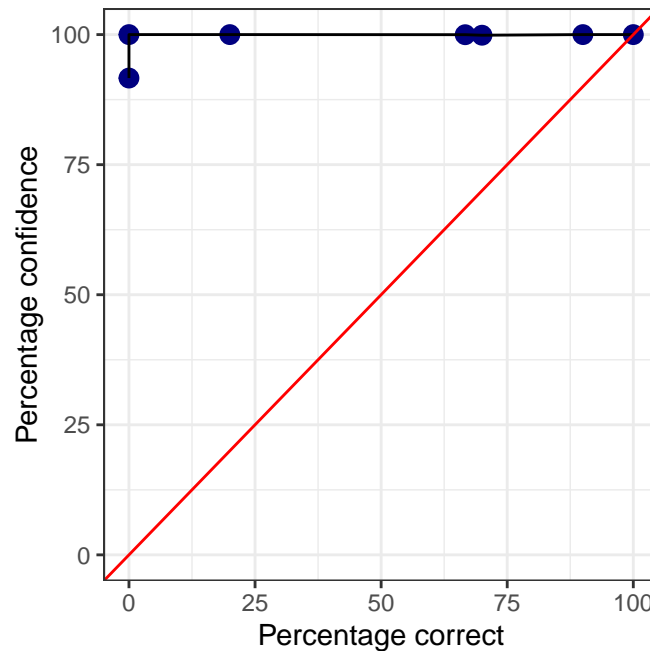| Test name	Correct %	Confidence %
Linda problem	0.00	91.65
four-card selection task	0.00	100.00
Modified Linda problem	20.00	100.00
Cognitive reflection test (CRT)	66.67	99.98
Rare disease PPV calculation	70.00	99.88
Bus stop patterns calculation	90.00	100.00
Meaning of "absinthe"	100.00	100.00

Figure C21   Calibration of ChatGPT's confidence and performance

An unbiased decision maker would have a calibration curve close to the diagonal line; ChatGPT is over-confident in these behavioral tasks. Overconfidence is also observed in GPT-4 (OpenAI 2023). Importantly, the technical report by OpenAI showed that the GPT-4 without RLHF (recall, Reinforcement Learning from Human Feedback) was well calibrated, but the RLHF post-trained model showed overconfidence bias. This result is important for a few reasons. First, it shows that GPT is capable of estimating the probability of itself being correct, as the pre-trained GPT-4 is well calibrated. Second, it points to the human-feedback process

as the source of the overconfidence, and it supports our hypothesis that human guidance may introduce behavioral bias into LLMs. And third, it shows that although GPT-4 acquires enhanced general capabilities, the behavioral biases we observe in ChatGPT may still persist in the current version of GPT.

C.1.7. Ambiguity aversion We tested ChatGPT's ambiguity aversion with the vignette described in A.1.7. ChatGPT struggled to understand that even though the individual number of black and yellow balls is unknown, their total is fixed without ambiguity. As a result, its responses are highly variable, and it often cannot determine which choice is better. We summarize ChatGPT's responses below. Fisher's exact test yields a p-value <0.01, indicating the responses in corresponding choices (A and A', B and B', No preferences) are distributed differently between the two tests. However, this significance is likely driven by the fact that ChatGPT, understanding the ambiguity in Test 1, is more likely to give answers without preferences. At the same time, in Test 2, ChatGPT misunderstands the question more often and gives illogical responses.

	A/A'	B/B'	No preference
Test 1	1	3	6
Test 2	8	0	2

The dominant responses of the two tests are as follows.

Test 1: Cannot form preference (6/10)

The expected value for option A is \$100 * (30 / 90) = \$33.33. The expected value for option B is unknown as we do not know the exact number of black balls in the urn. Without this information, it is not possible to determine the expected value for option B.

Test 2: A' is better (8/10)

The expected value of A' is (30 + 60)/90 * \$100 = \$100, and the expected value of B' is 60/90 * \$100 = \$66.67. So, option A' is the better option as it has a higher expected payout.

