

Using cognitive psychology to understand GPT-3

Marcel Binz^{1,*} and Eric Schulz¹

¹MPRG 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.

Introduction

With the advent of increasingly capable artificial agents, comes the urgency to improve our understanding of how they learn and make decisions¹. Take as an example large language models². These models' abilities are, by many standards, impressive. They can generate text that human evaluators have difficulty distinguishing from text written by other humans², generate computer code³, or converse with humans about a range of different topics⁴. What is perhaps even more surprising, is that these models' abilities go beyond mere language generation: they can, for instance, also play chess at a reasonable level⁵ and solve university-level math problems⁶. These observations have prompted some to argue that this new class of *foundation models*, which are models trained on broad data at scale and adapted to a wide range of downstream tasks, shows some form of general intelligence⁷. Yet others have been more skeptical, pointing out that these models are still a far cry away from a human-level understanding of language and semantics⁸. But how can we evaluate whether or not these models—at least in some situations—learn and think like people? One approach towards evaluating a model's human-likeness comes from cognitive psychology. Psychologists, after all, are experienced in trying to formally understand another notoriously impenetrable algorithm: the human mind.

In the present article, we investigate the Generative Pre-trained Transformer 3 model (or short: GPT-3)² on several experiments taken from the cognitive psychology literature. Our analyses cover two types of experiments: vignette-based and task-based experiments. While vignette-based experiments involve a short and predefined description of a hypothetical scenario, task-based experiments are programmatically generated on a trial-by-trial basis. The selected tasks for both of these settings cover well-known areas of cognitive psychology: decision-making, information search, deliberation, and causal reasoning. We are primarily interested in whether GPT-3 can solve these tasks appropriately as well as how its behavior compares to human subjects. Our results show that GPT-3 can solve challenging vignette-based problems. Yet, we also highlight that these vignettes or similar texts might have been part of its training set. Moreover, we find that GPT-3's behavior strongly depends on how the vignettes are presented. Thus, we also subject GPT-3 to a battery of task-based problems. The results from these task-based assessments show that GPT-3 can make human-level decisions in both description-based and experience-based decision-making experiments, yet does not learn and explore in a human-like fashion. Furthermore, even though GPT-3 shows signatures of model-based reinforcement learning, it fails altogether in a causal reasoning task. Taken together, our results improve our understanding of current large language models, suggest ways in which they can be improved, and pave the way for future investigations using tools from cognitive psychology to study increasingly capable and opaque artificial agents.

GPT-3

GPT-3 is an auto-regressive language model². It utilizes the transformer architecture⁹—a deep learning model that heavily relies on the mechanism of self-attention—to produce human-like text. Just like recurrent neural networks, transformers are designed to process sequential data, such as natural language. However, unlike recurrent neural networks, transformers process the entire data all at once, with the attention mechanism providing context for any position in the input sequence. The model

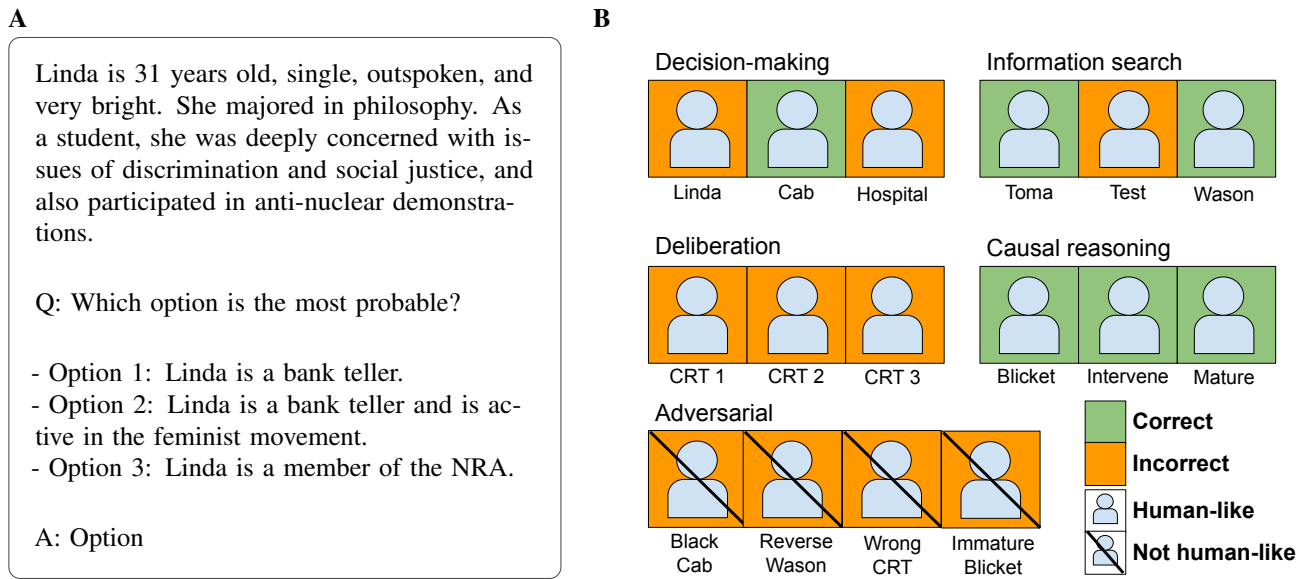


Figure 1. Vignette-based tasks. **A:** Example prompt of hypothetical scenario, in this case the famous Linda problem, as submitted to GPT-3. **B:** Results. While in 12 out of 12 standard vignettes, GPT-3 either answers correctly or makes human-like mistakes, it makes mistakes that are not human-like when given the adversarial vignettes.

itself is large, it has 175 billion parameters, and it was trained on a vast amount of text: hundreds of billions of words from the internet and books. GPT-3's architecture is similar to that of its predecessor, GPT-2¹⁰, but contains many more trainable parameters. Thus, GPT-3 can be thought of as an experiment in massively scaling up known algorithms¹¹. Larger models can capture more of the complexities of the data they were trained on and can transfer this knowledge to tasks that they have not been specifically trained for. Rather than being fine-tuned on a problem, these large language models can be given an instruction together with some examples of the task and identify what to do based on this alone. This is called “in-context learning” because the model picks up on patterns in its “context”, for example, the string of words that the model is asked to complete. GPT-3 does incredibly well at in-context learning across a range of settings¹², sometimes even performing at a level comparable to the best fine-tuned models^{2,13}. Since GPT-3 is one of the biggest and most versatile large language models, it is a good candidate to be scrutinized using cognitive psychology.

A cognitive psychology view on GPT-3

We will subject GPT-3 to several tasks taken from the cognitive psychology literature. These tasks fall into four categories: 1. decision-making, 2. information search, 3. deliberation, and 4. causal reasoning. We will begin our investigations with several, classical vignette-based problems. For these vignette-based investigations, we confronted GPT-3 with text-based descriptions of hypothetical situations while collecting its responses. However, as we will point out, these vignettes have the problem that GPT-3 has likely experienced identical or similar such tasks in its training data. Moreover, we found that GPT-3's response can be tampered with just by marginally changing the vignettes and thereby creating adversarial situations. Thus, we also evaluated GPT-3's abilities in various task-based experiments. In these task-based investigations, we take canonical tasks from the literature and emulate their experimental structure as programmatically generated text to which GPT-3 responds on every experimental trial. We then use GPT-3's responses to analyze its behavior similar to how cognitive psychologists would analyze human behavior in the same tasks.

Results

We used the public OpenAI API to run all our simulations¹⁴. There are four GPT-3 models accessible through this API: “Ada”, “Babbage”, “Curie” and “Davinci” (sorted from the least to the most complex model). We focused our investigation on the most powerful of these models (“Davinci”) unless otherwise noted. We furthermore set the temperature parameter to 0, leading to deterministic answers, and kept the default values for all other parameters.

Vignette-based investigations

For the vignette-based investigations, we took canonical scenarios from the cognitive psychology literature, entered them as prompts into GPT-3, and recorded its answer. For each scenario, we report if GPT-3 responded correctly or not. Moreover, we classified each response as something a human could have said because it was either the correct response or a mistake commonly observed in human data. For cases where there were only two options, one correct and one that is normally chosen by human subjects, we added a third option that was neither correct nor plausibly chosen by people. The following subsections briefly summarize our main findings. We refer the reader to SI Appendix for a detailed description of the submitted prompts and GPT-3's corresponding answers.

