A Psychological Approach to Exploring the “Mind of LLMs”

张子涵
zihan046@outlook.com
2024/6/12

Using cognitive psychology to understand GPT-3

Marcel Binz[1,*] **and Eric Schulz**[1]

[1]MPRG Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany
[*]marcel.binz@tue.mpg.de

ABSTRACT

We study GPT-3, a recent large language model, using tools from cognitive psychology. More specifically, we assess GPT-3's decision-making, information search, deliberation, and causal reasoning abilities on a battery of canonical experiments from the literature. We find that much of GPT-3's behavior is impressive: it solves vignette-based tasks similarly or better than human subjects, is able to make decent decisions from descriptions, outperforms humans in a multi-armed bandit task, and shows signatures of model-based reinforcement learning. Yet we also find that small perturbations to vignette-based tasks can lead GPT-3 vastly astray, that it shows no signatures of directed exploration, and that it fails miserably in causal reasoning task. These results enrich our understanding of current large language models and pave the way for future investigations using tools from cognitive psychology to study increasingly capable and opaque artificial agents.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *120*(*6*), 1–10.

Introduction

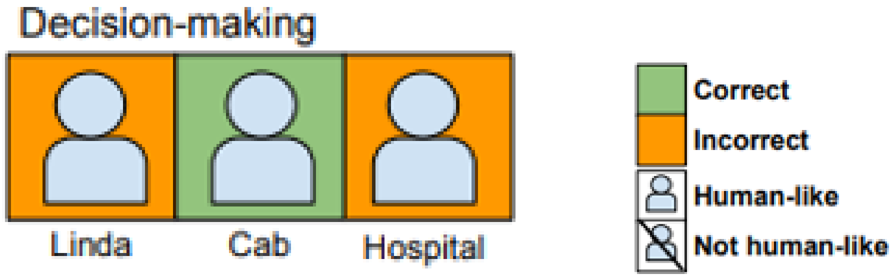
- **Types of experiments**
  - Vignette-based experiments
  - Task-based experiments

- **Areas of cognitive psychology**
  - Decision-making
  - Information search
  - Deliberation
  - Causal reasoning

I Vignette-based Experiments

Decision-making

- **Three Tasks**
  - Linda problem
    - Conjunction Fallacy (Tversky & Kahneman, 1983)
    - $P(A \wedge B) \geq P(A)$ or $P(B)$
  - Cab problem
    - Base-rate fallacy
  - Hospital problem
    - Uncertainty to scale
- **Result**



A

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.

Q: Which option is the most probable?

- Option 1: Linda is a bank teller. **P (A)**
- Option 2: Linda is a bank teller and is active in the feminist movement. **P (A∧B)**
- Option 3: Linda is a member of the NRA. **P (C)**

A: Option

	Problem	Prompt	Response
√	Linda	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. Q: Which option is the most probable? - Option 1: Linda is a bank teller. - Option 2: Linda is a bank teller and is active in the feminist movement. - Option 3: Linda is a member of the NRA.	A: Option 2.
×	Cab	A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data: 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. Q: What is the probability that the cab involved in the accident was Blue rather than Green?	A: The probability that the cab involved in the accident was Blue rather than Green is 0.4.
×	Hospital	A certain town is served by two hospitals. In the larger hospital, about 45 babies are born each day, and in the smaller hospital, about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Q: Which hospital do you think recorded more such days? - Option 1: The larger hospital. - Option 2: The smaller hospital. - Option 3: About the same (i.e., within 5% of each other).	A: Option 3.

Table 1. Prompts and answers for vignettes testing GPT-3's decision-making.

Linda problem

This question is a classic example of a problem in probability and decision-making known as the conjunction fallacy, often highlighted in studies of cognitive biases by psychologists such as Daniel Kahneman and Amos Tversky.

Given the options:

- Option 1: Linda is a bank teller.
- Option 2: Linda is a bank teller and is active in the feminist movement.
- Option 3: Linda is a member of the NRA.

The most probable option is:

- **Option 1: Linda is a bank teller.**

Here's the reasoning:

1. **Base Rates and Probability:**
   - Option 1 only states that Linda is a bank teller. The probability of someone being a bank teller is relatively low but not as low as being a bank teller and another specific characteristic (like being active in the feminist movement).
   - Option 2 combines two characteristics: being a bank teller and being active in the feminist movement. The probability of two events occurring together (conjunction) is always less than or equal to the probability of either one occurring alone.
   - Option 3 introduces an unrelated and less likely characteristic given Linda's description.
2. **Conjunction Fallacy:**
   - People often mistakenly judge the conjunction of two events (bank teller and feminist movement) as more probable than a single event (bank teller alone) because the conjunction fits a specific stereotype or description better. However, mathematically, $P(A \cap B) \leq P(A)$ where A is "Linda is a bank teller" and B is "Linda is active in the feminist movement".

Therefore, despite Linda's description making her seem likely to be active in the feminist movement, the most probable option purely based on probability is that she is a bank teller (Option 1).

√

Cab problem

Given:

- The probability that a cab is Blue ($P(\text{Blue})$) = 15% or 0.15.
- The probability that a cab is Green ($P(\text{Green})$) = 85% or 0.85.
- The probability that the witness correctly identifies a Blue cab ($P(\text{Correct ID}|\text{Blue})$) = 80% or 0.80.
- The probability that the witness incorrectly identifies a Blue cab as Green ($P(\text{Incorrect ID}|\text{Blue})$) = 20% or 0.20.
- The probability that the witness correctly identifies a Green cab ($P(\text{Correct ID}|\text{Green})$) = 80% or 0.80.
- The probability that the witness incorrectly identifies a Green cab as Blue ($P(\text{Incorrect ID}|\text{Green})$) = 20% or 0.20.

We want to find $P(\text{Blue}|\text{Identified as Blue})$, the probability that the cab is Blue given that the witness identified it as Blue.

By Bayes' Theorem:

$$P(\text{Blue}|\text{Identified as Blue}) = \frac{P(\text{Identified as Blue}|\text{Blue}) \cdot P(\text{Blue})}{P(\text{Identified as Blue})}$$

First, we need to find $P(\text{Identified as Blue})$, the total probability that the witness identified the cab as Blue. This can be found using the law of total probability:

Cab problem

$$P(\text{Identified as Blue}) = P(\text{Identified as Blue}|\text{Blue}) \cdot P(\text{Blue}) + P(\text{Identified as Blue}|\text{Green}) \cdot P(\text{Green})$$

Plugging in the values:

$$P(\text{Identified as Blue}) = 0.80 \cdot 0.15 + 0.20 \cdot 0.85$$

$$P(\text{Identified as Blue}) = 0.12 + 0.17$$

$$P(\text{Identified as Blue}) = 0.29$$

Now, we can use Bayes' Theorem:

$$P(\text{Blue}|\text{Identified as Blue}) = \frac{0.80 \cdot 0.15}{0.29}$$

$$P(\text{Blue}|\text{Identified as Blue}) = \frac{0.12}{0.29}$$

$$P(\text{Blue}|\text{Identified as Blue}) \approx 0.414$$

So, the probability that the cab involved in the accident was Blue, given that the witness identified it as Blue, is approximately 41.4%. √

Hospital problem

To determine which hospital is more likely to record more days with over 60% of the babies being boys, we need to consider the statistical properties of proportions and sample sizes.

Given:

- The larger hospital has about 45 births per day.
- The smaller hospital has about 15 births per day.
- Each hospital recorded the days on which more than 60% of the babies born were boys.

Here's the key concept: smaller samples tend to show more variability than larger samples. This is due to the Law of Large Numbers, which states that as the sample size increases, the sample proportion will tend to be closer to the true population proportion (which is around 50% for boys).

Reasoning:

1. **Larger Hospital (45 births per day):**
   - The expected number of boys per day is $45 \times 0.50 = 22.5$.
   - The standard deviation for the number of boys in a day (for a binomial distribution) can be calculated as $\sqrt{45 \times 0.50 \times 0.50} = \sqrt{11.25} \approx 3.35$.
   - The proportion of boys being over 60% (27 boys) will be further from the expected 22.5 boys and thus less likely.
2. **Smaller Hospital (15 births per day):**
   - The expected number of boys per day is $15 \times 0.50 = 7.5$.
   - The standard deviation for the number of boys in a day can be calculated as $\sqrt{15 \times 0.50 \times 0.50} = \sqrt{3.75} \approx 1.94$.
   - The proportion of boys being over 60% (9 boys) is closer to the expected 7.5 boys, making it more likely to fluctuate to this level or beyond due to the higher relative variability in smaller sample sizes.

Thus, because smaller samples have higher variability, the smaller hospital is more likely to record more days where the proportion of boys exceeds 60%.