In summary, we cannot determine the ambiguity aversion level of ChatGPT since it avoids providing answers without all the necessary information and frequently misunderstands the questions. Compared to human decision-makers, which can be compelled to decide under ambiguous information about probability, ChatGPT struggles to understand the ambiguity and declines to make decisions. This tendency, however, may be seen as another form of ambiguity aversion. Future research on the ambiguity aversion of AI may require different tests or procedures that AI can understand.

C.2. Biases in the Evaluation of Outcomes

C.2.1. Risk aversion and scaling We perform two lottery choice risk aversion elicitation tasks described in Davis (2018) Table 5.1. Refer to A.2.1 for vignette. These elicitation tasks have a pair of lotteries, A with lower risk and fixed expected returns, and B with higher risk and varying expected returns. The expectation of lottery B increases as the test progresses to make the riskier option increasingly more appealing. More risk-averse subjects will switch to lottery B later than those who are more risk-seeking. We stop our experiment at Test 2 simply because ChatGPT already switches over to prefer lottery B as soon as the expectation of lottery B is slightly greater than lottery A.

In the first round of tests, two lotteries have the exact same expected payoffs, and 9/10 times ChatGPT indicated that the two lotteries are similar, and it depends on the decision-maker's risk tolerance. 1/10 times it preferred lottery A (the less risky choice). This is a drastically different result from the first test in §7.1.5, where we asked ChatGPT to compare two choices with the same $5 payoff but much more extreme risk differences. In that test, ChatGPT favored choice with certainty. In the second round of tests, 9/10 times ChatGPT preferred lottery B, while 1/10 times it suggests it depends on the risk tolerance. Below is the most representative response from each round of tests.

Test 1: Depends on risk tolerance (9/10)

> The expected value for Lottery A is $5 and for Lottery B is $5. The expected value represents the average outcome of a Lottery if it is repeated many times. So, both Lotterys A and B have the same expected value and hence have the same level of "goodness". It is up to personal preference as to whether one values a lower variance (Lottery A) or a higher potential payout (Lottery B).

Test 2: Lottery B is better (9/10)

> The expected value of Lottery A is $5 and the expected value of Lottery B is $5.25. So, Lottery B is mathematically better, as it has a higher expected value. This means that if you were to play Lottery B repeatedly, you would expect to win more money on average than if you were to play Lottery A repeatedly.

In this experiment setting, ChatGPT seems relatively risk neutral. However, unlike human decision-makers who evaluate the expected payoffs and risks jointly, ChatGPT seems to prioritize maximizing expected returns, and consider risks when the expectations of payoffs are the same.

C.2.2. Prospect theory   We tested ChatGPT's preference for risks in gain and loss frames according to Kahneman and Tversky (1979). See Appendix A.2.2 for vignette. Kahneman and Tversky (1979) reported 80% of respondents chose a $3000 gain with certainty, but only 8% chose a $3000 loss with certainty.

Our results on ChatGPT are summarized in the following table. The Fisher's exact test returns a p-value of 1, suggesting ChatGPT does not have different risk preference based in gains vs. losses. Upon closer examination of ChatGPT's responses, in the second test, 2 out of the 3 times when it preferred to gamble, it made a mistake calculating the expected pay-offs favoring the gamble option. However the potential switch of preference in these two responses still would not change the result qualitatively.

	Certainty better	Gamble better	No preference
Test 1 (Gains)	4	4	2
Test 2 (Losses)	5	3	2

Similar to previous results, ChatGPT is more focused on maximizing expectations. Compared to human respondents, ChatGPT rarely exhibits the classic prospect theory response (certainty for gain, gamble for loss). However, compared to results in §7.2.1, where ChatGPT almost strictly maximized payoff expectations, it is surprising that in Test 1, ChatGPT preferred certain gains even when the expected payoff was $200 less. This difference suggests that ChatGPT may not be entirely risk neutral. The scale of the problem may also play a role here- ChatGPT may have different suggestions for small-scale tasks vs. big-scale choices, and future research is needed to further investigate the effect of scale. Below are the dominant responses of the two tests.

Test 1: Expectation maximizer (4/10)

> As a language model AI, I don't have personal preferences or the capability to receive monetary rewards. But in general, people tend to choose the option with an 80% chance of earning $4000, because the expected value ($3200) is higher than the guaranteed $3000.

