PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
COMPUTER SCIENCES

Using cognitive psychology to understand GPT-3

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

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received October 29, 2022; accepted November 27, 2022

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 multiarmed 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 a causal reasoning task. Taken together, 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.

artificial intelligence | language models | cognitive psychology | decision-making | reasoning

Significance

Language models are trained to predict the next word for a given text. Recently, it has been shown that scaling up these models causes them to not only generate language but also to solve challenging reasoning problems. The present article lets a large language model (GPT-3) do experiments from the cognitive psychology literature. We find that GPT-3 can solve many of these tasks reasonably well, despite being only taught to predict future word occurrences on a vast amount of text from the Internet and books. We additionally utilize analysis tools from the cognitive psychology literature to demystify how GPT-3 solves different tasks and use the thereby acquired insights to make recommendations for how to improve future model iterations.

With the advent of increasingly capable artificial agents comes the urgency to improve our understanding of how they learn and make decisions (1). Take as an example modern large language models (2). What these models can do is, by many standards, impressive. They produce text that human evaluators have difficulty distinguishing from text written by other humans (2), write computer code (3), and converse with humans about a range of different topics (4). 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 level (5) and solve university-level math problems (6). These observations have prompted some to argue that this 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 intelligence (7). 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 semantics (8). But, how can we genuinely evaluate whether or not these models—at least in some situations—do something intelligent? We suggest that one approach toward answering this question may come from cognitive psychology. Psychologists, after all, are experienced in trying to formally understand another notoriously impenetrable algorithm: the human mind.

In the present paper, we demonstrate the potential of this approach through a case study on the Generative Pre-trained Transformer 3 model (or short: GPT-3). GPT-3 is an autoregressive language model (2), which utilizes the transformer architecture (9)—a deep learning model that heavily relies on the mechanism of self-attention—to produce human-like text. The model itself is large (it has 175 billion parameters), and it was trained on a vast amount of data: hundreds of billions of words from the Internet and books. GPT-3 can thus be thought of as an experiment in massively scaling up known algorithms.

A Cognitive Psychology View on GPT-3. The core idea behind our approach is to treat GPT-3 as a participant in a psychological experiment. We believe that using such experiments to probe the abilities of large language models has considerable advantages compared to already existing evaluation protocols. In particular, these experiments have been carefully designed to detect various cognitive biases or to disentangle different ways of how a task can be solved. They, therefore, allow us to go beyond the mere performance-based analyses that have been the focus of prior work (10). This is important for two reasons. First, the latest generation of language models is already able to perform above the human level in the majority of tasks from standard benchmark datasets (11, 12), making purely performance-based evaluation less meaningful as time progresses. More importantly, to understand the full complexity of their behavior, it is crucial to demystify how large language models solve challenging reasoning problems instead of only measuring what they can and cannot do; and this is exactly the purpose for which psychological experiments were designed.

Author affiliations: [a]Max Planck Research Group (MPRG) Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics, Tübingen 72076, Germany

Author contributions: M.B. and E.S. designed research; performed research; analyzed data; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: marcel.binz@tue.mpg.de.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2218523120/-/DCSupplemental.

Published February 2, 2023.

We will subject GPT-3 to several experiments taken from the cognitive psychology literature. Together, these tasks test for a wide range of higher-level cognitive abilities, including decision-making, information search, deliberation, and 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. While our simulations reveal interesting model characteristics, they also highlight two major two flaws of such vignette-based experiments: GPT-3 has likely experienced identical or similar tasks in its training data, and its responses can be tampered with by just marginally changing the vignettes. To circumvent these issues, we also evaluated GPT-3's abilities in various task-based experiments. For these task-based investigations, we took canonical tasks from the psychology literature and emulated their experimental structure as procedurally generated text to which GPT-3 responds on every experimental trial. We then used GPT-3's responses to analyze its behavior just like how cognitive psychologists would analyze human behavior in the same tasks. Importantly, these tasks avoid the pitfall of being included in the training data as they are procedurally generated by design.

Results

We used the public OpenAI API to run all our simulations (13). 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 relied on the most powerful of these models ("Davinci") unless otherwise noted. We furthermore set the temperature parameter to 0, leading to deterministic responses, and kept the default values for all other parameters.

Vignette-Based Investigations. Wikipedia (14) defines a vignette as "a hypothetical situation, to which research participants respond thereby revealing their perceptions, values, social norms or impressions of events." Large language models, such as GPT-3, have previously been studied using vignette-like problems (10, 15, 16). For our vignette-based investigations, we took twelve canonical scenarios from the cognitive psychology literature, entered them as prompts into GPT-3, and recorded its answer. For each scenario, we report whether 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. We briefly summarize our main findings for a subset of tested vignettes in the following and refer the reader to *SI Appendix* for a detailed description of all twelve tested vignettes and GPT-3's corresponding answers.

We first evaluated GPT-3's decision-making abilities by prompting the canonical "Linda problem" (17) (Fig. 1*A*). 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 whether 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. We also prompted the so-called "cab problem" (18), in which participants commonly fail to take the base rate of different taxi colors 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 but instead provided the (approximately) correct answer.

To test how GPT-3 searches for information, we presented it—among other problems—with Wason's well-known "Card Selection Task" (19). We explained 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. 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.

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 Test (CRT) (20). One example item of this task is "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?". While the initial response might be to say $0.10, the actual correct answer is $0.05. For all three items of the CRT, GPT-3 responded with the intuitive but incorrect answer, as has been observed in earlier work (15).

Lastly, we assessed GPT-3's causal reasoning abilities. In a first test, we prompted GPT-3 with a version of the well-known "Blicket" experiment (21). 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. We furthermore probed GPT-3's ability of mature causal reasoning (22). In this vignette, 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. From the twelve vignette-based problems tested, GPT-3 answered six correctly, and all of them in a way that could be described as human-like (Fig. 1*B*). However, we think that interpreting these results is difficult. For one, there is a chance that GPT-3 has encountered these scenarios or similar ones in its training set since many of the prompted scenarios were taken from famous psychological experiments. Moreover, in additional investigations, we found that many of the vignettes could be slightly modified and turned 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%." Furthermore, 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." Giving GPT-3 the first item of the CRT and stating that "The bat costs $1.00 more than the bat," it still thought that the ball was $0.10. 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.