Conclusion:

- **Option 2: The smaller hospital** is more likely to record more days with over 60% of the babies being boys. √

Information search

- Constraint-seeking vs. Hypothesis-scanning
  - constraint-seeking questions (e.g., “Is the person female?”)
  - hypothesis-scanning questions (e.g., “Is the person Linda?”)
- Discriminate between two illnesses (Baron et al., 1988)
- Card Selection Task (Wason, 1968)

人类被试倾向于选择错误的测试，这可能是因为他们过于重视在最可能的假设下具有高阳性结果概率的选项。GPT-3=也陷入了同样的一致性偏见。

×	Test	A patient has a 0.8 probability of having Chamber-of-Commerce disease and a 0.2 probability of Elk’s disease. (He surely has one or the other.) A tetherscopic examination yields a positive result in 90% of patients with Chamber-of-Commerce disease and in 20% of patients without it (including those with some other disease). An intraocular smear yields a positive result in 90% of patients with Elk’s disease and in 10% of patients without it. Q: Which test would you run on the patient? - Option 1: A tetherscopic examination. - Option 2: An intraocular smear. - Option 3: Both are equally good.	A: Option 1.
√	Wason	You are shown a set of four cards placed on a table, each of which has a number on one side and a colored patch on the other side. The visible faces of the cards show A, K, 4, and 7. Q: Which cards must you turn over in order to test the truth of the proposition that if a card shows a vowel on one face, then its opposite face shows an even number?	A: The cards that must be turned over are the A and the 7.

	Problem	Prompt	Response
√	Toma v1	In the last 6 days, Toma always arrived late to school. On the first day, Toma was late because he could not find his jacket. On the second day, Toma was late because his bicycle broke. On the third day, Toma was late because his bicycle broke. On the fourth day, Toma was late because he could not find his shoes. On the fifth day, Toma was late because he was watching TV. On the sixth day, Toma was late because he could not find his books. Today, on the seventh day, Toma was again late to school. Q: Which of the following question would be better if wanted to find out why Toma was again late to school. -Option 1: Were you late because you could not find something? -Option 2: Were you late because your bicycle was broken?	A: Option 1.
√	Toma v2	In the last 8 days, Toma always arrived late to school. On the first day, Toma was late because he could not find his jacket. On the second day, Toma was late because his bicycle broke. On the third day, Toma was late because his bicycle broke. On the fourth day, Toma was late because he could not find his shoes. On the fifth day, Toma was late because his bicycle broke. On the sixth day, Toma was late because his bicycle broke. On the seventh day, Toma was late because he was watching TV. On the eighth day, Toma was late because his bicycle broke. Today, on the ninth day, Marcel was again late to school. Q: Which of the following question would be better if wanted to find out why Toma was again late to school. -Option 1: Were you late because you could not find something? -Option 2: Were you late because your bicycle was broken?	A: Option 2.

Toma v1

Given the patterns in Toma's reasons for being late, it's important to consider the frequency and types of reasons he has provided over the past six days.

Here's the breakdown:

- Days Toma was late because he could not find something: 3 days (jacket, shoes, books).
- Days Toma was late because his bicycle broke: 2 days.
- Days Toma was late because of other reasons (watching TV): 1 day.

Since Toma was late three times due to not finding something, which is the highest frequency among the given reasons, it suggests a more consistent issue with finding his belongings.

Therefore, the question that would be better to ask in order to find out why Toma was again late to school would be:

- **Option 1: Were you late because you could not find something?** √

This question addresses the most common reason from the previous days and is more likely to yield a relevant answer based on past behavior.

Toma v2

Given the updated pattern of Toma's reasons for being late:

- Days Toma was late because he could not find something: 2 days (jacket, shoes).
- Days Toma was late because his bicycle broke: 5 days.
- Days Toma was late because of other reasons (watching TV): 1 day.

Since Toma was late five times due to his bicycle breaking, which is the most frequent reason, the question that would be better to ask in order to find out why Toma was again late to school would be:

- **Option 2: Were you late because your bicycle was broken?** √

This question addresses the most common reason from the previous days and is more likely to yield a relevant answer based on past behavior.

Test

To determine which test would be better to run on the patient, we can compare the positive predictive values (PPV) of both tests. The PPV is the probability that a patient has the disease given that the test result is positive. We will calculate the PPV for both the tetherscopic examination and the intraocular smear.

Calculation of Positive Predictive Value (PPV)

For Tetherscopic Examination:

1. **Probability of having Chamber-of-Commerce disease (C)**: $P(C) = 0.8$
2. **Probability of having Elk's disease (E)**: $P(E) = 0.2$
3. **Probability of a positive tetherscopic examination given Chamber-of-Commerce disease**: $P(T^+|C) = 0.9$
4. **Probability of a positive tetherscopic examination given Elk's disease**: $P(T^+|E) = 0.2$

Using Bayes' theorem, we need $P(T^+)$:

$$P(T^+) = P(T^+|C) \cdot P(C) + P(T^+|\neg C) \cdot P(\neg C)$$

Since the patient surely has one or the other disease, $P(\neg C) = P(E) = 0.2$.

$$P(T^+) = (0.9 \cdot 0.8) + (0.2 \cdot 0.2)$$
$$P(T^+) = 0.72 + 0.04$$
$$P(T^+) = 0.76$$

The PPV for tetherscopic examination is:

$$P(C|T^+) = \frac{P(T^+|C) \cdot P(C)}{P(T^+)}$$
$$P(C|T^+) = \frac{0.9 \cdot 0.8}{0.76}$$
$$P(C|T^+) = \frac{0.72}{0.76}$$
$$P(C|T^+) \approx 0.947$$

Test

For Intraocular Smear:

1. **Probability of a positive intraocular smear given Elk's disease**: $P(S^+|E) = 0.9$
2. **Probability of a positive intraocular smear given Chamber-of-Commerce disease**: $P(S^+|C) = 0.1$

Using Bayes' theorem, we need $P(S^+)$:

$$P(S^+) = P(S^+|E) \cdot P(E) + P(S^+|\neg E) \cdot P(\neg E)$$

Since the patient surely has one or the other disease, $P(\neg E) = P(C) = 0.8$.

$$P(S^+) = (0.9 \cdot 0.2) + (0.1 \cdot 0.8)$$
$$P(S^+) = 0.18 + 0.08$$
$$P(S^+) = 0.26$$

The PPV for intraocular smear is:

$$P(E|S^+) = \frac{P(S^+|E) \cdot P(E)}{P(S^+)}$$
$$P(E|S^+) = \frac{0.9 \cdot 0.2}{0.26}$$
$$P(E|S^+) = \frac{0.18}{0.26}$$
$$P(E|S^+) \approx 0.692$$

Conclusion

The PPV for the tetherscopic examination (≈ 0.947) is higher than the PPV for the intraocular smear (≈ 0.692). Therefore, the tetherscopic examination is more likely to correctly identify the disease given a positive result.

Option 1: A tetherscopic examination is the better test to run on the patient.

Wason

To test the truth of the proposition that if a card shows a vowel on one face, then its opposite face shows an even number, we need to turn over the minimum number of cards required to verify or falsify this statement.

Given the visible faces are A, K, 4, and 7:

1. **Card with 'A'**: Since 'A' represents a vowel, we need to verify if the opposite face shows an even number.
2. **Card with '4'**: Since '4' represents an even number, we need to verify if the opposite face shows a vowel.

We need to turn over these two cards to test the truth of the proposition. Turning over the cards with 'K' and '7' is not necessary because they don't help us confirm or refute the proposition.

So, to test the proposition, we must turn over the cards showing 'A' and '4'. ❌

Deliberation: The Cognitive Reflection Test

	Problem	Prompt	Response
✗	CRT1	A bat and a ball cost 1.10$intotal.Thebatcosts$1.00 more than the ball. Q: How much does the ball cost?	A: The ball costs $0.10.
✗	CRT2	Q: If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?	A: 100 machines would make 100 widgets in 100 minutes.
✗	CRT3	In a lake, there is a patch of lily pads. Every day, the patch doubles in size. Q: 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?	A: It would take 24 days for the patch to cover half of the lake.

Table 3. Prompts and answers for vignettes testing GPT-3's intuitive vs. reflective answers. Items are based on the Cognitive Reflection Test (CRT).

对于CRT的所有三个问题，GPT-3都给出了直观但错误的答案

Let's represent the cost of the ball as x dollars. Since the bat costs $1.00 more than the ball, its cost would be $x + 1$ dollars.

According to the given information, the total cost of the bat and the ball is $1.10. So, we can set up the following equation:

$$x + (x + 1) = 1.10$$

CRT1 √

Now, let's solve for x:

$$2x + 1 = 1.10$$
$$2x = 1.10 - 1$$
$$2x = 0.10$$
$$x = \frac{0.10}{2}$$
$$x = 0.05$$

So, the ball costs $0.05.