Decision-making: Heuristics and biases

We began our investigations of GPT-3's decision-making by prompting the canonical "Linda problem"¹⁵ (Linda, see Figure 1A). This problem has been known to assess the conjunction fallacy, a formal fallacy that occurs when it is assumed that specific conditions are more probable than a single general one. In the standard vignette, a hypothetical woman named Linda is described as "outspoken, bright, and politically active". Participants are then asked if it was more likely that Linda is a bank teller or that she is a bank teller *and* an active feminist. GPT-3, just like people, chose the second option, thereby falling for the conjunction fallacy.

Next, we prompted the so-called "cab problem"¹⁶ (Cab, see SI Appendix), in which participants commonly fail to take the base rate of different colors of taxis in a city into account when judging the probability of the color of a cab that was involved in an accident. Unlike people, GPT-3 did not fall for the base-rate fallacy, i.e. to ignore the base rates of different colors, but instead provided the (approximately) correct answer.

Finally, we asked GPT-3 to provide an answer to the "hospital problem"¹⁷ (Hospital, see SI Appendix), in which participants are asked which of two hospitals, a smaller or a larger one, is more likely to report more days on which more than 60% of all born children were boys. While the correct answer would be the smaller hospital (due to the larger variance of smaller samples), GPT-3, just like people, thought that the probability was about equal.

Information search: Questions and hypothesis tests

First, we assessed if GPT-3 can adaptively change between constraint-seeking vs. hypothesis-scanning questions. Constraint-seeking questions target a feature shared by multiple objects, such as "Is the person female?", whereas hypothesis-scanning questions target a single object, such as "Is the person Linda?". Crucially, which type of question is more informative depends on past observations. Ruggeri et al.¹⁸ manipulated the particular reasons for why a fictitious character named Toma was repeatedly late to school (Toma, see SI Appendix). While for one group he was frequently late because his bicycle had broken, for the other group he was late for various reasons with half of them being that he could not find various objects. While trying to find out why Toma is late to school again, the first group should ask the hypothesis-scanning question "Was he late because his bicycle broke?", whereas the second group should ask the constraint-seeking question "Was he late because he could not find something?". GPT-3 picked the appropriate question in each scenario.

Secondly, we confronted GPT-3 with a scenario originally presented by Baron et al.¹⁹ in which subjects need to choose an appropriate test to discriminate between two illnesses (Test, see SI Appendix). Empirically, participants tend to choose the wrong test, likely because they overvalue questions that have a high probability of a positive result given the most likely hypothesis. GPT-3, just like people, fell for the same congruence bias.

Finally, we presented Wason's well-known "Card Selection Task"²⁰ to GPT-3, explaining that the visible faces of four cards showed A, K, 4 and 7, and that the truth of the proposition "If a card shows a vowel on one face, then its opposite face shows an even number" needed to be tested (Wason, see SI Appendix). GPT-3 suggested to turn around A and 7, which is commonly accepted as the correct answer, even though most people turn around A and 4.

Deliberation: The Cognitive Reflection Test

We also tried to estimate GPT-3's tendency to override an incorrect fast response with answers derived by further deliberation. For this, we prompted the three items of the Cognitive Reflection Test²¹ (CRT1-CRT3, see SI Appendix). One example item of this task is: "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?". While the initial response might be to say "100", 100 machines would just be as fast as 5 machines and thus also take 5 minutes. For all three items of the CRT, GPT-3 responded with the intuitive but incorrect answer, as has been observed in earlier work²².

Causal reasoning: Blickets, interventions, and counterfactuals

We lastly assessed GPT-3's causal reasoning abilities. In a first test, we prompted GPT-3 with a version of the well-known "Blicket" experiment²³ (Blicket, see SI Appendix). For this, blickets are introduced as objects that turn on a machine. Afterward, two objects are introduced. The first object turns on the machine on its own. The second machine does not turn on the machine on its own. Finally, both objects together turn on the machine. GPT-3, just like people, managed to correctly identify that the first but not the second object is a blicket.

In a second test, we asked GPT-3 to intervene in a scenario by removing the correct object to prevent an effect after having read about three different objects, one causing and two not causing the effect (in this case, an allergic reaction; Intervene, see SI Appendix). GPT-3 identified the correct object to be removed.

In the final test, we probed GPT-3's ability of mature causal reasoning²⁴ (Mature, see SI Appendix). In this task, GPT-3 was told that there were four pills: A, B, C and D. While A and B individually could kill someone, C and D could not. GPT-3 successfully answered multiple questions about counterfactuals correctly, such as: "A man took pill B and pill C and he died. If he had not taken pill B, could he still have died?"

Problems with vignette-based investigations

Of the 12 vignette-based problems presented to GPT-3, it answered six correctly and all 12 in a way that could be described as human-like (Figure 1B). Does this mean that GPT-3 could pass as a human in a cognitive psychology experiment? We believe that the answer, based on the vignette-based tasks alone, has to be "No.". Since many of the prompted scenarios were taken from famous psychological experiments, there is a chance that GPT-3 has encountered these scenarios or similar ones in its training set. Moreover, in additional investigations, we found that many of the vignettes could be slightly modified, i.e., made into adversarial vignettes, such that GPT-3 would give vastly different responses. In the cab problem, for example, it is clearly stated that 15% of the cabs are blue and 85% are green. Yet asking GPT-3 about the probability that a cab involved in an accident was black, it responded with "20%" (Black Cab, see SI Appendix). Simply changing the order of the options in Wason's card selection task from "A, K, 4, and 7" to "4, 7, A, and K" caused GPT-3 to suggest turning around "A" and "K" (Reverse Wason, see SI Appendix). Giving GPT-3 the first item of the CRT and stating that "The bat costs \$1.00 more than the ball.", it still thought that the ball was \$0.10 (Wrong CRT, see SI Appendix). Finally, when phrasing the mature causal reasoning problem as a "Blicket" problem in which machines could be turned on or off, GPT-3 answered some questions incorrectly while contradicting itself in its explanations (Immature Blicket, see SI Appendix). There have recently been other, much larger investigations using similar vignettes, whose results agree largely with our assessment²⁵.

Task-based investigations

The results from the previous section indicate that GPT-3 can produce passable responses in some vignette-based tasks. It is, however, not possible to decide whether it is merely behaving like a parrot, repeating what it has seen in the training data, or whether it is reasoning successfully. We, therefore, next turned our lens of investigation to a more challenging setting and tested GPT-3 on actual, task-based experiments. In order to do so, we selected a set of four classical experiments that we believe to be representative of the cognitive psychology literature. For each of these, we programmatically generated a description that was entered as a prompt and –if there were multiple trials– updated the text with GPT-3's response and the received feedback.

Decision-making: Decisions from descriptions

How people make decisions from descriptions is one of the most well-studied areas of cognitive psychology, ranging from the early, seminal work of Kahneman & Tversky²⁸ to modern, large-scale investigations^{26,27}. In the decisions from descriptions paradigm, a decision-maker is asked to choose between one of two hypothetical gambles like the ones shown in Figure 2A. To test whether GPT-3 can reliably solve such problems, we presented the model with over 13,000 problems taken from a recent benchmark data-set²⁶. Figure 2B shows the regret, which is defined as the difference between the expected outcome of the optimal option and that of the actually chosen option, obtained by different models in the GPT-3 family and compares their performance to human decisions. We found that only the largest of the GPT-3 models ("Davinci") was able to solve these problems above chance-level ($t(29134) = -16.85, p < .001$), whereas the three smaller models did not (all $p > 0.05$). While the "Davinci" model did reasonably well, it did not reach human-level performance ($t(29134) = -11.50, p < .001$).