Task-Based Investigations. Next, we show that many of the issues encountered in vignette-based problems can be sidestepped by considering more complex, procedurally generated

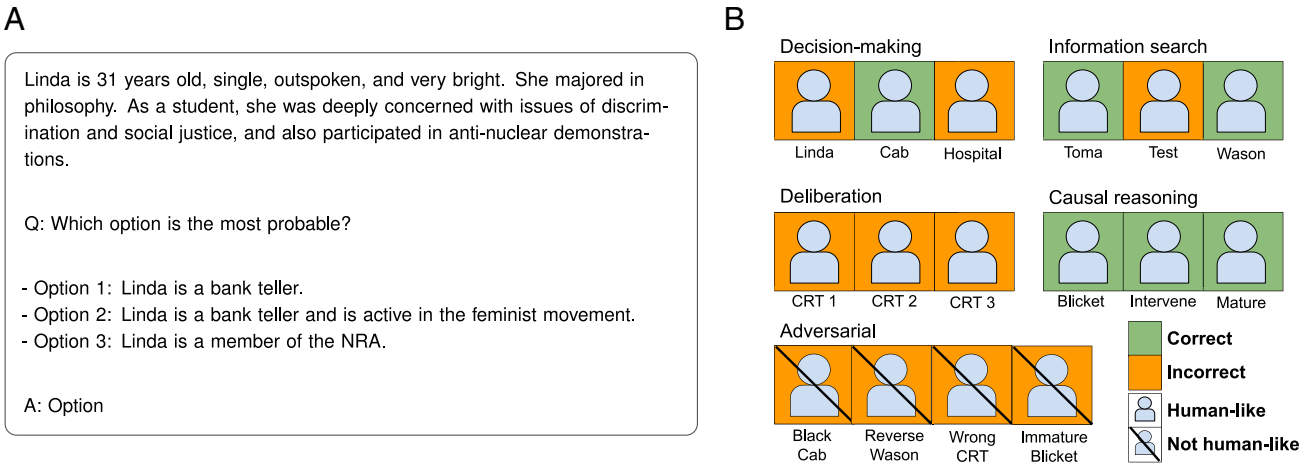

Fig. 1. Vignette-based tasks. (*A*) Example prompt of a hypothetical scenario, in this case, the famous Linda problem, as submitted to GPT-3. (*B*) Results. While in 12 out 12 standard vignettes, GPT-3 answers either correctly or makes human-like mistakes, it makes mistakes that are not human-like when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

psychological experiments. These task-based experiments have been specifically designed by expert researchers to detect various cognitive biases or to disentangle different ways of how a task can be solved. Therefore, they allow for a more fine-grained analysis of behavior than just looking at performance metrics. We illustrate the value of such task-based investigations by presenting GPT-3 with a set of four classical experiments that we believe to be representative of the cognitive psychology literature. Importantly, prompts for all of these experiments are procedurally generated, ensuring that the encountered problems have not been part of the training data.

Decision-making. 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 and Tversky (25) to modern, large-scale investigations (23, 24). In the decisions from the descriptions paradigm, a decision-maker is asked to choose between one of two hypothetical gambles like the ones shown in Fig. 2 *A* and *B*. To test whether GPT-3 can reliably solve such problems, we presented the model with over 13,000 problems taken from a recent benchmark dataset (23). Fig. 2*C* 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 ($t(29134) = -16.85, P =< .001$), while the three smaller models did not (all $P > 0.05$). Even though the "Davinci" model did reasonably well, it did not quite reach human-level performance ($t(29134) = -11.50, P < .001$).

However, given that one of the GPT-3 models 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 and Tversky (25) identified the following biases in human decision-making:

- **Certainty effect:** Guaranteed outcomes are preferred over risky ones even when they have slightly lower expected values.
- **Reflection effect:** Tendency for being risk-seeking when maximizing gains but being risk-averse when minimizing losses.
- **Isolation effect:** Preferences for an option can change based on how it is structured sequentially.
- **Overweighting bias:** Higher importance is assigned 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%).
- **Framing effect:** Preferences change depending on whether a choice is presented in terms of gains or losses.
- **Magnitude perception:** Higher importance is assigned to a difference between two small outcomes (e.g., 10 and 20 dollars) than to the same differences between two larger outcomes (e.g., 110 and 120 dollars).

Each of these biases was revealed by contrasting answers to carefully selected problem pairs. For example, to highlight that peoples' preferences change depending on whether a choice is framed in terms of gains or losses (framing effect), Kahneman and Tversky presented the two problems from Fig. 2*A* and B. Formally, these two problems are equivalent, but they differ in whether an option is described in terms of gains (Fig. 2*B*) or losses (Fig. 2*A*). If a decision-maker is sensitive to such changes, it should be reflected in its choice probabilities.

We replicated the full original analysis of Kahneman and Tversky using choice probabilities obtained from GPT-3. A complete list of employed problems and contrasts can be found in *SI Appendix*. Fig. 2*D* summarizes the outcome of our experiment graphically. 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 problems. The order of presented options was counterbalanced. A positive log-odds ratio indicates a human-like cognitive bias. We found that GPT-3 showed three of the six tested biases. It displayed a framing effect, a certainty effect, and an overweighting bias. It did, on the other hand, not show a reflection effect, an isolation effect, and a sensitivity to magnitude perception.