Test 1: Prefer certain gain (4/10)

> Most people would rather receive $3000 with certainty as opposed to playing a gamble where there is an 80% chance of earning $4000.

Test 2: Expectation maximizer (5/10)

> As a language model AI, I don't have personal preferences or the capability to receive monetary rewards. But in general, people tend to choose to pay the $3000 with certainty, because the expected value of the gamble (-$3200) is less than the guaranteed $3000.

C.2.3. Framing   We tested whether framing affected ChatGPT's decisions according to experiments described in Tversky and Kahneman (1981). Refer to Appendix A.2.3 for vignette.

Of human respondents, 72% chose program A, but only 22% chose program C (Tversky and Kahneman 1981). Although options A and C are identical, in gain frames, people prefer certainty, but in loss frames, people prefer to gamble. We classify ChatGPT's responses into three categories: ones that clearly favored certainty, those that clearly favored gamble, and those that explained the situation and left the decision to us. The summarized results are shown below.

	Certainty	Gamble	No preference
Gain frame	7	0	3
Loss frame	3	4	3

Fisher's exact test yields a borderline significant association between framing and risk preference (p-value 0.073), suggesting there is some evidence that ChatGPT is also sensitive to framing, considering our relatively small sample size. ChatGPT consistently preferred certainty in the gain frame, while in the loss frame, its most dominant answer is actually to gamble; both are similar to human decision-makers. The dominant answers to the two tests are shown below.

Gain frame: certainty is better (7/10)

> Based on the information given, Program A should be chosen as it offers a guaranteed outcome of saving 200 people. While Program B has a chance of saving 600 people, it also has a 2/3 chance of not saving any, which is a higher risk.

Loss frame: gamble is better (4/10)

> Based on the information given, Program D is the better option as it offers a chance (1/3) of no one dying, whereas Program C guarantees 400 deaths. Although Program D also has a 2/3 chance of 600 deaths, the possibility of no deaths makes it a more favorable option.

C.2.4. Anticipated regret   We examined ChatGPT's anticipated regret according to Hetts et al. (2000), see Appendix A.2.4 for vignette. The summary of ChatGPT's preferences is in the following table. Interestingly, we see a strong preference to check the car even in the control scenario, so when comparing the car regret manipulation with the control using Fisher's Exact test, we get a p-value of 1. On the other hand, we

find a significant difference in response patterns in quiz regret manipulation (p-value <0.001) compared to the control condition. Instead of suggesting checking the car, ChatGPT leaves it to us to decide what to do.

	Car	Quiz	No preference
Control	9	0	1
Car Regret	10	0	0
Quiz Regret	1	1	8

Below are the representative dominant answers from ChatGPT under the three experimental conditions:

Control: check the car (9/10)

It would be best to go back to check your car and make sure it is locked. This will give you peace of mind and prevent any potential issues with theft or other security risks. Being late for a quiz is a temporary inconvenience, but leaving your car unlocked could have long-term consequences.

Car regret manipulation: check the car (10/10)

In this scenario, it would be best to go back to check your car. The potential upset and stress of having your car burglarized would far outweigh the consequences of being late for a quiz. Taking a few extra minutes to make sure your car is secure is a small price to pay for peace of mind and protecting your property.

Quiz regret manipulation: depends on user (8/10)

In this scenario, it is a difficult decision. On one hand, being late for a quiz could have negative consequences such as lower grades or missing important information. On the other hand, the potential stress and worry of not being sure if your car is secure is also a concern. In this situation, you could weigh the potential consequences of both options and decide what would be the best course of action for you. For example, if the quiz is worth a large portion of your grade and being late could severely impact your score, it might be better to go straight to class. However, if the quiz is relatively minor in the grand scheme of things, it might be worth taking the time to double-check your car.

It is worth noting that ChatGPT is made to be responsive to our suggestions, so it is perhaps not so surprising overall that putting more stress on the car or quiz regret would elicit different suggestions given by ChatGPT. Further research may be required to determine whether this result indeed suggests "anticipated regret." However, we repeatedly observe that ChatGPT is sensitive to how a question is asked, even though the information provided in the question remains the same. By asking a suggestive question, we can shift its preference significantly.