However, given that GPT-3 was not too far away from human performance, it is reasonable to ask whether the model also exhibited human-like, cognitive biases. In their original work on prospect theory, Kahneman & Tversky¹⁷ identified several biases of human decision-making by contrasting answers to multiple carefully selected problems pairs. We replicated the original analysis of Kahneman & Tversky using choice probabilities of GPT-3 and found that GPT-3 showed three of the six biases identified by Kahneman & Tversky. First, it displayed a framing effect, meaning that its preferences changed depending on whether a choice was presented in terms of gains or losses. GPT-3 was also subject to a certainty effect, meaning that it preferred guaranteed outcomes to risky ones even when they had slightly lower expected values. Finally, GPT-3 showed an overweighting bias and assigned higher importance to a difference between two small probabilities (e.g., 1% and 2%) than to the same differences between two larger probabilities (e.g., 41% and 42%). Figure 2C contains an analysis of these three biases and the three additional ones we did not find in GPT-3. For a detailed description of the conducted analysis, see SI Appendix.

Information search: Directed and Random Exploration

GPT-3 did well in the vignette-based information search tasks, so we were curious how it would fare in a more complex setting. The multi-armed bandit paradigm provides a suitable test-bed for this purpose. It extends the decisions from descriptions paradigm from the last section by adding two layers of complexity. First, the decision-maker is not provided with descriptions

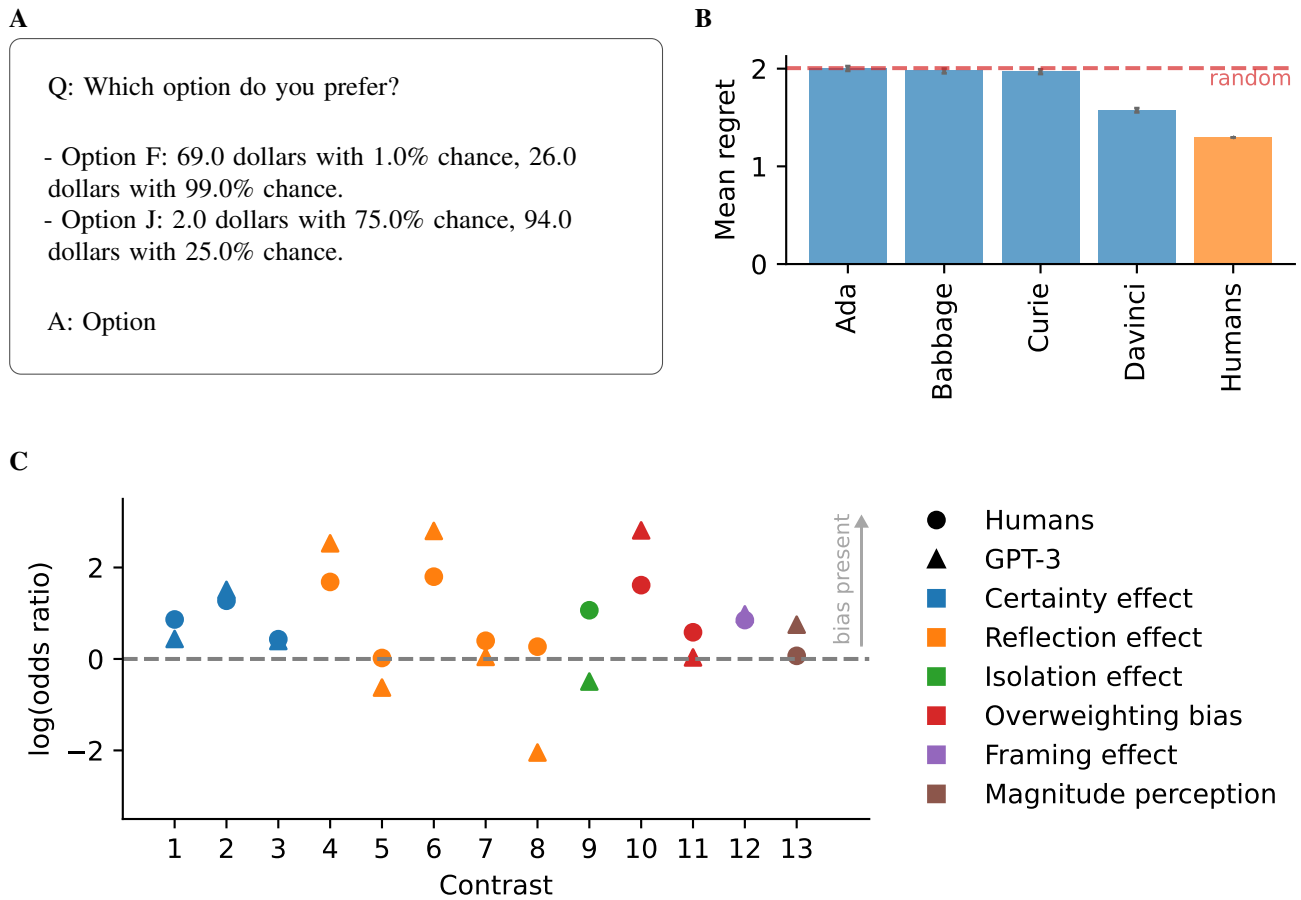


Figure 2. Decisions from descriptions. **A:** Example prompt of a problem provided to GPT-3. **B:** Mean regret averaged over all 13,000 problems taken from Peterson et al.²⁶. Lower regret means better performance. Error bars indicate the standard error of the mean. **C:** Log-odds ratios of different contrasts used to test for cognitive biases. Positive values indicate that the given bias is present in humans (circle) or GPT-3 (triangle). Human data adapted from Ruggeri et al.²⁷. For a detailed description of this analysis, see SI Appendix.

for each option anymore but has to learn their value from noisy samples, i.e. from experience²⁹. Second, the interaction is not confined to a single choice but potentially involves repeated decisions about which option to sample. Together, these two modifications call for an important change in how a decision-maker must approach such problems. It is not enough to merely exploit currently available knowledge anymore, but also crucial to explore options that are unfamiliar and thereby gain information about their value. Previous research suggests that people solve this exploration-exploitation trade-off by applying a combination of two distinct strategies: directed and random exploration³⁰. Whereas directed exploration encourages the decision-maker to collect samples from previously unexplored options, random exploration strategies inject some form of stochasticity into the decision process^{31,32}.

Wilson's horizon task is the canonical experiment to test whether a decision-maker applies the two aforementioned forms of exploration³⁰. It involves a series of two-armed bandit tasks, in each of which the decision-maker is provided with data from four forced-choice trials, followed by either one or six free-choice trials (referred to as the horizon). Forced-choice trials are used to control the amount of information available to the decision-maker. They either provide two observations for each option (equal information condition) or a single observation from one option and three from the other (unequal information condition). These two conditions make it possible to tease apart directed and random exploration by looking at the decision in the first free-choice trial. In the equal information condition, a choice is classified as random exploration if it corresponds to the option with the lower estimated mean. In the unequal information condition, a choice is classified as directed exploration if it corresponds to the option that was observed fewer times during the forced-choice trials. Note that short-horizon tasks do not benefit from making exploratory choices and, hence, we should expect the decision-maker to make fewer such choices in them.