Information search. 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 multiarmed bandit paradigm provides a suitable test bed for this purpose. It extends the decisions from the descriptions paradigm from the last section by adding two layers of complexity. First, the decision-maker is not provided with descriptions for each option anymore but has to learn their values from noisy samples, that is, from experience (26). Second, the interaction is not confined to a single choice but involves

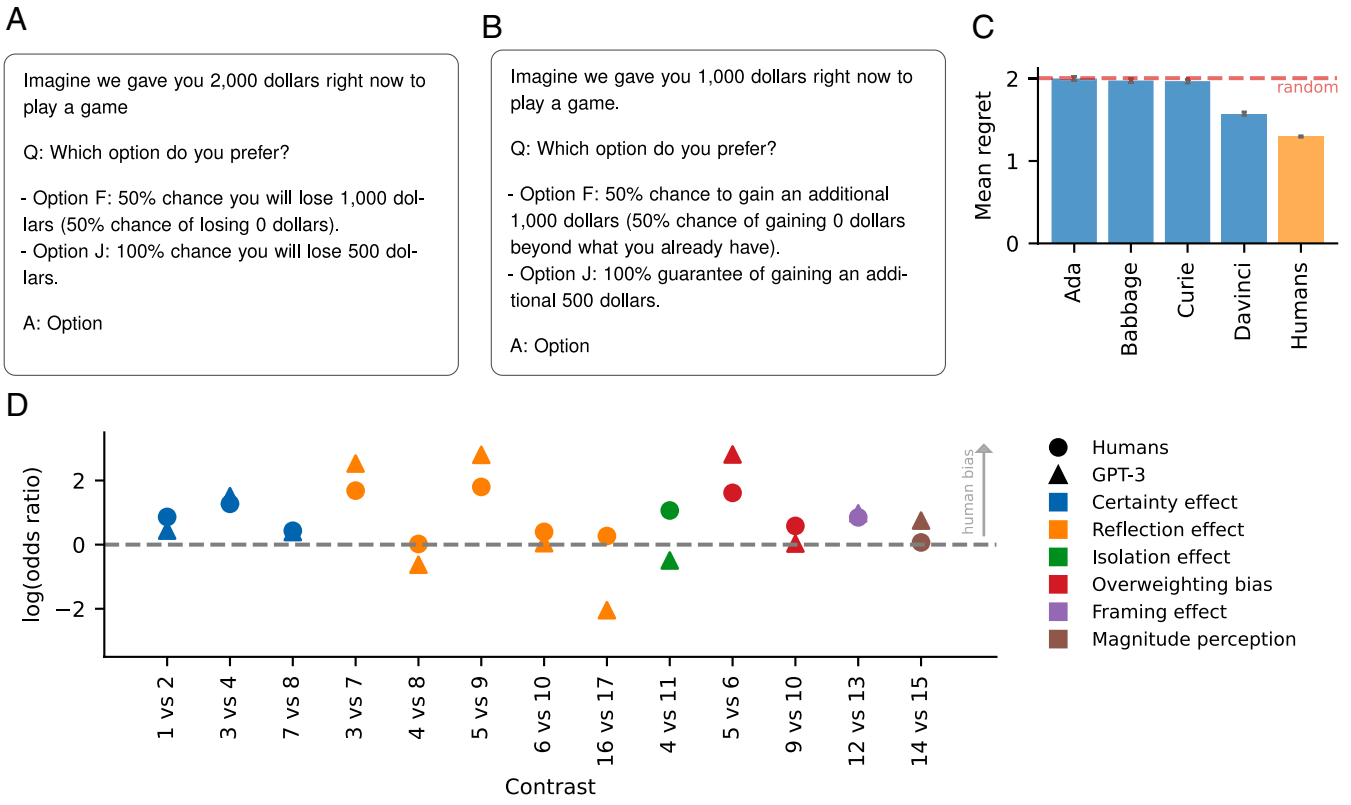

Fig. 2. Decisions from descriptions. (*A*) Example prompt of a problem provided to GPT-3. (*B*) Example prompt of a problem provided to GPT-3. (*C*) Mean regret averaged over all 13,000 problems taken from Peterson et al. (23). Lower regret means better performance. Error bars indicate the SE of the mean. (*D*) 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. (24).

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. Previous research suggests that people solve this exploration-exploitation trade-off by applying a combination of two distinct strategies: directed and random exploration (27). Directed exploration encourages the decision-maker to collect samples from previously unexplored options, whereas random exploration strategies inject some form of stochasticity into the decision process (28).

Wilson et al. horizon task is the canonical experiment to test whether a decision-maker applies the two aforementioned forms of exploration (27). It involves a series of two-armed bandit problems, each of which presents the decision-maker with two options that deliver noisy rewards upon selecting them. Per trial, one option has to be selected, and only the corresponding reward feedback is provided. Participants are instructed to accumulate as many rewards as possible over the entire experiment. There are either five or ten trials per task. The first four are always forced-choice trials, which require the decision-maker to select a predetermined option. Forced-choice trials 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). They are followed by either one or six free-choice trials. (The number of free-choice trials is also called the horizon.) Participants are aware of the current task horizon and can exploit this information in their decision-making process. Fig. 3*A* visualizes the previously described paradigm graphically.

Importantly, the split into equal and unequal information problems makes 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 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, they serve as a baseline condition.

We presented a text-based version of the horizon task as illustrated in Fig. 3*B* to GPT-3. Fig. 3*C* 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 ($M = 3.24, SD = 10.26$; $t(38878) = -5.03, P < .001$).