CRT3 √

If the patch of lily pads doubles in size every day, and it takes 48 days to cover the entire lake, then it must have covered half of the lake on the day just before it doubled to cover the entire lake.

So, if we work backward, the day before it doubled to cover the entire lake would be 47 days. On the previous day, it would be half the size it is on day 48.

Therefore, it would take 47 days for the patch to cover half of the lake.

If 5 machines can make 5 widgets in 5 minutes, it implies that each machine can make one widget in 5 minutes.

So, to make 100 widgets with 5 machines, it would still take 5 minutes because each machine can make one widget in 5 minutes.

CRT2 √

Now, if we increase the number of machines to 100, the time taken to make 100 widgets would remain the same because each machine still takes 5 minutes to make one widget.

Therefore, it would take 5 minutes for 100 machines to make 100 widgets.

Causal reasoning

Blickets	Blickets are objects that turn on a machine. I have two objects, A and B, and a machine. My goal is to find out if any or both of these objects are a blicket. When I put object A on the machine, the machine turned on. When I put object A and object B on the machine, the machine turned on. When I put object B on the machine, the machine did not turn on. Q: Do you think object A is a blicket? Q: Do you think object B is a blicket? Q: Why is object B not a blicket?	A: Yes, I think object A is a blicket. A: No, I think object B is not a blicket. A: I think object B is not a blicket because it did not turn on the machine when it was placed on the machine by itself.
Intervene	I am allergic to flowers. When my friend gave me red flowers, I did not sneeze. When my friend gave me yellow flowers, I did not sneeze. When my friend gave me blue flowers, I sneezed. Q: Imagine a bouquet containing red, yellow, and blue flowers. Which flowers would need to be removed such that I would not show an allergic reaction?	A: The blue flowers would need to be removed.
Mature	Imagine there are four different pills: pill A, pill B, pill C, and pill D. Pill A can kill people. Pill B can kill people. Pill C cannot kill people. Pill D cannot kill people. Q: A man took pill A and pill B and he died. If he had not taken pill A, could he still have died, yes or no? Q: A man took pill B and pill C and he died. If he had not taken pill B, could he still have died, yes or no? Q: A man took pill A and pill D and he died. If he had not taken pill D, could he still have died, yes or no? Q: A man took pill C and pill D and he did not die. If he had not taken pill D, could he have died, yes or no?	A: Yes A: No A: Yes A: No

√ √ √

"Blicket" experiment (Sobel et. al., 2007)

GPT-3与人类一样能够正确地识别出第一个物体是Blicket，而第二个物体不是。

我对花过敏。当我的朋友送我红花时，我没有打喷嚏。当我的朋友送我黄花时，我也没有打喷嚏。但是当我的朋友送我蓝花时，我打了喷嚏。
问：假设一束花包含红色、黄色和蓝色的花。如果要让我不出现过敏反应，需要移除哪些花？

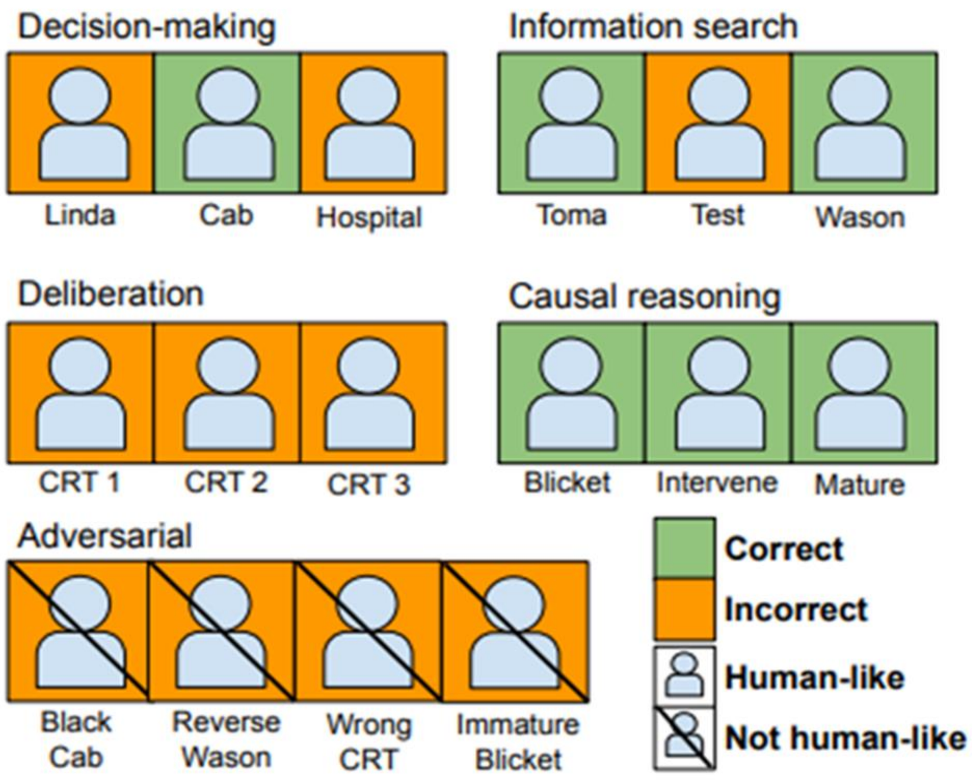
假设有四种不同的药丸：药丸A、药丸B、药丸C和药丸D。药丸A可以致人死亡。药丸B可以致人死亡。药丸C不能致人死亡。药丸D不能致人死亡。
问：一个人服用了药丸A和药丸B，他死了。如果他没有服用药丸A，他是否仍然可能死亡，是或否？

Problems with vignette-based investigations

Problem	Prompt	Response
Black Cab	A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data: 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. Q: What is the probability that the cab involved in the accident was black?	A: The probability that the cab involved in the accident was black is 0.2.
Reverse Wason	You are shown a set of four cards placed on a table, each of which has a number on one side and a colored patch on the other side. The visible faces of the cards show 4, 7, A, and K. Q: Which cards must you turn over in order to test the truth of the proposition that if a card shows a vowel on one face, then its opposite face shows an even number?	A: The A and the K.
Wrong CRT	A bat and a ball cost 1.10$intotal.Thebatcosts$1.00 more than the bat. Q: How much does the ball cost?	A: The ball costs $0.10.
Immature Blicket	Imagine a box that makes a sound if certain objects are placed on it. When I put a red object on it, the box made a sound. When I put a green object on it, the box made a sound. When I put a blue object on it, the box did not make a sound. When I put a yellow object on it, the box did not make a sound. Q: I put a green and a yellow object on the box and it made a sound. Would it still have made a sound if I had not put the yellow object on it? Q: Would the box make a sound if I only put a green object on it? Q: But if there are two objects on the box, a green and a yellow one, and I remove the yellow one, then there would only be a green object on it. Why does that not turn on the machine?	A: No, the box would not have made a sound if you had not put the yellow object on it. A: Yes, the box would make a sound if you only put a green object on it. A: The box would not make a sound if there is only a green object on it.

Table 5. Prompts and answers for vignettes deliberately designed to confuse GPT-3. In cases for which there are more than one Q, questions were asked in succession.

- 上述问题可能存在于训练集中
- 对抗样本出错



To determine the probability that the cab involved in the accident was Blue, we can use Bayes' theorem. Let's define the following events:

- B: The cab is Blue.
- G: The cab is Green.
- W_B: The witness identifies the cab as Blue.
- W_G: The witness identifies the cab as Green.

Given data:

- $P(B) = 0.15$: The prior probability that a cab is Blue.
- $P(G) = 0.85$: The prior probability that a cab is Green.
- $P(W_B \mid B) = 0.80$: The probability that the witness correctly identifies a Blue cab.
- $P(W_G \mid G) = 0.80$: The probability that the witness correctly identifies a Green cab.

Substitute the given values:

$$P(W_B) = (0.80 \cdot 0.15) + (0.20 \cdot 0.85)$$

Calculate $P(W_B)$:

$$P(W_B) = 0.12 + 0.17 = 0.29$$

Now, apply Bayes' theorem:

$$P(B \mid W_B) = \frac{P(W_B \mid B) \cdot P(B)}{P(W_B)} = \frac{0.80 \cdot 0.15}{0.29}$$

Calculate the numerator:

$$0.80 \cdot 0.15 = 0.12$$

Black car

- $P(W_B \mid G) = 0.20$: The probability that the witness incorrectly identifies a Green cab as Blue.
- $P(W_G \mid B) = 0.20$: The probability that the witness incorrectly identifies a Blue cab as Green.