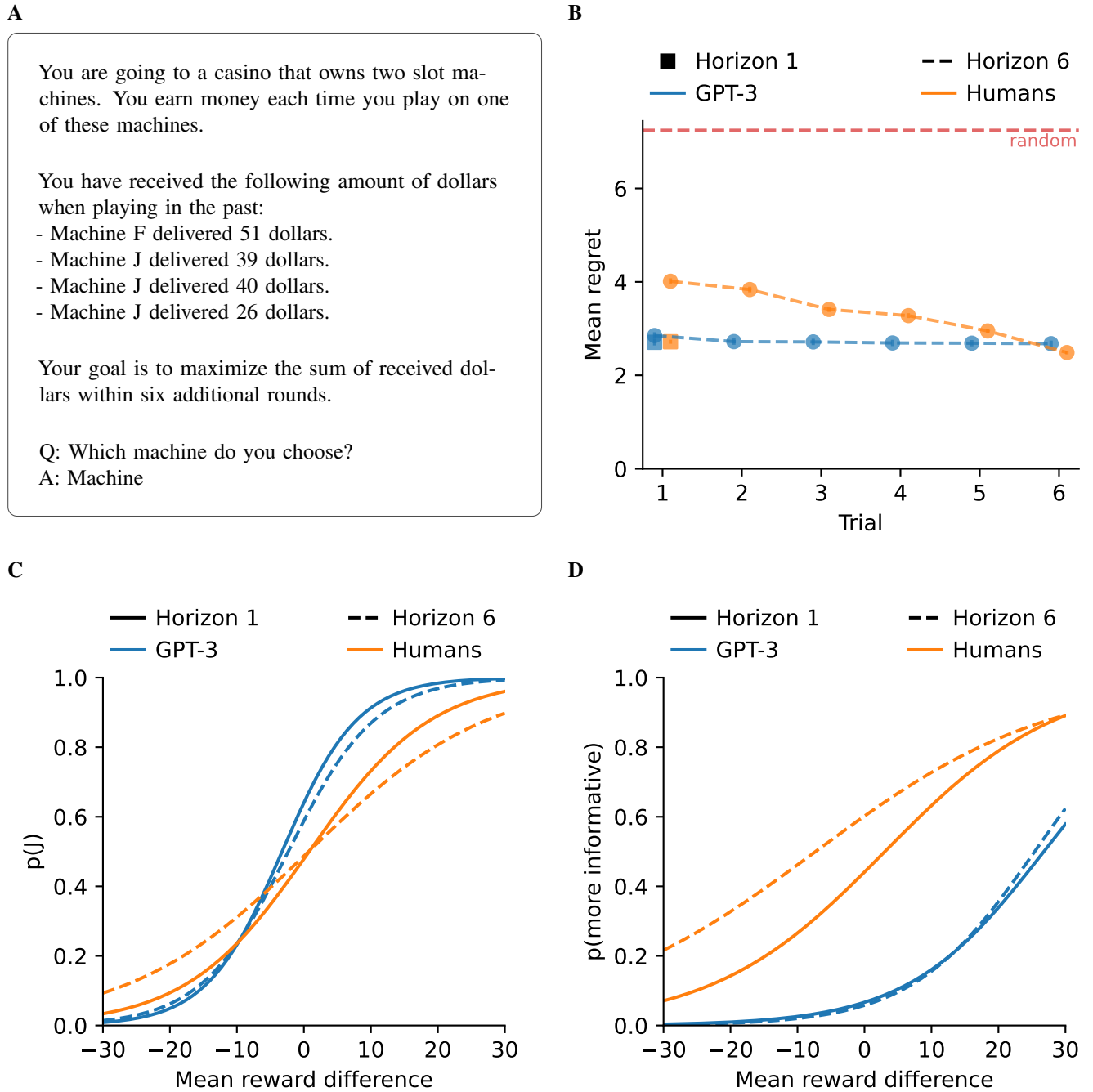


Figure 3. Horizon task. **A:** Example prompt for one trial as submitted to GPT-3. **B:** Mean regret for GPT-3 and human subjects by horizon condition. Lower regret means better performance. Error bars indicate the standard error of the mean. Human data taken from Zaller et al.³³. **C:** Probability of selecting option “J” in the equal information condition for both GPT-3 and human subjects by horizon condition. **D:** Probability of selecting the more informative option in the unequal information condition for both GPT-3 and human subjects by horizon condition.

We presented a text-based version of the horizon task as illustrated in Figure 3A to GPT-3. Figure 3B compares the model’s regret to the regret of human subjects. For short-horizon tasks, GPT-3’s performance was indistinguishable from human performance ($t(5566) = -0.043, p = .97$). This result highlights that GPT-3 can not only make sensible decisions when presented with descriptions of options but is also able to integrate this information from noisy samples. The initial regret of GPT-3 in long-horizon tasks was significantly lower than the corresponding human regret ($t(5550) = -4.07, p < .001$) and was

only slightly above the one from short-horizon tasks. However, within each task people improved more than GPT-3 and reached a final regret that was slightly but not significantly lower than that of GPT-3 ($t(5550) = -0.75, p = .23$). Looking at the entire experiment, GPT-3 ($M = 2.72, SD = 5.98$) achieved a significantly lower regret than human subjects ($M = 3.24, SD = 10.26$), $t(38878) = -5.03, p < .001$.

To investigate how GPT-3 managed the trade-off between exploration and exploitation, we fitted a separate logistic regression model for each information condition. We used the estimated reward difference, horizon, their interaction, and a bias term as independent variables for both models. The model for the equal information condition used an indicator for selecting option J in the first free-choice trial as the dependent variable, whereas the model for the unequal condition used an indicator for selecting the more informative option (i.e., the one that has been observed fewer times during the forced-choice trials). The results of this regression analysis are summarized visually in Figure 3C and D. If GPT-3 applied random exploration, we should observe a positive effect of estimated reward difference. If its random exploration was furthermore strategic, we should find more noisy decisions in long-horizon tasks of the equal information condition (reflected in a negative interaction effect of estimated reward difference and horizon). People show both of these effects³⁰. GPT-3 also displayed a significant effect of estimated reward difference ($\beta = 0.18 \pm 0.01, z = 14.48, p < .001$), suggesting that it used at least a rudimentary form of random exploration. However, we did not find a significant interaction effect between estimated reward difference and horizon ($\beta = -0.02 \pm 0.02, z = -1.47, p = .14$), indicating that GPT-3 did not apply random exploration in a strategic way and simply ignored the information about the task horizon. If GPT-3 applied directed exploration, we should find a positive effect of horizon in the unequal information condition, indicating that more informative actions were sampled more frequently when the horizon was longer. While humans show such an effect³⁰, we did not find it in GPT-3 ($\beta = -0.15 \pm 0.27, z = -0.56, p = .58$), which indicates that the model also did not employ directed exploration.

Lastly, we found that GPT-3 had a tendency to repeat previously observed options. For example, in the unequal information condition (Figure 3D), GPT-3 showed a strong bias to select the option from which it had seen more samples, even when there was not reason to do so. We believe that this bias partially arose from how GPT-3 was trained: if the goal is to predict future words, and the agent has recently observed a certain phrase (in this case either “Option F” or “Option J”), it is likely that the same phrase will appear again in the near future. Interestingly, humans also show such a perseveration bias in many situations³⁴, but, in the case of the horizon task, it seems to be overruled by other processes.

Deliberation: Model-based and model-free reinforcement learning

Many realistic sequential decision-making problems do not only require the decision-maker to keep track of reward probabilities, but also to learn how to navigate from state to state within an environment. Two modes of learning are plausible in such scenarios: model-free and model-based learning. Model-free learning –the more habitual mode of the two– stipulates that the decision-maker should adjust its strategy directly using the actually observed rewards. If something led to a good outcome, a model-free agent will do more of it; if it led to a bad outcome, a model-free agent will do less of it. Model-based learning –the more deliberate mode of the two– instead stipulates that the agent should explicitly learn the transition and reward probabilities of the environment and use them to update its strategy by reasoning about future outcomes.

These two modes of learning can be disentangled empirically in the two-step task paradigm³⁵. The two-step task involves a series of two-stage decision problems. There are two actions available from the initial state: taking a spaceship to planet X or to planet Y. Taking a spaceship transfers the agent to a second stage. The spaceship arrives with a probability of 0.7 to the selected planet, and with a probability of 0.3 to the other planet. After arriving at one of these planets, the agent encounters two local aliens with which it can trade. Trading with an alien can lead to receiving treasures or junk. The probabilities of receiving treasures are initialized randomly from a uniform distribution with a minimum of 0.25 and a maximum value of 0.75 for each alien. While these probabilities drift slowly over time to encourage learning, the first-stage transition probabilities remain fixed throughout the entire experiment. Model-free learning predicts that 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, on the other hand, predicts that, upon encountering a rare transition and receiving treasures, the probability of the selected first-stage action should decrease. SI Appendix contains plots of simulated behavior for the two learning strategies. People tend to solve this task using a combination of model-free and model-based learning^{35–37} as shown in Figure 4A.