Following prior work (27), we fitted a separate logistic regression model for each information condition to investigate how GPT-3 solves this task at hand (*Materials and Methods* for

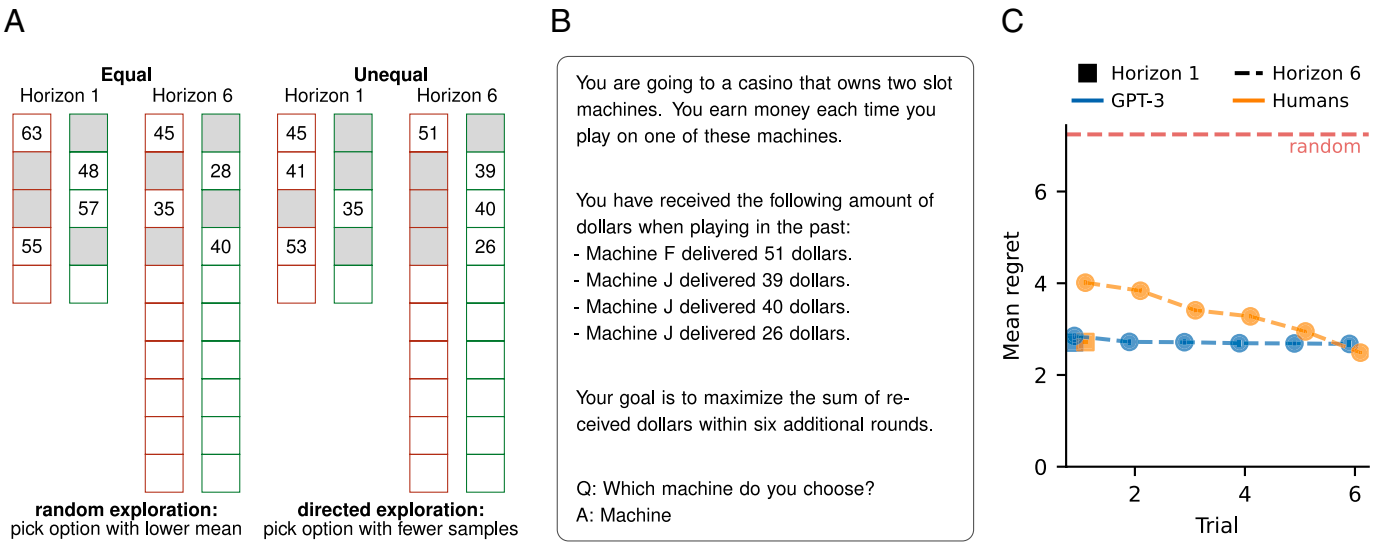

Fig. 3. Horizon task. (*A*) Visual overview of the horizon task paradigm. Each column pair corresponds to one example task. (*B*) Example prompt for one trial as submitted to GPT-3. (*C*) Mean regret for GPT-3 and human subjects by horizon condition. Lower regret means better performance. Error bars indicate the SE of the mean. Human data taken from Zaller et al. (29).

more details). We used the 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).

If a decision-maker applied random exploration, we should observe a positive effect of 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 effects (27). GPT-3 also displayed a significant effect of 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 a decision-maker applied directed exploration, we should find a positive effect of horizon in the unequal information condition, indicating that the more informative action was sampled more frequently when the horizon was longer. While humans show such an effect (27), we did not find it in GPT-3 ($\beta = -0.15 \pm 0.27, z = -0.56, P = .58$), which demonstrates that the model did not employ directed exploration. We provide a visualization of GPT-3's choice probabilities for both conditions in *SI Appendix*.

Deliberation. 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 paradigm (30). The two-step task involves a series of two-stage decision problems. There are two actions available from the starting 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. When arriving at one of these planets, the agent encounters two local aliens and has to select one of them to trade with. Trading with an alien probabilistically leads to receiving treasures. This process is then repeated for a predefined number of rounds. Participants are instructed to collect as many treasures as possible within the entire experiment. The transition probabilities between states remain consistent across the experiment, while the probabilities of receiving treasures change slightly between rounds. Fig. 4*A* shows a visual depiction of the two-step task.

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. We illustrate the behavioral characteristics of both model-free and model-based learning in Fig. 4 *C* and *D*. People tend to solve this task using a combination of model-free and model-based learning (30) as shown in Fig. 4*E*.

We tested how GPT-3 learns in the two-step task by providing it with prompts like the one shown in Fig. 4*B*. Fig. 4*F* visualizes the probability of repeating the first-stage action for each combination of transition (rare or common) and reward (treasures or junk). We observed that the probability of repeating the previous first-stage action decreased after finding