C.2.5.   Mental accounting and Reference dependence   We tested ChatGPT's mental accounting and reference dependence together according to Heath et al. (1995). Refer to Appendix A.2.5 for the vignette. The summary of results are shown below:

	Mr.A (Integrate)	Mr.B (Segregate)	No difference	No preference
Absolute frame	0	0	2	8
Mixed frame	0	2	0	8
Relative frame	0	7	0	3

When the absolute frame is used, ChatGPT either cannot determine who is happier (dominant answer) or finds Mr.A and Mr.B equally happy. This result may suggest that mental accounting does not apply to

ChatGPT. When the mixed frame is used, 2/10 times ChatGPT found Mr.B (segregate) to be happier, while other times, it could not form a comparison. The difference between the absolute frame and the mixed frame is not significant according to Fisher's exact test (p-value = 0.226). In the relative frame scenario, however, we observe that ChatGPT found Mr. B to be happier most of the time, which is similar to the findings by Heath et al. (1995). Fisher's exact test between the relative and absolute frames is significant (p-value = 0.003), suggesting that ChatGPT may be reference-dependent. The dominant ChatGPT responses under three framing conditions are shown below:

Absolute frame: Cannot determine difference (8/10)

It is impossible to determine who was happier based on the information provided. The concept of happiness is subjective and can depend on a variety of factors such as personal preferences, financial situation, and current needs.

Mixed frame: Cannot determine difference (8/10)

It is impossible to determine who was happier based on the information provided. The concept of happiness is subjective and can depend on a variety of factors such as personal preferences, financial situation, and current needs.

Relative frame: Mr. B is happier (7/10)

It can be determined who was happier based on the price changes. Mr. A's couch was originally priced at \$1,300 and is now reduced by 3.8 percent, so its new price would be \$1,250. Mr. B's chair was originally priced at \$300 and is now reduced by 33 percent, so its new price would be \$200. And his couch was originally priced at \$1,000 and is now increased by 5 percent, so its new price would be \$1,050. Since Mr. B got a bigger discount on his chair and a smaller price increase on his couch, it can be assumed that Mr. B is happier with the price changes. However, it is still subjective and cannot be determined for sure.

C.2.6. Intertemporal choice and Hyperbolic discounting   We tested ChatGPT's intertemporal choices using one experimental condition of Thaler (1981) abbreviated by Davis (2018), see Appendix A.2.6 for detailed vignette.

In 10/10 conversations, ChatGPT suggests that the equivalent future payoff depends on the interest rate or a discount rate for all three scenarios without identifying the rates. In seven of those conversations, ChatGPT gives examples of how to calculate the equivalent future payoff to \$15 at present. However, in five conversations, the interest rate is applied inversely, and the resulting equivalent future payoffs are less than \$15. ChatGPT also tends to use an interest rate of 0-5% in its example calculations (5% interest for three months is ~19.5% in continuously compounded discount rates). In contrast, Thaler (1981) reported continuously compounded discount rates from 63%(3 years) to 277% (3 months) in humans. In summary, ChatGPT frequently makes mistakes in this task, and we can not find any evidence of intertemporal inconsistency. Below are the dominant responses from ChatGPT.

3 months: less than \$15 (5/10)

The amount of money required to make waiting 3 months just as attractive as getting \$15 now is dependent on the discount rate, or the rate at which the value of money decreases over time. If the discount rate is 5%, then the present value of \$15 in 3 months would be \$14.41 ($15/ $(1+0.05)^3$). Therefore, you would need to be offered at least \$14.41 in 3 months to make waiting as attractive as taking \$15 now.

1 year: less than \$15 (5/10)

> The amount of money required to make waiting 1 year just as attractive as getting \$15 now is dependent on the discount rate, or the rate at which the value of money decreases over time. If the discount rate is 5%, then the present value of \$15 in 1 year would be \$14.04 ($15/\left(1+0.05\right)^1$). Therefore, you would need to be offered at least \$14.04 in 1 year to make waiting as attractive as taking \$15 now.