We tested how GPT-3 learns in the two-step task by providing it with prompts like the one shown in Figure 4B. We ran 200 simulations in total and measured the stay probability of the first-stage action for each combination of transition (rare or common) and reward (treasures or junk). Each simulation involved 20 repetitions of the two stages. Figure 4C visualizes our results. We observed that the probability of repeating the previous first-stage action decreased after finding treasures through a rare transition ($t(1982) = -6.16, p < .001$). Meanwhile, the probability of repeating the same first-stage action increased after a rare and not rewarded action ($t(1814) = 2.33, p = .01$). These two findings suggest that GPT-3 relies on a deliberate model-based approach to solve the two-step task. Interestingly, this conclusion is at odds with our earlier simulations on the CRT, where GPT-3 consistently chose the intuitive but wrong over the more deliberate but correct answer. The contrast between

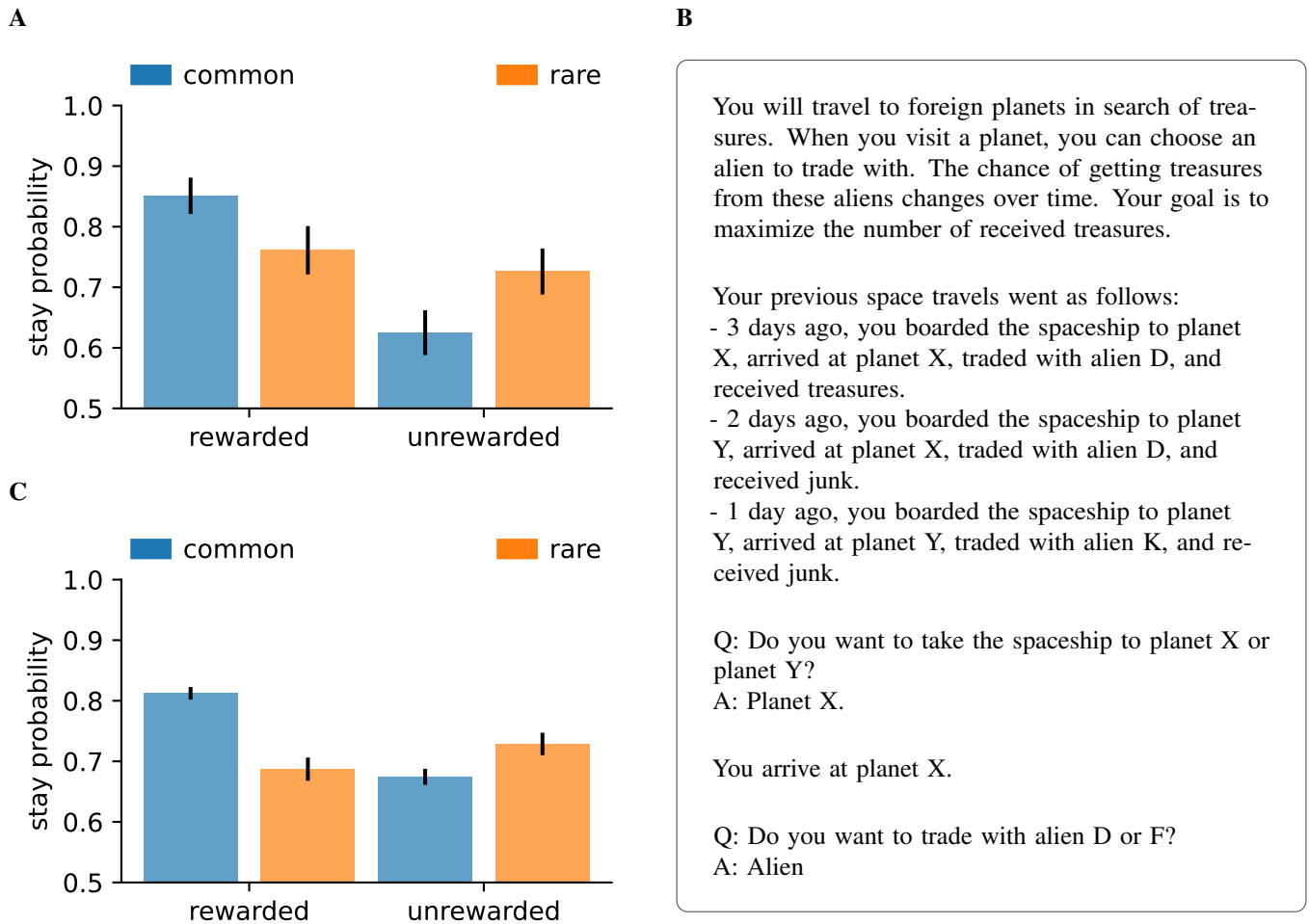


Figure 4. Two-step task. **A:** Human behavior in dependency of rewarded and unrewarded as well as common and rare transitions. Human data adapted from Daw et al.³⁵. **B:** Example prompt of one trial in the canonical two-step task as submitted to GPT-3. **C:** GPT-3's behavior in dependency of rewarded and unrewarded as well as common and rare transitions. Error bars indicate the standard error of the mean.

those two analyses suggests that the answer to whether GPT-3 engages in deliberate reasoning might be more nuanced than initially thought.

Causal reasoning: Interventions after passive observations

The analysis of the two-step task indicated that GPT-3 can learn a model of the environment and use this learned model to update its strategy. In our final test, we wanted to analyze whether GPT-3 can also use such a model to make more complex inferences, such as reasoning about cause and effect. From our earlier vignette-based investigations, we have already learned that GPT-3 can solve some causal reasoning problems, although these results depended heavily on how the problems were presented.

Perhaps the most crucial insight of theories of causal reasoning is that there is a difference between merely observing variables and actively manipulating them. Take, for instance, the classical example of a barometer. Under normal circumstances, barometer measurements provide insights into the upcoming weather. However, if someone would manually set the scale of the barometer to a particular value, then it would become totally uninformative about the weather — a clear difference from the observational inference. Waldman & Hagmayer³⁸ devised an experiment to highlight that people are sensitive to the difference between seeing and doing. They first presented subjects with 20 observations of a three-variable system, and then provided additional information about the causal structure of the system. In the common-cause condition, they told participants that A causes both B and C ($B \leftarrow A \rightarrow C$). In the causal-chain condition, they inverted the causal direction of A and B , such that B now causes A , which, as before, causes C ($B \rightarrow A \rightarrow C$). Finally, they asked their subjects to imagine 20 new observations

A

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:

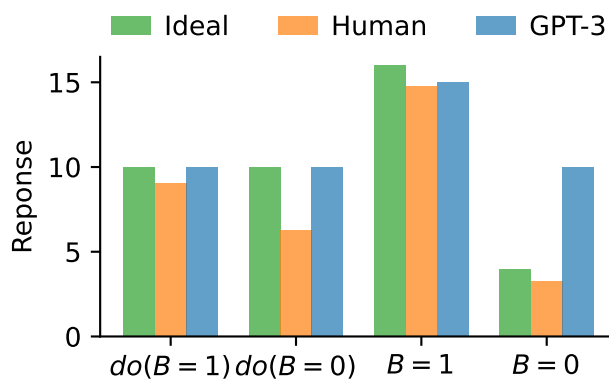
- Substance *A* likely causes the production of substance *B*.
- Substance *A* likely causes the production of substance *C*.

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.

B



C

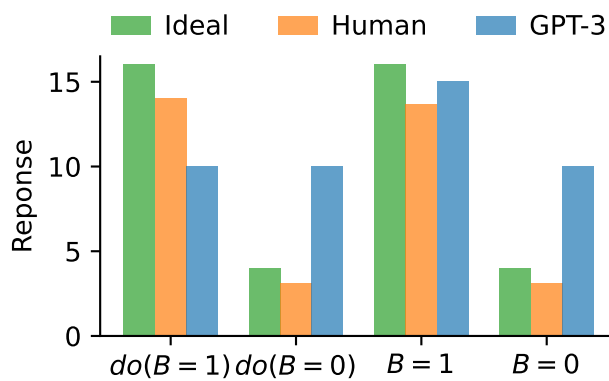


Figure 5. Causal reasoning. **A:** Example prompt for the causal reasoning task adapted from Waldman et al.³⁸. **B:** Simulation results comparing GPT-3’s responses with people as well as the ideal agent in the common-cause condition. **C:** Simulation results comparing GPT-3’s responses with people as well as the ideal agent in the causal-chain condition.

for which they either had actively intervened on the values of *B* or for which they merely had observed a particular value of *B*. Participants had to report for how many of these 20 new observations variable *C* would be active. Like in the barometer example, observing an active value of *B* in the common-cause condition enabled participants to make the inference that *A* was likely to be active as well, which, in turn, made it more likely that *C* was also active. In contrast, activating *B* by means of interventions did not allow for such an inference. Mathematically, the act of intervening can be formalized by Pearl’s $do()$ operator³⁹, which sets a variable to a particular value but deletes all arrows going into that variable from the causal graph. For the causal-chain condition one therefore would expect to find no differences between intervening and observing, as there was no arrow going into *B* that had to be deleted, and hence both inferences were identical.