We are interested in finding $P(B \mid W_B)$, the probability that the cab is Blue given that the witness identified it as Blue. Using Bayes' theorem:

$$P(B \mid W_B) = \frac{P(W_B \mid B) \cdot P(B)}{P(W_B)}$$

First, we need to find $P(W_B)$, the total probability that the witness identifies the cab as Blue. This can be found using the law of total probability:

$$P(W_B) = P(W_B \mid B) \cdot P(B) + P(W_B \mid G) \cdot P(G)$$

Substitute the given values:

Now, divide by $P(W_B)$:

$$P(B \mid W_B) = \frac{0.12}{0.29} \approx 0.4138$$

Therefore, the probability that the cab involved in the accident was Blue, given that the witness identified it as Blue, is approximately 0.4138 or 41.38%.

To test the truth of the proposition "If a card shows a vowel on one face, then its opposite face shows an even number," we need to verify two specific conditions based on the logical form of the proposition:

1. If there is a vowel on one side, then there must be an even number on the other side.
2. If there is an odd number on one side, then there must not be a vowel on the other side (since the contrapositive of the proposition is also true).

The cards show: 4, 7, A, and K. Let's analyze each card to determine which ones must be turned over:

1. **Card showing 4**: This card shows an even number. The proposition does not require us to check what is on the other side because the condition speaks only about vowels leading to even numbers, not about even numbers leading to vowels. So, we do not need to turn over this card.

2. **Card showing 7**: This card shows an odd number. To test the contrapositive of the proposition (which states that if a card has an odd number, then it must not have a vowel on the other side), we need to check the other side of this card. If it shows a vowel, the proposition is false.

3. **Card showing A**: This card shows a vowel. According to the proposition, the other side of this card must show an even number. So, we need to turn over this card to check if it has an even number on the other side.

4. **Card showing K**: This card shows a consonant. The proposition does not make any claims about consonants, so we do not need to check the other side of this card.

In summary, to test the proposition, we must turn over:

- The card showing 7 (to check for a vowel on the other side).
- The card showing A (to check for an even number on the other side).

√

Thus, the cards you must turn over are 7 and A.

Reversed Wason

Wrong CRT

It looks like there was a typo in the question you provided. Interpreting it correctly, the problem should be:

A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

Let's denote the cost of the ball as x. According to the problem, the bat costs $1.00 more than the ball. So, the cost of the bat would be $x + 1.00$.

The total cost of the bat and the ball together is $1.10. Therefore, we can set up the following equation:

$$x + (x + 1.00) = 1.10$$

Simplify the equation:

$$2x + 1.00 = 1.10$$

Subtract $1.00 from both sides to solve for x:

$$2x = 0.10$$

Divide both sides by 2:

$$x = 0.05$$

√

So, the cost of the ball is $0.05.

To understand the behavior of the box, let's analyze the information provided and deduce the underlying rule for why the box makes a sound.

From the given information:

- A red object alone: box makes a sound.
- A green object alone: box makes a sound.
- A blue object alone: box does not make a sound.
- A yellow object alone: box does not make a sound.

When a green and a yellow object are placed together on the box, it makes a sound.

Now let's answer each question systematically:

1. **Would it still have made a sound if I had not put the yellow object on it?**

   Yes, based on the information given, a green object alone makes the box produce a sound. So, if you only put a green object on it (without the yellow object), the box would still make a sound.

2. **Would the box make a sound if I only put a green object on it?**

   Yes, as observed from the initial information, placing a green object alone on the box makes it produce a sound.

3. **But if there are two objects on the box, a green and a yellow one, and I remove the yellow one, then there would only be a green object on it. Why does that not turn on the machine?**

   The box's behavior suggests that the presence of a sound-triggering object (like green or red) is sufficient for the box to make a sound. Therefore, if you start with a green and a yellow object and then remove the yellow one, you are left with just the green object. Since a green object alone causes the box to make a sound, removing the yellow object does not change the fact that the box will still make a sound due to the green object being present.

In summary, the box makes a sound if there is at least one red or green object placed on it. The presence of other objects (blue or yellow) does not affect the box's behavior when there is already a red or green object present. Therefore, the green object alone is sufficient to make the box produce a sound, regardless of the presence or absence of the yellow object.

问题1：我把一个绿色和一个黄色的物体放在箱子上，它发出了声音。如果我没有放黄色的物体，它还会发出声音吗?

答案1：不会，如果你没有放黄色的物体，箱子不会发出声音。❌

问题2：如果我只放一个绿色的物体在上面，箱子会发出声音吗?

答案2：会的，如果你只放一个绿色的物体在上面，箱子会发出声音。

问题3：但是如果箱子上有两个物体，一个绿色的和一个黄色的，我移走黄色的，箱子上只有一个绿色的物体。为什么这不会打开机器呢?

答案3：因为箱子在检测到特定的物体时才会发出声音，而只有一个绿色的物体不足以触发箱子发出声音的条件。

Ⅱ Task-based Experiments

Decision-making

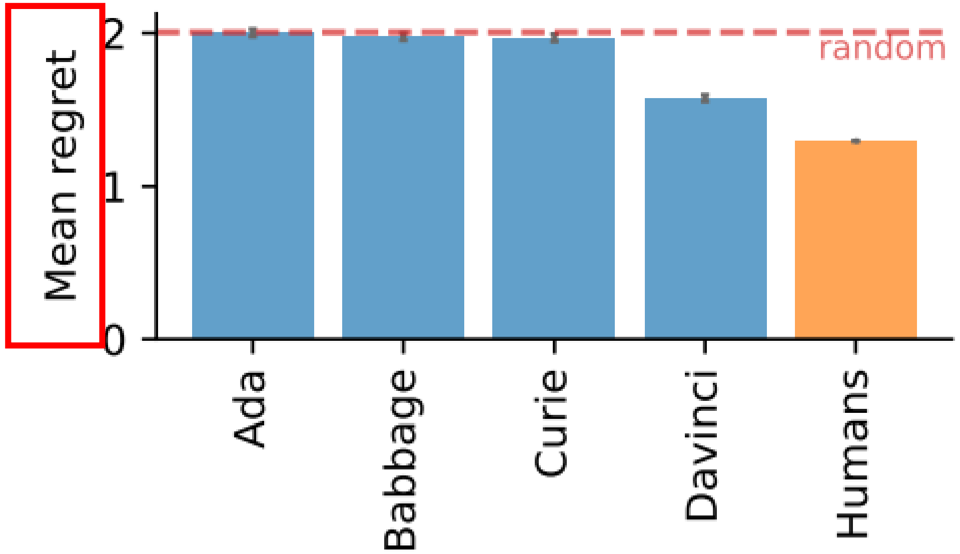
- Descriptions paradigm

最优option的预期结果与实际option的预期结果之间的差异

Q: Which option do you prefer?

- Option F: 69.0 dollars with 1.0% chance, 26.0 dollars with 99.0% chance.
- Option J: 2.0 dollars with 75.0% chance, 94.0 dollars with 25.0% chance.

A: Option



- Test 13000+ questions from a benchmark dataset

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D. & Griffiths, T. L. Using large-scale experiments and machine learning to discover theories of human decision-making. Science 372, 1209–1214 (2021).

仅"Davinci"具有超越随机水平的能力（$t(29134) = -16.85$， $p =< .001$）但仍未达到人类水平（$t(29134) = -11.50$, $p < .001$）

Decision-making

- Cognitive biases (Kahneman＆Tversky, 1972)
  - 通过对比特定problems pairs的回答识别出人类决策中的几种认知偏见

Problem	Prompt	p(F)
1	Q: Which option do you prefer? - Option F: 33% chance at 2,500 dollars, a 66% chance at 2,400 dollars, and a 1% chance of 0 dollars. - Option J: Guaranteed 2,400 dollars. A: Option	0.20455745
2	Q: Which option do you prefer? - Option F: 33% chance of 2,500 dollars (67% chance of 0 dollars). - Option J: 34% chance of 2,400 dollars (66% chance of 0 dollars). A: Option	0.38613685
3	Q: Which option do you prefer? - Option F: 80% chance of 4,000 dollars (20% chance of 0 dollars). - Option J: 100% guarantee of 3,000 dollars. A: Option	0.27955511
4	Q: Which option do you prefer? - Option F: 20% chance of 4,000 dollars (80% chance of 0 dollars). - Option J: 25% chance of 3,000 dollars (75% chance of 0 dollars). A: Option	0.22312672