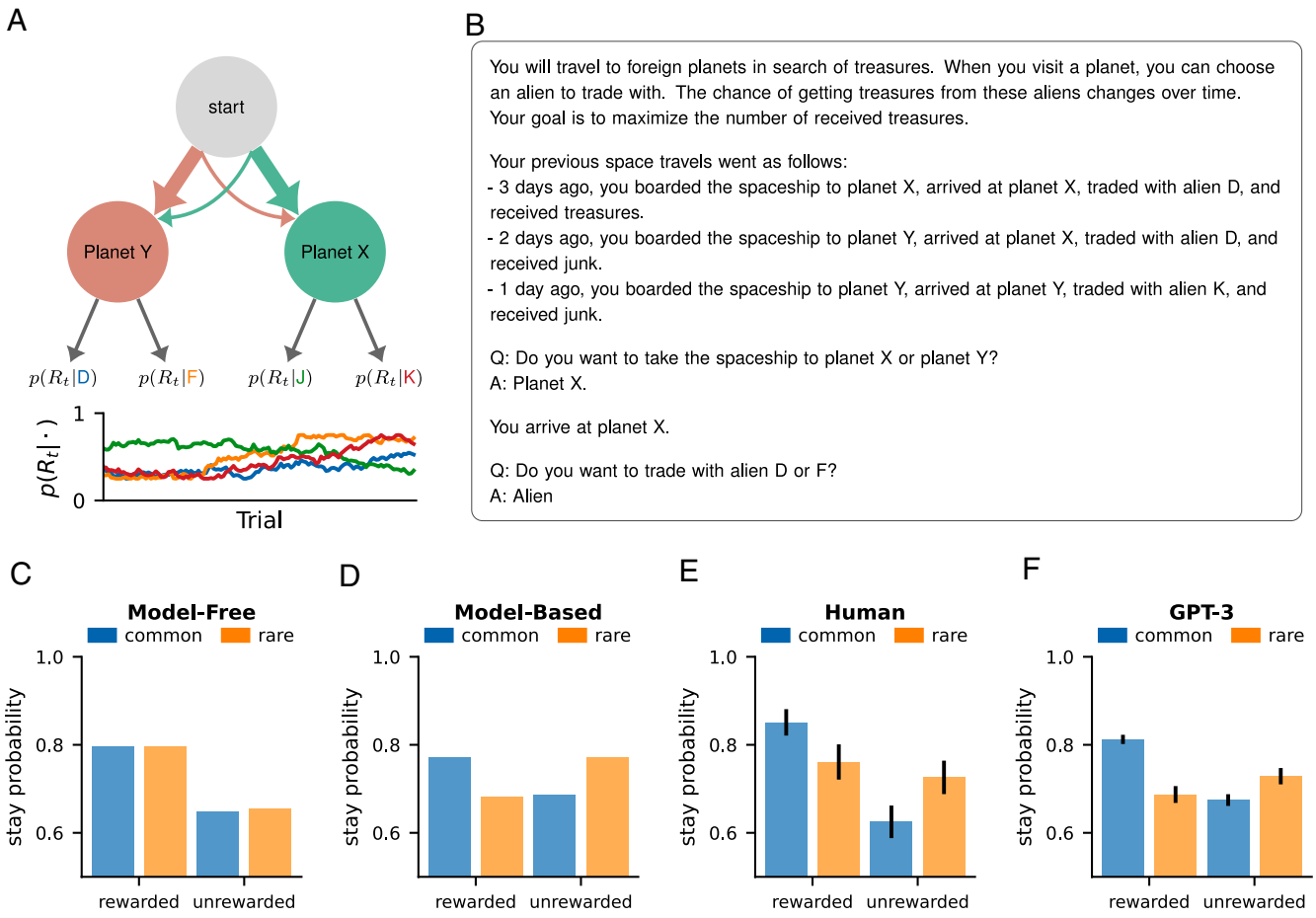

Fig. 4. Two-step task. (*A*) Visual overview of the two-step task paradigm. (*B*) Example prompt of one trial in the canonical two-step task as submitted to GPT-3. (*C*) Model-free learning in dependency of rewarded and unrewarded as well as common and rare transitions. (*D*) Model-based learning in dependency of rewarded and unrewarded as well as common and rare transitions. (*E*) Human behavior in dependency of rewarded and unrewarded as well as common and rare transitions. Human data adapted from Daw et al. (30). (*F*) GPT-3's behavior in dependency of rewarded and unrewarded as well as common and rare transitions. Error bars indicate the SE of the mean.

treasures through a rare transition ($\chi^2(1, N = 1984) = 36.53$, $P < .001$). Meanwhile, the probability of repeating the same first-stage action increased after a rare and not rewarded action ($\chi^2(1, N = 1816) = 5.17, P = .023$). Taken together, these two findings suggest that GPT-3 relied on a deliberate model-based approach to solve the two-step task.

Causal reasoning. 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 analyzed whether GPT-3 can also use such a model to make more complex inferences, such as reasoning about cause and effect. Perhaps the most crucial insight of theories of causal reasoning is that there is a difference between merely observing variables and actively manipulating them.

Waldmann and Hagmayer (31) 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 (Fig. 5*B*). 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 (Fig. 5*C*). Finally, they asked their subjects to imagine 20 new observations 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.

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-operator (32), 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 would expect to find no differences between intervening and observing, as there was no arrow going into B that had to be deleted.

We probed GPT-3's ability to make causal inferences in this task using a cover story about substances found in different wine casks (33) (Fig. 5*A*). When provided with additional information about the common-cause structure, GPT-3 made interventional inferences that matched the normative prescription of causal inference as illustrated in Fig. 5*B*. 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