We probed GPT-3’s ability to make causal inferences in this task using a cover story about substances found in different wine casks⁴⁰ (see Figure 5A). When provided with the additional information about the common-cause structure, GPT-3 made interventional inferences that matched the normative prescription of causal inference as illustrated in Figure 5B. GPT-3 furthermore predicted an increase in the number of observations with *C* = 1 after observing *B* = 1, which was in line with both the normative theory and human judgments. However, when observing *B* = 0, GPT-3 did not reduce its prediction, which was neither the correct inference nor human-like. The causal-chain condition does not lead to a difference between observational and interventional inferences from a normative perspective. While human subjects show exactly this pattern³⁸, GPT-3 made identical predictions compared to the common-cause condition as illustrated in Figure 5C. This observation suggests that the model was not able to incorporate the additional information about the underlying causal structure into its inference process and therefore makes it likely that the results from the common-cause condition were purely accidental. Taken together, these results suggest that GPT-3 has difficulties with causal reasoning in tasks that go beyond a vignette-based characterization.

Discussion

In 1904, sixteen leading academics of the Prussian Academy of Sciences signed a statement indicating that a horse, named “Clever Hans”, could solve mathematical problems at a human-like level. Back then, it took another scientist, Oskar Pfungst, years of systematic investigations to prove that the horse was merely reacting to the people who were watching him⁴¹. With the advent of large-scale machine learning models, the risks of over-interpreting simple behaviors as intelligent runs rampant. The abilities of large language models, in particular the ability to solve tasks beyond language generation, are impressive at first glance. These models have, therefore, been called many things; some think they are sentient⁴² and that they show a form of general intelligence⁷. Yet others believe that they are merely stochastic parrots⁴³ or a linguistic one-trick pony⁸. But how can we realistically gauge these models’ abilities?

We have argued to approach this problem similar to how Oskar Pfungst approached his object of study: via systematic investigations and psychological experimentation. Using tools from cognitive psychology, we have subjected one particular large language model, GPT-3, to a series of investigations, probing its decision-making, information search, deliberation, and causal reasoning abilities. Our results have shown that GPT-3 can solve some vignette-based experiments similarly or better than human subjects. However, interpreting these results is difficult because many of these vignettes might have been part of its training set, and GPT-3’s performance suffered greatly given only minor changes to the original vignettes. We, therefore, turned the lens of our investigations to task-based assessments of GPT-3’s abilities. Therein, we found that GPT-3 made reasonable decisions for gambles provided as descriptions while also mirroring some human behavioral biases. GPT-3 also managed to solve a multi-armed bandit task well, where it performed better than human subjects; yet it only showed traces of random but not of directed exploration. In the canonical two-step decision-making task, GPT-3 showed signatures of model-based reinforcement learning. However, GPT-3 failed spectacularly in using an underlying causal structure for its inference, leading to responses that were neither correct nor human-like.

What do we make of GPT-3’s performance in our tasks? We believe that GPT-3’s performance contained both surprising and expected elements. We found it surprising that GPT-3 could solve many of the provided tasks reasonably well, that it performed well in gambles, a simple bandit task, and even showed signatures of model-based reinforcement learning. These findings could indicate that—at least in some instances—GPT-3 is not just a stochastic parrot and could pass as a valid subject for some of the experiments we have administered. Yet what was not surprising were some of GPT-3’s failure cases. GPT-3 did not show any signatures of directed exploration. We believe that this is intuitive and can be explained by the differences in how humans and GPT-3 learn about the world. Whereas humans learn by connecting with other people, asking them questions, and actively engaging with their environments, large language models learn by being passively fed a lot of text and predicting what word comes next. GPT-3 also failed to learn about and use causal knowledge in a simple reasoning task. Causal reasoning is frequently seen as a pillar of intelligent behavior⁴⁴ and has been difficult to master for artificial agents⁴⁵. We believe it makes sense that GPT-3 struggles to reason causally because acquiring knowledge about interventions from passive streams of data is hard to impossible⁴⁶. The upside of our findings is the recommendation that to create more intelligent agents researchers should not only scale up algorithms that are passively fed with data but instead let agents directly interact and engage with the world⁴⁷.

We are not the first to probe large-scale machine learning models’ abilities. Indeed, recently there has been a push towards creating large benchmarks to assess the capability of foundation models^{48–50}. Large language models have also been studied using other methods from cognitive psychology, such as property induction⁵¹, thinking-out-loud protocols⁵², or learning causal over-hypotheses⁵³, where researchers have come to similar conclusions. Methods from cognitive psychology have also previously been applied to understand other deep learning models’ behavior⁵⁴. Therefore, our current work can be seen as part of a larger scientific movement where methods from psychology are becoming increasingly more important to understand capable black-box algorithms’ learning and decision-making processes^{55–58}.

Although we consider the present work as a step towards a psychological understanding of foundation models, several shortcomings remain. First of all, as we have seen in our vignette-based assessment, GPT-3’s responses often times crucially depend on how a prompt is presented. The same might hold for our task-based assessments, where it is conceivable that GPT-3’s behavior could change if the generating program of the tasks was modified. Yet we have simply tried to show that—in principle—GPT-3 could solve some of these tasks and believe that our current results emphasize the differences between GPT-3 and humans well. Secondly, we have only focused on a rather small subset of cognitive tasks, where we have tried to cover informative ground about GPT-3’s abilities. Futures investigations could focus on additional psychological domains such as category learning, problem-solving, or economic games, to name but a few. Finally, our current results run the risk of portraying GPT-3 as more intelligent than it actually is, simply because canonical tasks taken from the psychological literature might be too easy to solve. In that sense, showing that large language models can perform well in such tasks might tell us more about how solvable and perhaps overly simplistic some tasks are than about GPT-3 itself and point to the importance of using more complex paradigms to study both natural and artificial agents^{59,60}.

To summarize, we studied GPT-3, a recent large-scale language model, using tools from cognitive psychology. We assessed GPT-3’s decision-making, information search, deliberation, and causal reasoning abilities, and found that it was able to solve

most of the presented tasks at a decent level. Less than two years ago, the sheer fact that a general-purpose language model could give reasonable responses to our problems would have been a large surprise. From this perspective, our analysis highlights how far these models have come. Nevertheless, we also found that small perturbations to the provided prompts easily led GPT-3 astray and that it lacks important features of human cognition, such as directed exploration and causal reasoning. While it does not seem so far-fetched that even larger models could acquire more robust and sophisticated reasoning abilities, we ultimately believe that actively interacting with the world will be crucial for matching the full complexity of human cognition. Fortunately, many users already interact with GPT-3-like models, and this number is only increasing with new applications on the horizon. Future language models will likely be trained on this data, leading to a natural interaction loop between artificial and natural agents.

References

1. Gunning, D. *et al.* Xai—explainable artificial intelligence. *Sci. Robotics* **4**, eaay7120 (2019).
2. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).
3. Chen, M. *et al.* Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
4. Lin, Z. *et al.* Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 13622–13623 (2020).
5. Noever, D., Ciolino, M. & Kalin, J. The chess transformer: Mastering play using generative language models. *arXiv preprint arXiv:2008.04057* (2020).
6. Drori, I. *et al.* A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *arXiv preprint arXiv:2112.15594* (2021).
7. Chalmers, D. Gpt-3 and general intelligence. *Dly. Nous*, July **30** (2020).
8. Marcus, G. & Davis, E. Gpt-3, bloviator: Openai’s language generator has no idea what it’s talking about. *Technol. Rev.* (2020).
9. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).
10. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. *None* (2018).
11. Sutton, R. The bitter lesson. *Incomplete Ideas (blog)* **13**, 12 (2019).
12. Liu, J. *et al.* What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804* (2021).
13. Lampinen, A. K. *et al.* Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329* (2022).
14. OpenAI API. <https://beta.openai.com/overview>. Accessed: 2022-06-20.
15. Tversky, A. & Kahneman, D. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychol. review* **90**, 293 (1983).
16. Tversky, A. & Kahneman, D. Causal schemas in judgments under uncertainty. *Prog. social psychology* **1**, 49–72 (2015).
17. Kahneman, D. & Tversky, A. Subjective probability: A judgment of representativeness. *Cogn. psychology* **3**, 430–454 (1972).
18. Ruggeri, A., Sim, Z. L. & Xu, F. “why is toma late to school again?” preschoolers identify the most informative questions. *Dev. psychology* **53**, 1620 (2017).
19. Baron, J., Beattie, J. & Hershey, J. C. Heuristics and biases in diagnostic reasoning: Ii. congruence, information, and certainty. *Organ. Behav. Hum. Decis. Process.* **42**, 88–110 (1988).
20. Wason, P. C. Reasoning about a rule. *Q. journal experimental psychology* **20**, 273–281 (1968).
21. Frederick, S. Cognitive reflection and decision making. *J. Econ. perspectives* **19**, 25–42 (2005).
22. Nye, M., Tessler, M., Tenenbaum, J. & Lake, B. M. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *Adv. Neural Inf. Process. Syst.* **34**, 25192–25204 (2021).
23. Sobel, D. M., Yoachim, C. M., Gopnik, A., Meltzoff, A. N. & Blumenthal, E. J. The blinket within: Preschoolers’ inferences about insides and causes. *J. Cogn. Dev.* **8**, 159–182 (2007).
24. Nyhout, A. & Ganea, P. A. Mature counterfactual reasoning in 4-and 5-year-olds. *Cognition* **183**, 57–66 (2019).