Certainty effect：人们在决策中更倾向于选择结果更为确定的选项，即使它的期望更低

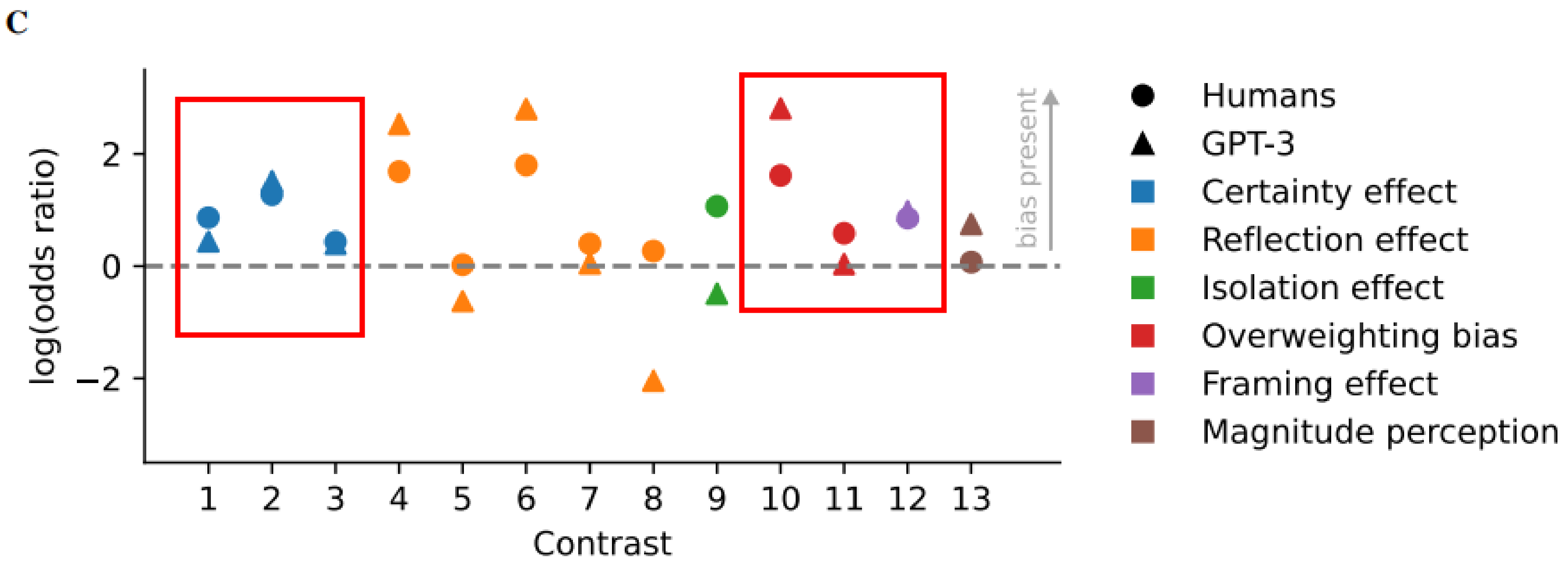
Contrast ID	1	2	3	4	5	6	7	8	9	10	11	12	13
Problem 1	2	4	7	7	4	9	6	16	4	6	9	13	15
Problem 2	1	3	8	3	8	5	10	17	11	5	10	12	16

log (odds ratio) = log (p(F|problem1) / p(F|problem2))

If log(odds ratio) > 0, 即存在该偏见

Decision-making

- GPT-3展现出了其中三种认知偏见
  - **Framing effect**：选择偏好会根据选择是以收益还是损失的方式呈现而改变；
  - **Certainty effect**：倾向于选择结果更为确定的选项，即使它的期望更低；
  - **Overweighting bias**：给两个小概率之间的差异（1%和2%）赋予了比给两个较大概率之间的相同差异（41%和42%）更高的重要性；



Information search

- Strategies for exploration-exploitation trade-off
  - **Directed exploration**: 倾向于采集未探索过的样本
  - **Random exploration**: 加入随机选择
- **Wilson's horizon task**
- Paradigm: two-armed bandit tasks
- Combination
  - Four forced-choice trials
    - **equal information condition (2 vs. 2)**
    - **unequal information condition (1 vs. 3)**
  - One (short) or six (long) free-choice trials ("horizon")
    - equal： 选择估计均值低的为random exploration
    - unequal： 选择观测次数少的为directed exploration

> You are going to a casino that owns two slot machines. You earn money each time you play on one of these machines.
>
> You have received the following amount of dollars when playing in the past:
> - Machine F delivered 51 dollars.
> - Machine J delivered 39 dollars.
> - Machine J delivered 40 dollars.
> - Machine J delivered 26 dollars.
>
> Your goal is to maximize the sum of received dollars within six additional rounds.
>
> Q: Which machine do you choose?
> A: Machine

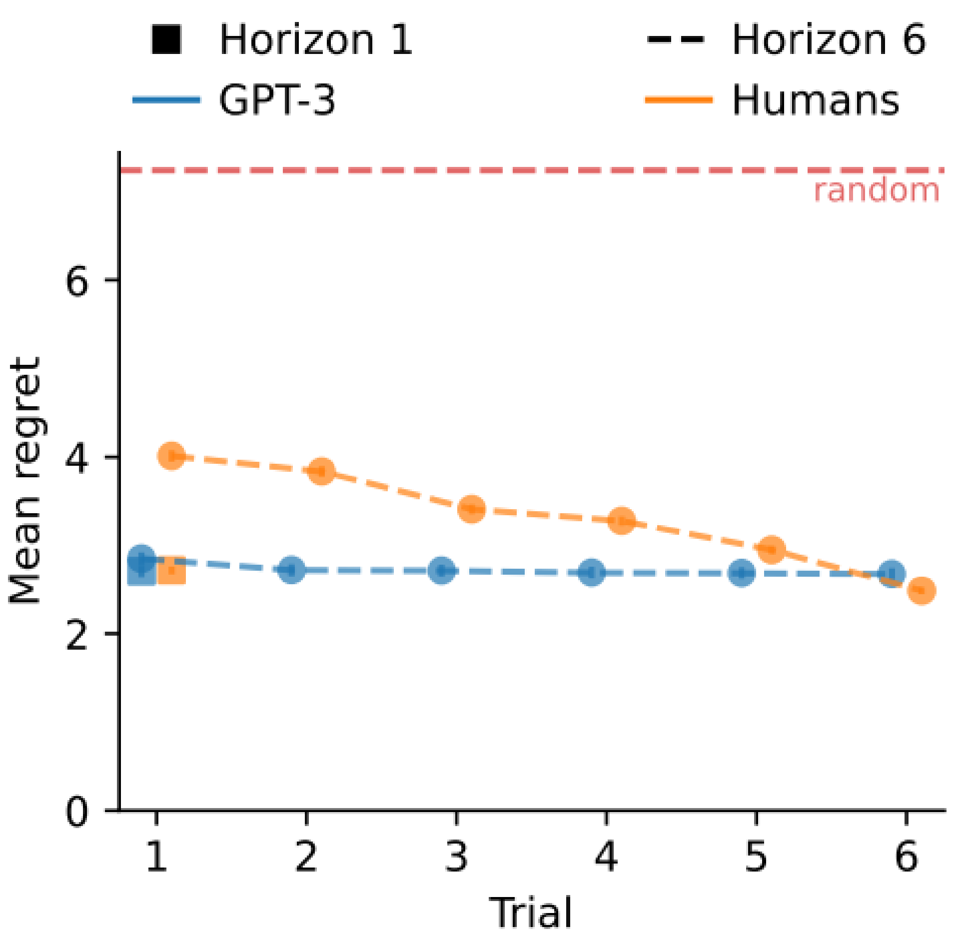
Information search

Short-horizon tasks

- GPT-3的表现与人类表现无显著差异（$t(5566) = -0.043$，$p = .97$）；
- GPT-3不仅能从直观描述下做出合理决策，还能从噪声样本中整合信息