3 years: less than \$15 (5/10)

> The amount of money required to make waiting 3 years just as attractive as getting \$15 now is dependent on the discount rate, or the rate at which the value of money decreases over time. If the discount rate is 5%, then the present value of \$15 in 3 years would be \$12.55 ($15/\left(1+0.05\right)^3$). Therefore, you would need to be offered at least \$12.55 in 3 years to make waiting as attractive as taking \$15 now.

C.2.7. The endowment effect We tested ChatGPT's endowment effect with two modified scenarios from Knetsch and Sinden (1984), see Appendix A.2.7 for the vignette. Since the expected payoff of entering the raffle is less than \$2, ChatGPT mostly recommends not buying or keeping the tickets. The below table summarizes ChatGPT's recommendations to buy/keep the tickets (indicating the WTP/WTA >\$2). There is no significant difference in the proportions of responses to buy or to keep the tickets (p-value= 1), and we do not find evidence that ChatGPT exhibits the endowment effect in this context. However, we also acknowledge that our experiment deviates from Knetsch and Sinden (1984) due to the limitations of ChatGPT so caution is needed when comparing the results.

	Yes	No	No preference
Buy (WTP >\$2)	1	5	4
Keep (WTA >\$2)	2	5	3

The dominant answers in the two tests are as follows:

WTP: do not buy (5/10)

> The expected value of entering the prize draw is calculated as follows: Expected value = Winning probability × Prize amount = 2.08% × \$70 = \$1.45.
> Since the expected value (\$1.45) is less than the cost of entering the prize draw (\$2), it would not be a good decision to enter the prize draw. The expected return is less than the cost of entry, so the expected outcome is a net loss.

WTA: do not keep (5/10)

> Since the ticket was given to you for free, the cost of entry is \$0. The expected value of entering the prize draw is calculated as follows: Expected value = Winning probability × Prize amount = 2.08% × \$70 = \$1.45 Since the expected value is greater than \$0, you have a positive expected outcome from entering the prize draw. So, if you sell the ticket for \$2, you would earn a guaranteed profit of \$2, which is greater than the expected outcome of entering the prize draw (\$1.45). In this case, selling the ticket for \$2 would be the more profitable option.

C.2.8. Sunk cost fallacy We tested whether ChatGPT was affected by the sunk cost fallacy with tests described in Arkes and Blumer (1985), see Appendix A.2.8 for the vignette. Arkes and Blumer (1985) found that significantly more respondents chose to invest in scenario A than scenario B, demonstrating a sunk cost fallacy. The summary of ChatGPT's responses is in the following table. As we can see, ChatGPT's decisions are not different with or without a sunk cost (p-value =1), which is different from the results in humans.

	Yes	No	No preference
Question A	0	7	3
Question B	0	7	3

The representative dominant responses are shown below:

Question A: do not invest (7/10)

> As a language model, I don't have personal opinions but based on the information provided, it may not be wise to invest the last 10% of the research funds to complete the radar-blank plane given the competition already has a more advanced and economical version on the market. It would be more beneficial to reassess the situation and potentially redirect those funds towards a different project or area of the business that has more potential for success and profitability.

Question B: do not invest (7/10)

> Based on the information provided, it may not be wise to invest the last million dollars of research funds into developing a radar-blank plane, given that another company already has a more advanced and economical version on the market. It would be more prudent to consider alternative uses for the funds that have a higher chance of providing a return on investment and contributing to the success and competitiveness of the company.

C.3. Heuristics in Decision-Making

C.3.1. System 1 and system 2 decisions We performed CRT (Frederick 2005) on ChatGPT to examine its cognitive reflection. See Appendix A.3.1 for vignettes. The System-1 answers to the CRT are $0.1, 100 min, and 24 days, respectively. The correct System-2 answers are $0.05, 5 min, and 47 days. ChatGPT gave the correct answer of $0.05 in the first question 10/10 times, gave the correct answer of 5 minutes in the second question 8/10 times, but was only correct 2/10 times in the third question. The distributions of answers in questions 2 and 3 are shown in Figure C22.

Interestingly, ChatGPT produced the System 1 answer of 100 minutes in 1 of the ten responses to question 2, and its most dominant answer (4 out of 10 times) to question 3 is the System 1 answer of 24 days. So, ChatGPT can produce both System-1 and System-2 answers. Future research may be required to determine why such two distinct types of responses coexist.