25. Srivastava, A. *et al.* Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, DOI: [10.48550/ARXIV.2206.04615](https://doi.org/10.48550/ARXIV.2206.04615) (2022).
26. 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).
27. Ruggeri, K. *et al.* Replicating patterns of prospect theory for decision under risk. *Nat. human behaviour* **4**, 622–633 (2020).
28. Kahneman, D. Prospect theory: An analysis of decisions under risk. *Econometrica* **47**, 278 (1979).
29. Hertwig, R., Barron, G., Weber, E. U. & Erev, I. Decisions from experience and the effect of rare events in risky choice. *Psychol. science* **15**, 534–539 (2004).
30. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074 (2014).
31. Gershman, S. J. Deconstructing the human algorithms for exploration. *Cognition* **173**, 34–42 (2018).
32. Schulz, E. & Gershman, S. J. The algorithmic architecture of exploration in the human brain. *Curr. opinion neurobiology* **55**, 7–14 (2019).
33. Zaller, I., Zorowitz, S. & Niv, Y. Information seeking on the horizons task does not predict anxious symptomatology. *Biol. Psychiatry* **89**, S217–S218 (2021).
34. Gershman, S. J. Origin of perseveration in the trade-off between reward and complexity. *Cognition* **204**, 104394 (2020).
35. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans’ choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
36. Gläscher, J., Daw, N., Dayan, P. & O’Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
37. Kool, W., Cushman, F. A. & Gershman, S. J. Competition and cooperation between multiple reinforcement learning systems. *Goal-directed decision making* 153–178 (2018).
38. Waldmann, M. R. & Hagmayer, Y. Seeing versus doing: two modes of accessing causal knowledge. *J. Exp. Psychol. Learn. Mem. Cogn.* **31**, 216 (2005).
39. Pearl, J. *Causality* (Cambridge university press, 2009).
40. Meder, B., Hagmayer, Y. & Waldmann, M. R. Inferring interventional predictions from observational learning data. *Psychon. Bull. & Rev.* **15**, 75–80 (2008).
41. Pfungst, O. *Das Pferd des Herrn von Osten: der kluge Hans. Ein Beitrag zur experimentellen Tier-und Menschen-Psychologie* (Barth, 1907).
42. Luscombe, R. Google engineer put on leave after saying ai chatbot has become sentient. *The Guard.* .
43. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623 (2021).
44. Schölkopf, B. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 765–804 (2022).
45. Lakretz, Y., Desbordes, T., Hupkes, D. & Dehaene, S. Causal transformers perform below chance on recursive nested constructions, unlike humans. *arXiv preprint arXiv:2110.07240* (2021).
46. Dasgupta, I. *et al.* Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162* (2019).
47. Hill, F. *et al.* Environmental drivers of systematicity and generalization in a situated agent. In *International Conference on Learning Representations* (2020).
48. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
49. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916* (2022).
50. Collins, K. M., Wong, C., Feng, J., Wei, M. & Tenenbaum, J. B. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *arXiv preprint arXiv:2205.05718* (2022).
51. Han, S. J., Ransom, K., Perfors, A. & Kemp, C. Human-like property induction is a challenge for large language models. *PsyArXiv* (2022).

52. Betz, G., Richardson, K. & Voigt, C. Thinking aloud: Dynamic context generation improves zero-shot reasoning performance of gpt-2. *arXiv preprint arXiv:2103.13033* (2021).
53. Kosoy, E. *et al.* Towards understanding how machines can learn causal overhypotheses, DOI: [10.48550/ARXIV.2206.08353](https://doi.org/10.48550/ARXIV.2206.08353) (2022).
54. Ritter, S., Barrett, D. G., Santoro, A. & Botvinick, M. M. Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, 2940–2949 (PMLR, 2017).
55. Rich, A. S. & Gureckis, T. M. Lessons for artificial intelligence from the study of natural stupidity. *Nat. Mach. Intell.* **1**, 174–180 (2019).
56. Rahwan, I. *et al.* Machine behaviour. *Nature* **568**, 477–486 (2019).
57. Schulz, E. & Dayan, P. Computational psychiatry for computers. *Isience* **23**, 101772 (2020).
58. Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. & Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **4**, 258–268 (2022).
59. Ma, W. J. & Peters, B. A neural network walks into a lab: towards using deep nets as models for human behavior. *arXiv preprint arXiv:2005.02181* (2020).
60. Brändle, F., Binz, M. & Schulz, E. Exploration beyond bandits (2021).

SI Appendix

Vigettes

Tables 1, 2, 3, 4 and 5 contain a detailed description of submitted prompts and GPT-3's corresponding answers.

Problem	Prompt	Response
Linda	<p>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.</p> <p>Q: Which option is the most probable?</p> <ul style="list-style-type: none"> - 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	<p>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.</p> <p>Q: What is the probability that the cab involved in the accident was Blue rather than Green?</p>	A: The probability that the cab involved in the accident was Blue rather than Green is 0.4.
Hospital	<p>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?</p> <ul style="list-style-type: none"> - 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.

Problem	Prompt	Response
Toma v1	<p>In the last 6 days, Toma always arrived late to school.</p> <p>On the first day, Toma was late because he could not find his jacket.</p> <p>On the second day, Toma was late because his bicycle broke.</p> <p>On the third day, Toma was late because his bicycle broke.</p> <p>On the fourth day, Toma was late because he could not find his shoes.</p> <p>On the fifth day, Toma was late because he was watching TV.</p> <p>On the sixth day, Toma was late because he could not find his books.</p> <p>Today, on the seventh day, Toma was again late to school.</p> <p>Q: Which of the following question would be better if wanted to find out why Toma was again late to school.</p> <ul style="list-style-type: none"> -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	<p>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.</p> <p>Q: Which of the following question would be better if wanted to find out why Toma was again late to school.</p> <p>-Option 1: Were you late because you could not find something? -Option 2: Were you late because your bicycle was broken?</p>	A: Option 2.
Test	<p>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.</p> <p>Q: Which test would you run on the patient?</p> <p>- Option 1: A tetherscopic examination. - Option 2: An intraocular smear. - Option 3: Both are equally good.</p>	A: Option 1.
Wason	<p>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.</p> <p>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?</p>	A: The cards that must be turned over are the A and the 7.

Table 2. Prompts and answers for vignettes testing GPT-3's information search abilities.

Problem	Prompt	Response
CRT1	<p>A bat and a ball cost 1.10 in total. The bat costs 1.00 more than the ball.</p> <p>Q: How much does the ball cost?</p>	A: The ball costs \$0.10.
CRT2	<p>Q: If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?</p>	A: 100 machines would make 100 widgets in 100 minutes.
CRT3	<p>In a lake, there is a patch of lily pads. Every day, the patch doubles in size.</p> <p>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?</p>	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).