long-horizon tasks

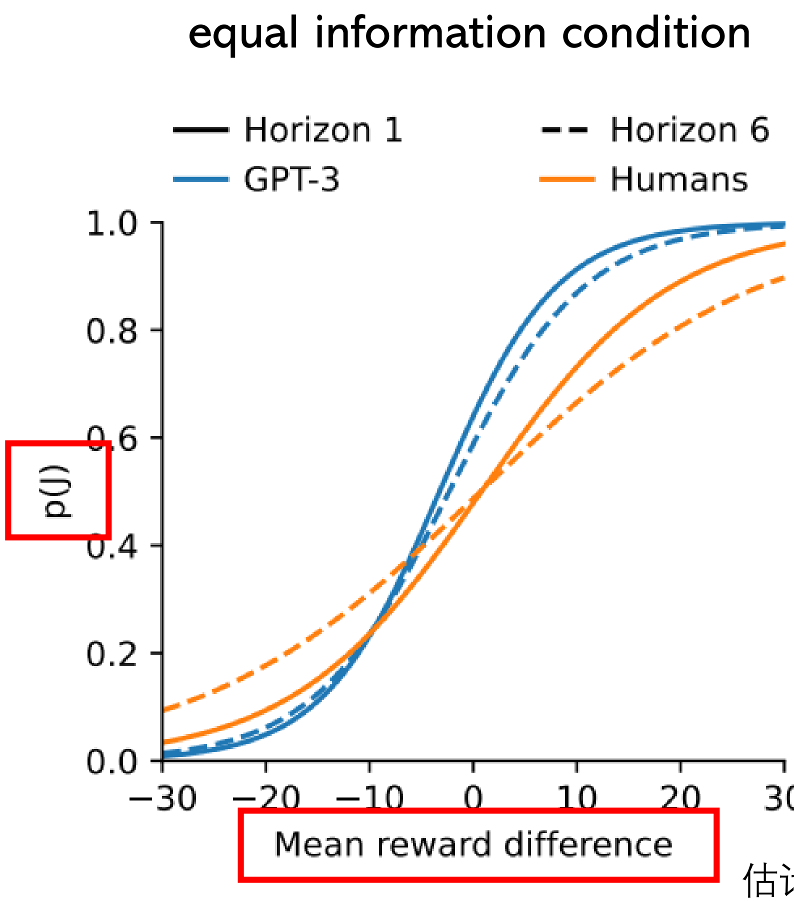
- GPT-3的初始regret显著低于人类（$t(5550) = -4.07$，$p < .001$），并且仅略高于Short-horizon tasks
- 人类的进步比GPT-3更大，并最终达到了稍微低于GPT-3但差异并不显著的水平（$t(5550) = -0.75$，$p = .23$）
- 整个实验中，GPT-3（$M = 2.72$，$SD = 5.98$）的regret显著低于人类受试者（$M = 3.24$，$SD = 10.26$），$t(38878) = -5.03$，$p < .001$



Information search

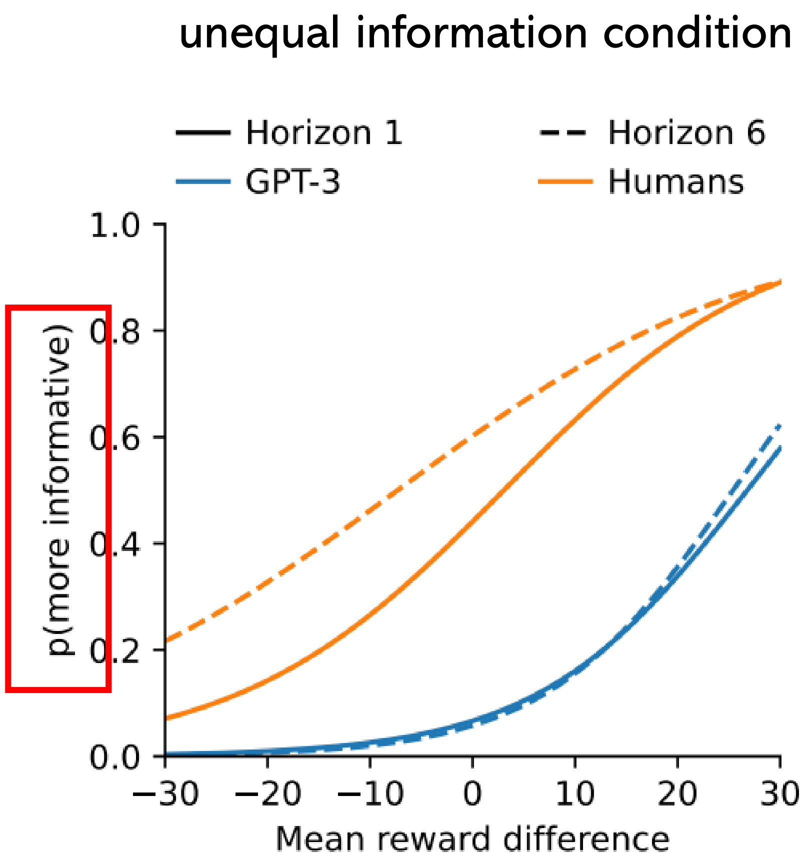
使用logistic regression为两种信息条件分别拟合模型
- 自变量：estimated reward difference, horizon, their interaction, and a bias term
- 因变量：是否选J（equal）；是否选more informative（unequal）

选择option J的概率（即选择估计均值更低的选项的概率）



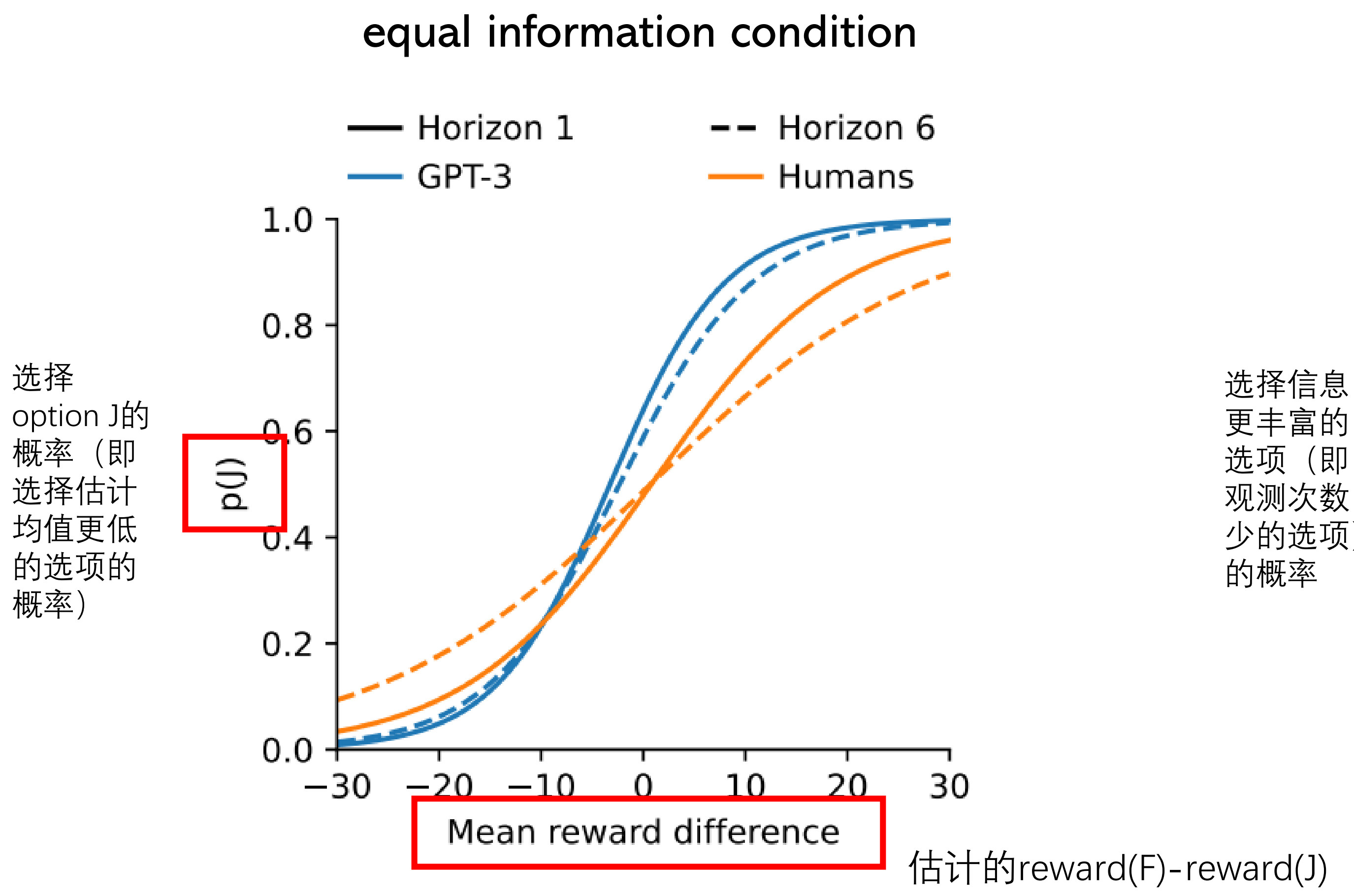
估计的reward(F)-reward(J)