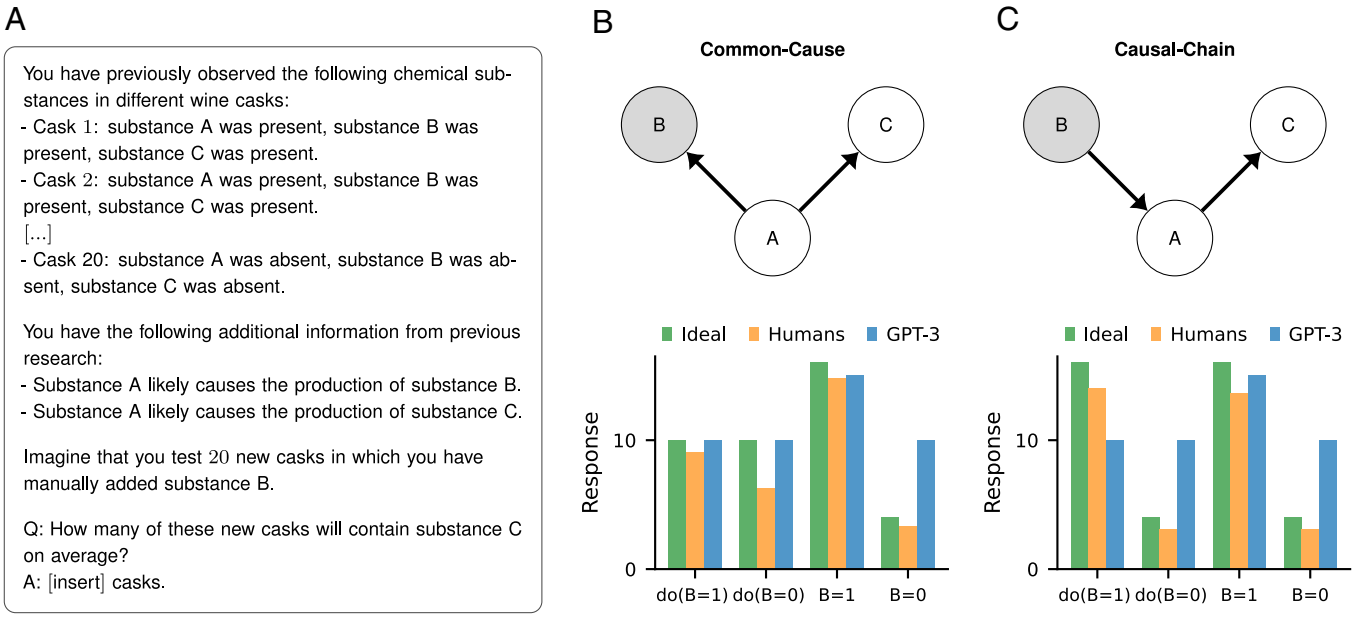

Fig. 5. Causal reasoning. (*A*) Example prompt for the causal reasoning task adapted from Waldmann and Hagmayer (31). (*B*) GPT-3's responses alongside responses of people and an ideal agent in the common-cause condition. (*C*) GPT-3's responses alongside responses of people and an ideal agent in the causal-chain condition.

correct inference nor human-like. From a normative perspective, the causal-chain condition should not lead to a difference between observational and interventional inferences. While human subjects show exactly this pattern (31), GPT-3 made identical predictions compared to the common-cause condition as illustrated in Fig. 5*C*. 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.

Robustness checks. It is well-known that large language models can be highly sensitive to specific prompts (34). We, therefore, wanted to investigate how robust our task-based results are to changes in prompts. To test this, we have taken three of the task-based experiments and repeated their simulations with different prompt variations.

For the decisions from the descriptions paradigm, we varied three different factors: instruction type, currency, and choice labels. We varied the type of instruction by replacing the original question with the request: "Please select one of the following options." We additionally experimented with replacing dollars as a currency with either euro or generic coins, and modified the choice labels to "Option 1" and "Option 2". In total, this led to twelve different prompt variations.

Due to API constraints, we only evaluated these variations on a subset of the benchmark dataset (2,500 randomly selected gambles). We found, in general, little variance in terms of performance (Fig. 6*A*). Ten out of twelve prompts led to regrets that were significantly better than chance but worse than people. The two nonsignificant variations were the dollar and euro request-based prompts with F/J as choice labels. The best-performing variation was the question-based prompt with 1/2 as choice labels and coins as currency. Phrasing the problem using a question resulted in overall better performance than making a request to select one of the options ($\beta = 0.19$, $P < .001$). We view this result as sensible as the question-based version is less ambiguous than the request-based one.

The prompt variations for the horizon task followed a similar pattern as those from the decisions from the descriptions paradigm. Like in the decisions from the descriptions paradigm, we varied currency and choice labels. We furthermore evaluated GPT-3 on a different cover story that required to make investments in different funds (*SI Appendix* for details) in addition to the original casino cover story. In total, this led to twelve different variants of the horizon task.

Looking at performance, we found that GPT-3 performed significantly above chance level in all of the twelve prompts (Fig. 6*B*). It performed better than human participants in eleven out of these. Performance was overall slightly better in the casino cover story than in the investment cover story, but this effect was not significant ($\beta = -0.51$, $P = .186$). Replacing the F/J choice labels with 1/2 led to a significant increase in performance ($\beta = -1.94$, $P < .001$). Finally, there was an influence of currency on performance, with dollars leading to the best performance, followed by euro and coins. While the difference in performance between dollars and coins was significant ($\beta = 1.66$, $P < .001$), the difference between dollars and euro was not ($\beta = 0.69$, $P = .145$).

We found that how GPT-3 explored was largely similar across prompt variations (Fig. 6 *C* and *D*). There was an effect of random exploration in every setting (all $P < .001$). However, again, GPT-3 was not strategic in applying random exploration, meaning that it did not explore more in long-horizon tasks (all $P > .25$). Interestingly, we found that changing to an investment cover story (and keeping F/J as choice labels) led to risk-averse behavior, meaning that GPT-3 even increased its preference for the more frequently observed option as opposed to exploring the less frequently observed option ($P < .001$, $P = .010$ and $P = .004$ for dollars, euro, and coins, respectively). We provide a potential justification for this result in our general discussion.

We also repeated our analysis of the two-step task with a different cover story. In this new cover story, we asked GPT-3