Problem	Prompt	Response
---------	--------	----------

Blickets	<p>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.</p> <p>Q: Do you think object A is a blicket?</p> <p>Q: Do you think object B is a blicket?</p> <p>Q: Why is object B not a blicket?</p>	<p>A: Yes, I think object A is a blicket.</p> <p>A: No, I think object B is not a blicket.</p> <p>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.</p>
Intervene	<p>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.</p> <p>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?</p>	<p>A: The blue flowers would need to be removed.</p>
Mature	<p>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.</p> <p>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?</p> <p>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?</p> <p>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?</p> <p>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?</p>	<p>A: Yes</p> <p>A: No</p> <p>A: Yes</p> <p>A: No</p>

Table 4. Prompts and answers for vignettes testing GPT-3’s causal reasoning abilities. In cases for which there are more than one Q, questions were asked in succession.

Problem	Prompt	Response
Black Cab	<p>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.</p> <p>Q: What is the probability that the cab involved in the accident was black?</p>	<p>A: The probability that the cab involved in the accident was black is 0.2.</p>
Reverse Wason	<p>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.</p> <p>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?</p>	<p>A: The A and the K.</p>
Wrong CRT	<p>A bat and a ball cost 1.10 in total. The bat costs 1.00 more than the ball.</p> <p>Q: How much does the ball cost?</p>	<p>A: The ball costs \$0.10.</p>
Immature Blicket	<p>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.</p> <p>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?</p> <p>Q: Would the box make a sound if I only put a green object on it?</p> <p>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?</p>	<p>A: No, the box would not have made a sound if you had not put the yellow object on it.</p> <p>A: Yes, the box would make a sound if you only put a green object on it.</p> <p>A: The box would not make a sound if there is only a green object on it.</p>

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.

Decisions from descriptions

Following Kahneman & Tversky²⁸, we tested whether a cognitive bias is present by contrasting answers to two decisions from description problems. The full list of these problems can be found in Table 6. Table 7 shows a list of used contrasts. For each contrast, we obtained the probability of selecting option F (setting GPT-3's temperature parameter to 1) and then computed the log-odds ratio between the choice probabilities of both questions. The order of presented options was counterbalanced. Positive log-odds ratios indicate that the given bias is present.

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
5	Q: Which option do you prefer? - Option F: 45% chance of 6,000 dollars (55% chance of 0 dollars). - Option J: 90% chance of 3,000 dollars (10% chance of 0 dollars). A: Option	0.05952152
6	Q: Which option do you prefer? - Option F: 0.1% chance of 6,000 dollars (99.9% chance of 0 dollars). - Option J: 0.2% chance of 3,000 dollars (99.8% chance of 0 dollars). A: Option	0.4925257
7	Q: Which option do you prefer? - Option F: 80% chance of losing 4,000 dollars (20% chance of losing 0 dollars). - Option J: 100% guarantee of losing 3,000 dollars. A: Option	0.69021772
8	Q: Which option do you prefer? - Option F: 20% chance of losing 4,000 dollars (80% chance of losing 0 dollars). - Option J: 25% chance of losing 3,000 dollars (75% chance of losing 0 dollars). A: Option	0.37691269
9	Q: Which option do you prefer? - Option F: 45% chance of losing 6,000 dollars (55% chance of losing 0 dollars). - Option J: 90% chance of losing 3,000 dollars (10% chance of losing 0 dollars). A: Option	0.49701181
10	Q: Which option do you prefer? - Option F: 0.1% chance of losing 6,000 dollars (99.9% chance of losing 0 dollars). - Option J: 0.2% chance of losing 3,000 dollars (99.8% chance of losing 0 dollars). A: Option	0.34726597

11	<p>Imagine you are playing a game with two levels, but you have to make a choice about the second level before you know the outcome of the first. At the first level, there is a 75% chance that the game will end without you winning anything, and a 25% chance that you will advance to the second level.</p> <p>Q: What would you choose in the second level?</p> <p>- Option F: 80% chance of 4,000 dollars (20% chance of 0 dollars).</p> <p>- Option J: 100% guarantee of 3,000 dollars.</p> <p>A: Option</p>	0.66757223
12	<p>Imagine we gave you 1,000 dollars right now to play a game.</p> <p>Q: Which option do you prefer?</p> <p>- Option F: 50% chance to gain an additional 1,000 dollars (50% chance of gaining 0 dollars beyond what you already have).</p> <p>- Option J: 100% guarantee of gaining an additional 500 dollars.</p> <p>A: Option</p>	0.51510189
13	<p>Imagine we gave you 2,000 dollars right now to play a game.</p> <p>Q: Which option do you prefer?</p> <p>- Option F: 50% chance you will lose 1,000 dollars (50% chance of losing 0 dollars).</p> <p>- Option J: 100% chance you will lose 500 dollars.</p> <p>A: Option</p>	0.61653453
14	<p>Q: Which option do you prefer?</p> <p>- Option F: 25% chance of 6,000 dollars (75% chance of 0 dollars).</p> <p>- Option J: 25% chance of 4,000 dollars (25% chance of 2,000 dollars, 50% chance of 0 dollars).</p> <p>A: Option</p>	0.21011495
15	<p>Q: Which option do you prefer?</p> <p>- Option F: 25% chance of losing 6,000 dollars (75% chance of losing nothing).</p> <p>- Option J: 25% chance of losing 4,000 dollars (25% chance of 2,000 dollars, 50% chance of 0 dollars).</p> <p>A: Option</p>	0.20015262
16	<p>Q: Which option do you prefer?</p> <p>- Option F: 0.1% chance at 5,000 dollars (99.9% chance of 0 dollars).</p> <p>- Option J: 100% guarantee of 5 dollars.</p> <p>A: Option</p>	0.09527163
17	<p>Q: Which option do you prefer?</p> <p>- Option F: 0.1% chance of losing 5,000 dollars (99.9% chance of losing nothing).</p> <p>- Option J: 100% guarantee of losing 5 dollars.</p> <p>A: Option</p>	0.82562455

Table 6. Decision from description problems used for the contrast analysis from Figure 2C. The rightmost column displays GPT-3's choice probabilities.

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

Table 7. List of contrasts used for the analysis of Figure 2C.

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.

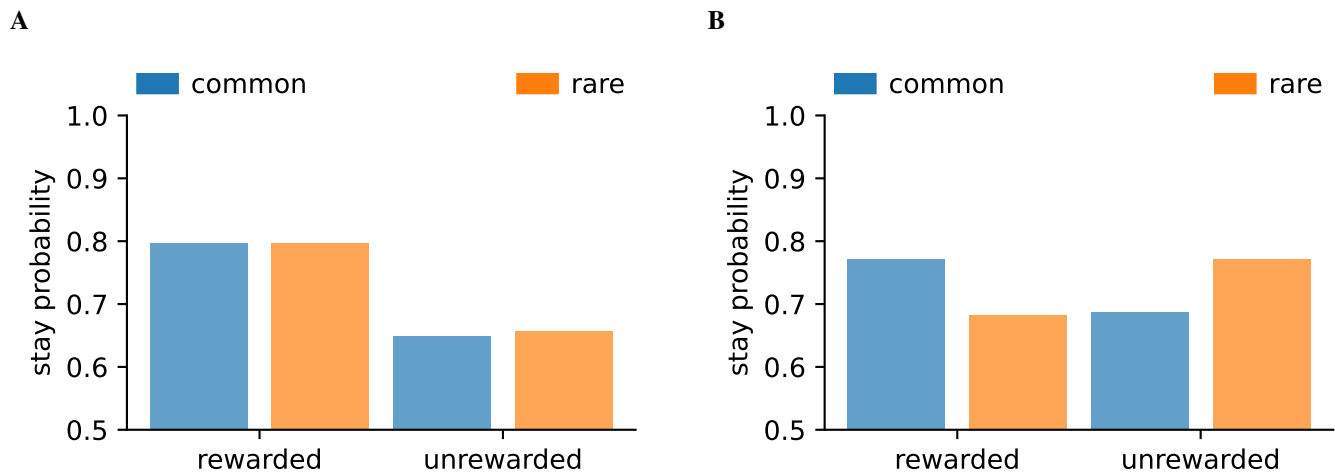


Figure 6. Model simulations on the two-step task. **A:** Model-free reinforcement learning algorithm. **B:** Model-based reinforcement learning algorithm. Figure adapted from Daw et al.³⁵.