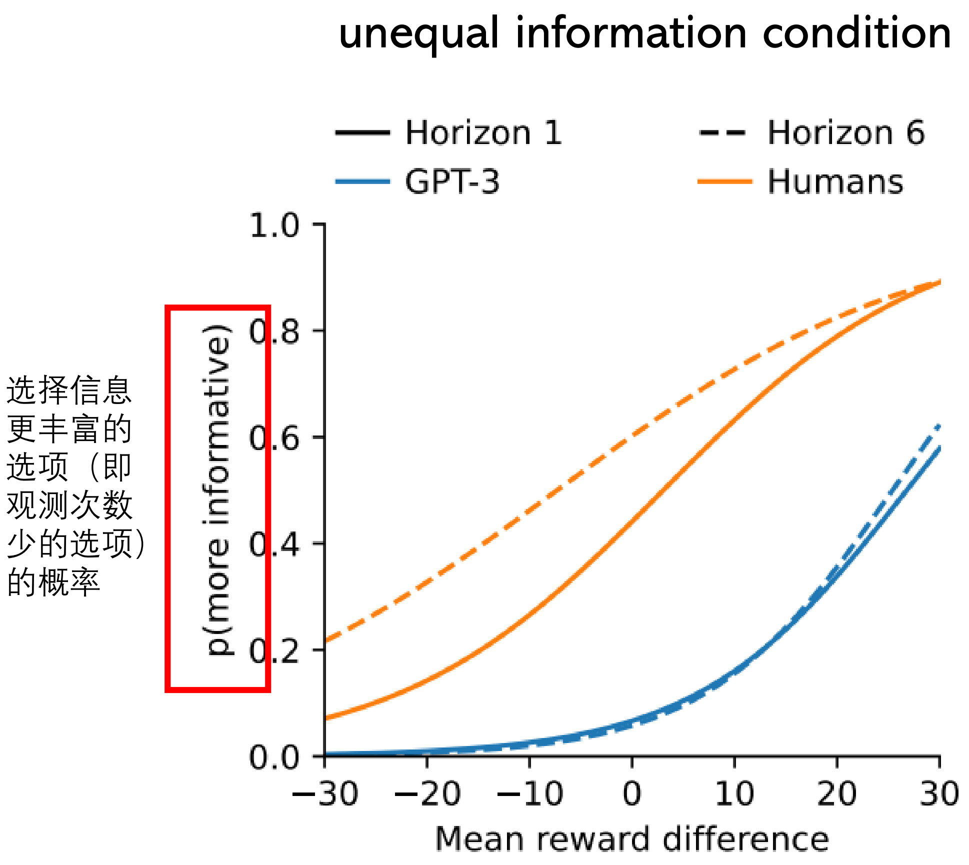
选择信息更丰富的选项（即观测次数少的选项）的概率



- Human：reward difference存在正面主效应，且与horizon存在负面交互效应；
- GPT-3：reward difference存在正面主效应，但不存在交互效应；
- 人类有策略地应用了random exploration，但GPT-3仅初级地使用了该策略，忽略了任务的horizon信息。

- Human：horizon存在正面主效应；
- GPT-3：horizon没有效应；
- GPT-3没有使用directed exploration





Deliberation

- **Two modes of learning**
  - Model-free learning (e.g., Q-learning)

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right)$$

  - Model-based learning (e.g., MCTS)

- **Two-step task paradigm**
  - **Model-free learning**: the probability of the selected first-stage action should increase upon receiving treasures in the second stage, regardless of whether the decision-maker experienced a rare or a common first-stage transition.
  - **Model-based learning**: upon encountering a rare transition and receiving treasures, the probability of the selected first-stage action should decrease.

You will travel to foreign planets in search of treasures. When you visit a planet, you can choose an alien to trade with. The chance of getting treasures from these aliens changes over time. Your goal is to maximize the number of received treasures.

Your previous space travels went as follows:
- 3 days ago, you boarded the spaceship to planet X, arrived at planet X, traded with alien D, and received treasures.
- 2 days ago, you boarded the spaceship to planet Y, arrived at planet X, traded with alien D, and received junk.
- 1 day ago, you boarded the spaceship to planet Y, arrived at planet Y, traded with alien K, and received junk.

Q: Do you want to take the spaceship to planet X or planet Y?
A: Planet X.

You arrive at planet X.

Q: Do you want to trade with alien D or F?
A: Alien

Deliberation

Two-step task

Figure 6 contains simulated behavior of a model-free and model-based reinforcement learning algorithm. For a detailed description of these algorithms, see Daw et al.[35].

2 (common/rare) * 2 (rewarded/unrewarded) design

保持first-stage action的概率

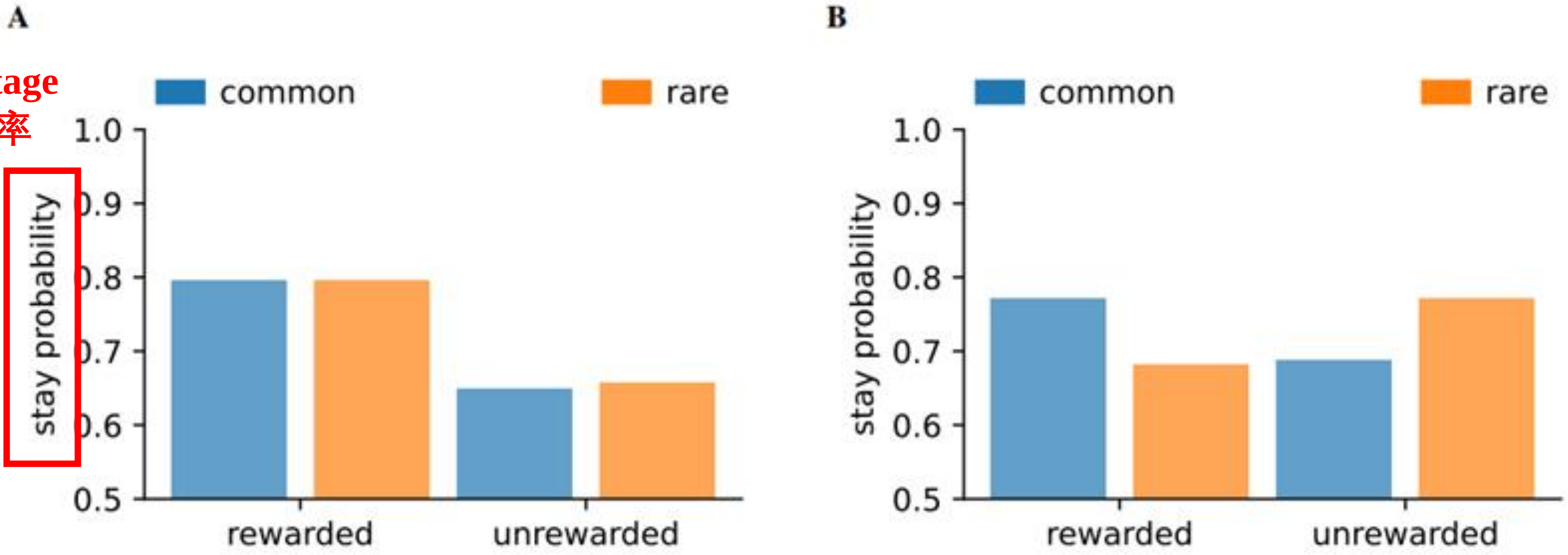
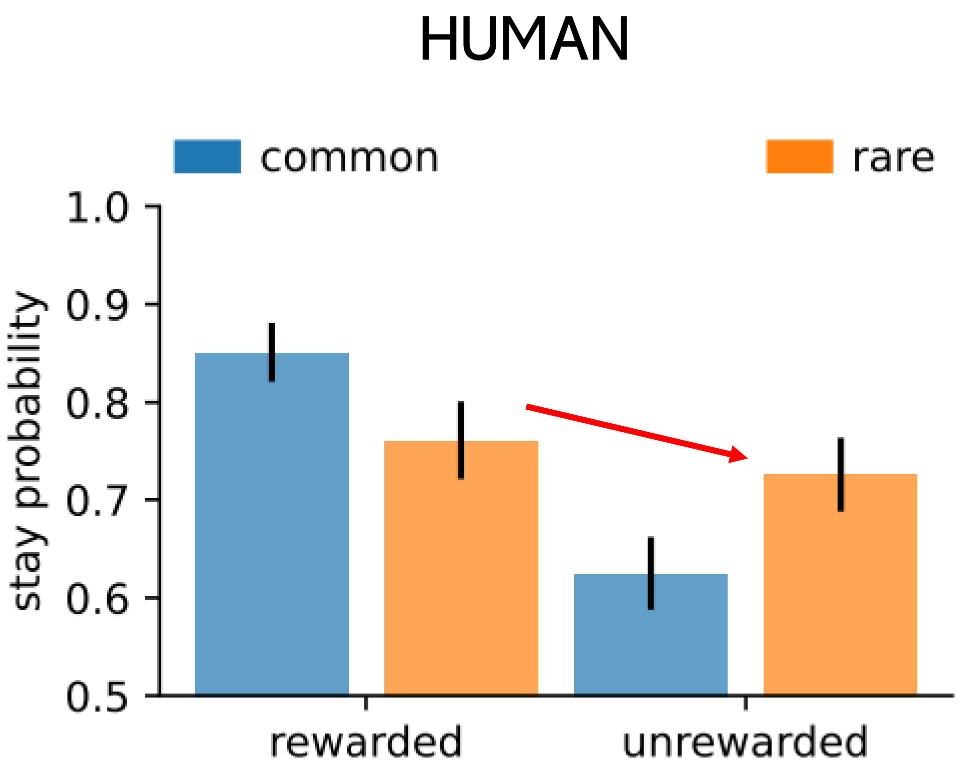
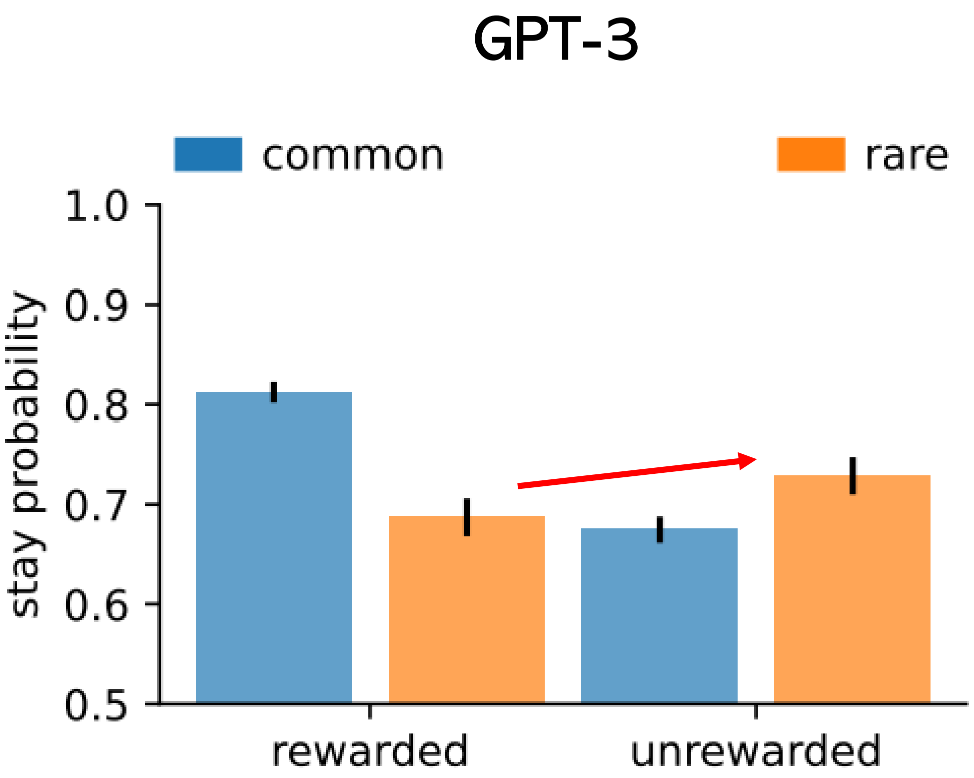