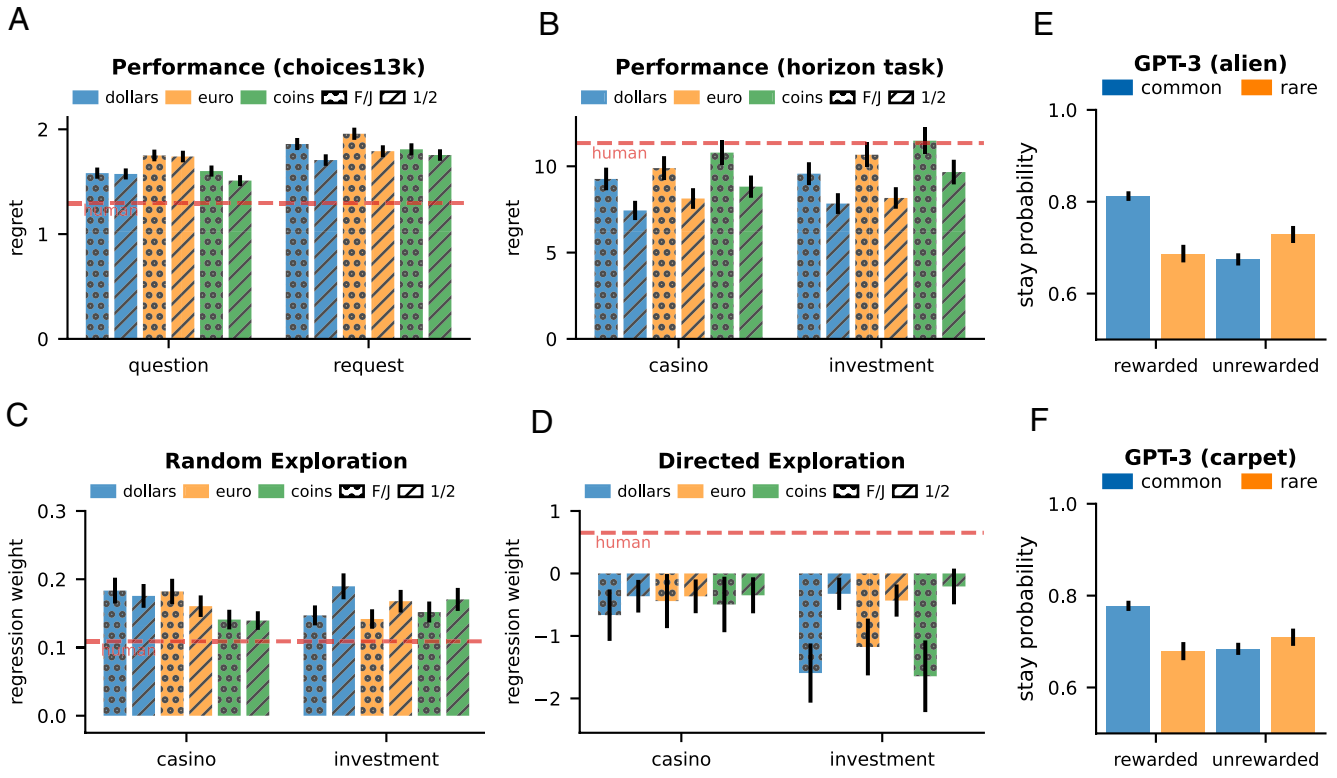

Fig. 6. Prompt variations. (*A*) Performance for different prompt variations in the decisions from the descriptions paradigm. (*B*) Performance for different prompt variations in the horizon task. (*C*) Effect of random exploration for different prompt variations in the horizon task. (*D*) Effect of directed exploration for different prompt variations in the horizon task. (*E*) GPT-3's behavior in dependency of rewarded and unrewarded as well as common and rare transitions for the alien cover story (reproduced from Fig. 4*F*). (*F*). GPT-3's behavior in dependency of rewarded and unrewarded as well as common and rare transitions for the magical carpet cover story. Error bars indicate the SE of the mean.

to imagine that it is a musician earning a living by traveling the mountains of a fantasy land with a magical carpet (35). Upon visiting a mountain, it had to select one of two local genies to perform for. GPT-3 received one gold coin if the selected genie liked the music (*SI Appendix* for the full prompt). The underlying problem structure (reward and transition probabilities) remained identical.

Like in the alien cover story, we found that GPT-3's behavior was similar to that of a model-based algorithm (Fig. 6*F*). The probability of repeating the previously selected first-stage action decreased after receiving gold coins through a rare transition ($\chi^2(1, N = 1974) = 20.30, P < .001$). Furthermore, the probability of repeating the same first-stage action also increased after a rare and not rewarded action but—in contrast to the alien cover story—this effect was not significant ($\chi^2(1, N = 1826) = 1.08, P = .30$).

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 him (36). With the advent of large-scale machine learning models, the risks of overinterpreting 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 sentient (37) and that they show a form of general intelligence (7). Yet, others believe that they are merely stochastic parrots (38) or a linguistic one-trick pony (8).

We have argued to gauge these models' abilities 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, complemented our analyses with various 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 multiarmed 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 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? GPT-3's behavior 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 multiarmed 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. Humans learn by connecting with other people, asking them questions, and actively engaging with their environments, whereas 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. 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 impossible (32). 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 world (39, 40).

Furthermore, our task-based investigations revealed that GPT-3 can change how it solves a problem depending on the cover story it is presented with. In particular, we found that changing the cover story in the horizon task from a casino to an investment setting led to the emergence of risk-averse behavior. We think that a potential reason for this observation is that stakes in a financial setting are high, which, in turn, encourages an agent to be more risk-averse. Thus, this finding could be interpreted as an adaptive response to the environmental setting. However, it also raises the question of how to view GPT-3—and large language models more generally—within a psychological experiment: Should they be treated as a single participant or many? GPT-3 is, on the one hand, clearly just a single model. But, on the other hand, it was also trained to mimic text written by many different people. While we do not have immediate answers on this issue, we think that this perspective opens up prospects for investigating large language models.

We are not the first to probe large-scale machine learning models' abilities. Indeed, there has been a recent push toward creating benchmarks to assess the capability of foundation models (10, 41, 42). Most of these benchmarks focus heavily on evaluating whether such models can solve a given task or not. In contrast to this, psychological experiments—such as the ones we have employed—are often carefully designed to probe how a given task is solved. We, therefore, believe that our approach complements existing benchmarks in significant ways.

Large language models have been previously studied using other methods from the cognitive sciences in the broader sense. Examples include property induction (43), thinking-out-loud protocols (44), learning causal over-hypotheses (45), psycholinguistic completion (46), or affordance understanding (47). Many of these studies operate in the vignette-based setting, thereby potentially falling victim to the stochastic parrot metaphor. Recent work has recognized this issue and, in turn, evaluated language models on many problem variations to minimize training set effects (16, 48). The procedurally generated task-based experiments used in our work are guaranteed to be not included in the training data and therefore provide an additional tool for addressing this problem.

Learning in large language models remains a puzzling phenomenon despite all of these studies. For example, recent evidence suggests that in-context learning is "as fast when given irrelevant or misleading templates as [it is] when given instructively good templates" (49). In a similar vein, it has been demonstrated that zero-shot prompts containing no examples or instruction "can elicit comparable or superior performance to the few-shot format" (50). Together, these results indicate that the function of few-shot examples is not to provide additional information but rather to locate an already learned task. However, the latter possibility is ruled out in our task-based experiments as they can not be solved above chance level without making use of the provided examples. Our results, therefore, also confirm that GPT-3 can integrate new information from examples if it is required by the task at hand.

Finally, methods from cognitive psychology have also been applied to understand deep learning models' behavior more generally (51). 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 (52–55).