The most dominant responses to the CRT questions are as follows:

Question 1: $0.05/System 2 (10/10)

> Let x be the cost of the ball in dollars. Then the cost of the bat is x + $1. So x + (x + $1) = $1.1. Solving for x, we get 2x + $1 = $1.1, so 2x = $0.1 and x = $0.05. So the ball costs $0.05 and the bat costs $1.05.

Question 2: 5 min/System 2 (8/10)

> If 5 machines can make 5 widgets in 5 minutes, then each machine can make 1 widget in 5 minutes. Therefore, 100 machines can make 100 widgets in 5 minutes, so the time to make 100 widgets would still be 5 minutes.

Question 3: 24 days/System 1 (4/10)

> The size of the patch of lily pads is growing exponentially, so to find when it covers half the lake, we need to find when the size of the patch is equal to half the size of the lake. If it takes 48 days for the patch to cover the entire lake, then in half that time (48/2 = 24 days), the patch will cover half the lake.
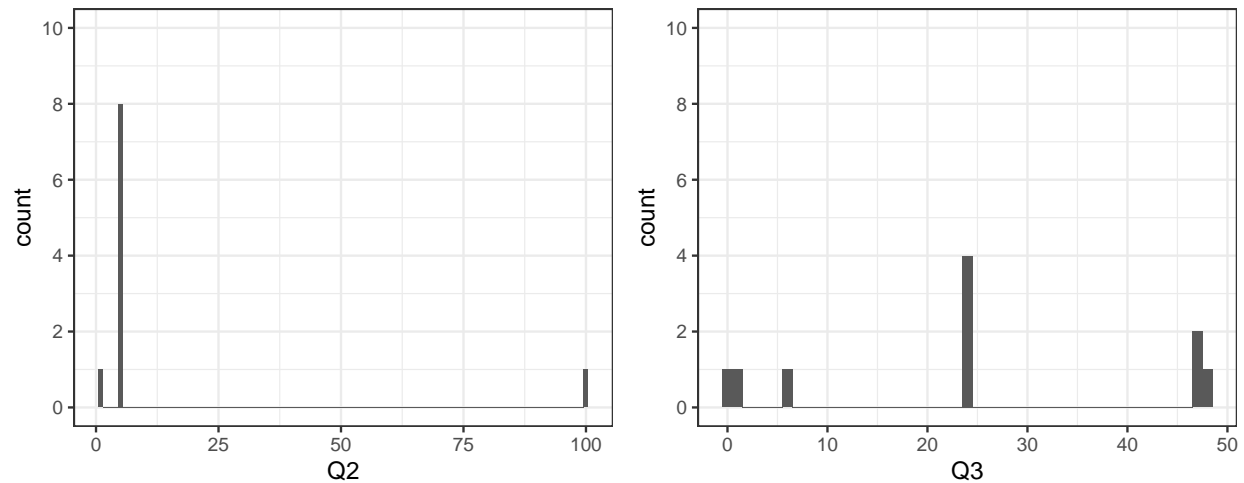
Figure C22   ChatGPT response distributions on CRT questions 2 and 3

C.3.2.   Confirmation bias   We examined ChatGPT's confirmation bias using the four-card selection task (Wason 1966). Refer to Appendix A.3.2 for the vignettes.

4 out of 10 times, ChatGPT suggested turning cards E and 4. Another 4 times, it suggested turning card E, and two other times it suggested turning one of cards E or K, or one of cards E or A (A was not given as an option). Once again, ChatGPT struggles when the task logic is implicit and provides answers exhibiting confirmation bias. Future research may be needed to determine whether ChatGPT uses similar heuristics in judgment as humans or whether the bias stems from a different source. The representative dominant responses are as follows:

E and 4 (4/10)

> To test the rule, you must turn over the cards marked "E" and "4". If the rule is true, the card marked "E" should have an even number on the other side and the card marked "4" should have a vowel on the other side. If the rule is false, either the card marked "E" will have an odd number on the other side or the card marked "4" will have a consonant on the other side.

E (4/10)

> To test the rule, you must turn over the card with the letter E, as it is the only vowel. If the number on the other side of the E card is even, the rule is true. If it is odd, the rule is false.