Figure 6. Model simulations on the two-step task. **A:** Model-free reinforcement learning algorithm. **B:** Model-based reinforcement learning algorithm. Figure adapted from Daw et al.[35].

Deliberation



人类倾向于使用model-free和model-based learning结合的方式解决此类问题



GPT-3倾向使用model-based的方式解决此类问题

与CRT的结果矛盾，需进一步探究复杂过程

Causal reasoning

Waldman & Hagmayer (2005)

- Show 20 observations of a three-variable system
- Show additional information about the causal structure of the system
  - **Common-cause condition**: A causes both B and C (B ← A → C)
  - **Causal-chain condition**: inverted the causal direction of A and B (B → A → C)
- Ask subjects to imagine 20 new observations (intervened B or observed B) and report for how many of these 20 new observations variable C would be active. (2*2 design)

You have previously observed the following chemical substances in different wine casks:
- Cask 1: substance *A* was present, substance *B* was present, substance *C* was present.
- Cask 2: substance *A* was present, substance *B* was present, substance *C* was present.
[...]
- Cask 20: substance *A* was absent, substance *B* was absent, substance *C* was absent.

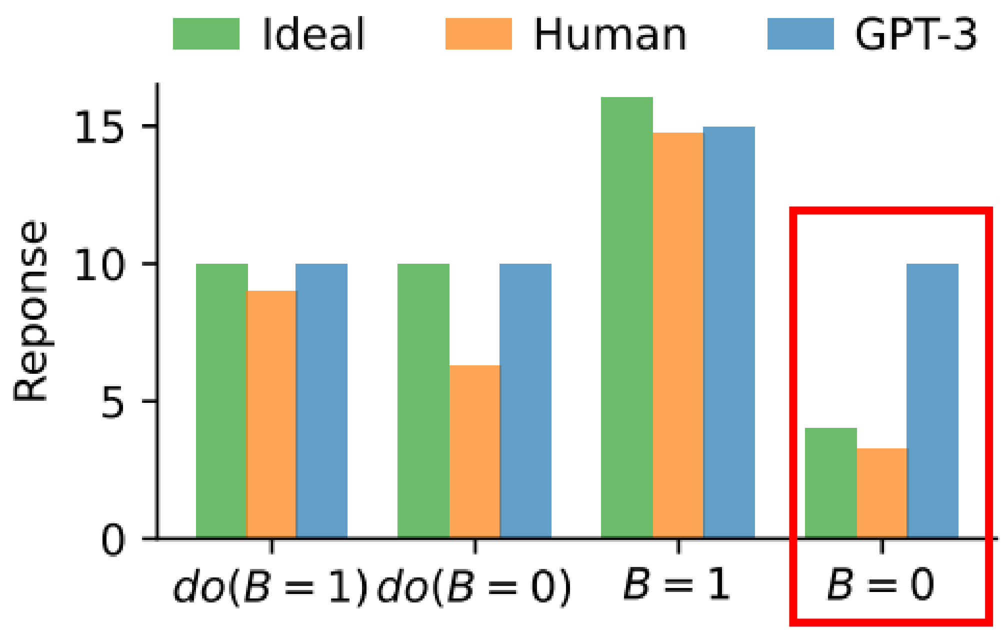
You have the following additional information from previous research:
- Substance *A* likely causes the production of substance *B*.
- Substance *A* likely causes the production of substance *C*.

Imagine that you test 20 new casks in which you have manually added substance *B*.

Q: How many of these new casks will contain substance *C* on average?
A: [insert] casks.

Causal reasoning

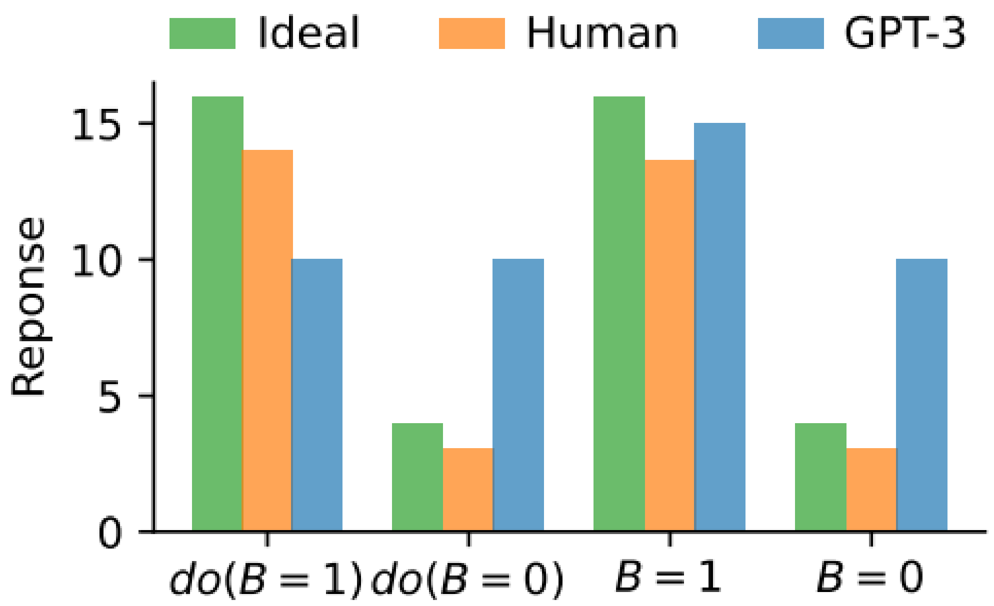
Common-cause condition (B ← A → C)



B=0时GPT-3的表现与Ideal Agent和人类均不符

人类的推断相较Ideal Agent更保守

Causal-chain condition (B → A → C)



四种条件下，GPT-3的表现均与Common-cause condition相同

GPT-3无法将关于潜在因果结构的额外信息纳入其推理过程