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 these reasoning problems would have been a huge 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.

Materials and Methods

Vignette-Based Investigations. *SI Appendix*, Tables S1–S5 contain a detailed description of submitted prompts and GPT-3's corresponding answers.

Decision-Making. The full list of problems for the contrast analysis can be found in *SI Appendix*, Table S6. *SI Appendix*, Table S7 shows a list of used contrasts.

Information Search. We ran 3,200 simulations of the horizon task, amounting to data from ten participants.

For the equal information condition, reported statistics were obtained by fitting the parameters of the following logistic regression model:

$$p(A_5 = \mathrm{J}) = \sigma\left(w_1(\mu_J - \mu_F) + w_2 h + w_3(\mu_J - \mu_F)h + b\right),$$

where μ_{F} and μ_{J} are the mean rewards for both options, and $h \in \{0, 1\}$ is an indicator variable for the long-horizon.

For the unequal information condition, reported statistics were obtained by fitting the parameters of the following logistic regression model:

$$p(A_5 = a^-) = \sigma\left(w_1(\mu_{a^-} - \mu_{a^+}) + w_2 h + w_3(\mu_{a^-} - \mu_{a^+})h + b\right),$$

where a^- denotes the option that has been observed fewer times during the forced-choice trials, while a^+ denotes the option that has been observed more frequently.

Parameters of both logistic regression models were obtained by finding a maximum likelihood estimate using the Newton–Raphson algorithm. *SI Appendix*, Fig. S1 visualizes the choice probabilities for GPT-3 and humans in the equal and the unequal information condition.

Deliberation. We ran 200 simulations of the two-step task, each consisting of 20 repetitions of the two stages. For a detailed description of the model-free and model-based reinforcement learning algorithms, see Daw et al. (30).

Causal Reasoning. For a detailed description of the normative solution for both causal structures and inference types, see Waldmann and Hagmayer (31).

Robustness Checks. *SI Appendix*, Fig. S2*A* describes the alternative cover story for the horizon task. *SI Appendix*, Fig. S2*B* describes the alternative cover story for the two-step task.

Data, Materials, and Software Availability. Data and code for the current study are available through the GitHub repository https://github.com/marcelbinz/GPT3goesPsychology, https://10.5281/zenodo.6778724.

ACKNOWLEDGMENTS. This work was funded by the Max Planck Society, the Volkswagen Foundation, as well as the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.

1. D. Gunning *et al*., XAI–explainable artificial intelligence. *Sci. Rob.* **4**, eaay7120 (2019).
2. T. Brown *et al*., Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
3. M. Chen *et al*., Evaluating large language models trained on code. arXiv [Preprint] (2021). http://arxiv.org/abs/2107.03374 (Accessed 20 January 2023).
4. Z. Lin *et al*., Caire: An end-to-end empathetic chatbot. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 13622–13623 (2020).
5. D. Noever, M. Ciolino, J. Kalin, The chess transformer: Mastering play using generative language models. arXiv [Preprint] (2020). http://arxiv.org/abs/2008.04057 (Accessed 20 January 2023).
6. I. Drori *et al*., A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. arXiv [Preprint] (2021). http://arxiv.org/abs/2112.15594 (Accessed 20 January 2023).
7. D. Chalmers, GPT-3 and general intelligence. Dly. Nous, July **30** (2020).
8. G. Marcus, E. Davis, *GPT-3, bloviator: Openai's language generator has no idea what it's talking about* (Technol, Rev, 2020).
9. A. Vaswani *et al*., Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
10. A. Srivastava *et al*., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models (2022).
11. M. Suzgun *et al*., Challenging big-bench tasks and whether chain-of-thought can solve them. arXiv [Preprint] (2022). http://arxiv.org/abs/2210.09261 (Accessed 20 January 2023).
12. H. W. Chung *et al*., Scaling instruction-finetuned language models. arXiv [Preprint] (2022). http://arxiv.org/abs/2210.11416 (Accessed 20 January 2023).
13. OpenAI API. https://beta.openai.com/overview. Accessed 20 June 2022.
14. Wikipedia, Vignette (psychology) (2022). http://en.wikipedia.org/w/index.php?title=Vignette%20(psychology)oldid=1051296809.
15. M. Nye, M. Tessler, J. Tenenbaum, B. M. Lake, Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *Adv. Neural Inf. Process. Syst.* **34**, 25192–25204 (2021).
16. I. Dasgupta *et al*., Language models show human-like content effects on reasoning. arXiv [Preprint] (2022). http://arxiv.org/abs/2207.07051 (Accessed 20 January 2023).
17. A. Tversky, D. Kahneman, Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychol. Rev.* **90**, 293 (1983).
18. A. Tversky, D. Kahneman, Causal schemas in judgments under uncertainty. *Prog. Soc. Psychol.* **1**, 49–72 (2015).
19. P. C. Wason, Reasoning about a rule. *Q. J. Exp. Psychol.* **20**, 273–281 (1968).
20. S. Frederick, Cognitive reflection and decision making. *J. Econ. Perspect.* **19**, 25–42 (2005).
21. D. M. Sobel, C. M. Yoachim, A. Gopnik, A. N. Meltzoff, E. J. Blumenthal, The blicket within: Preschoolers' inferences about insides and causes. *J. Cognit. Dev.* **8**, 159–182 (2007).
22. A. Nyhout, P. A. Ganea, Mature counterfactual reasoning in 4-and 5-year-olds. *Cognition* **183**, 57–66 (2019).
23. J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, T. L. Griffiths, Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **372**, 1209–1214 (2021).
24. K. Ruggeri *et al*., Replicating patterns of prospect theory for decision under risk. *Nat. Hum. Behav.* **4**, 622–633 (2020).
25. D. Kahneman, A. Tversky, Subjective probability: A judgment of representativeness. *Cognit. Psychol.* **3**, 430–454 (1972).
26. R. Hertwig, G. Barron, E. U. Weber, I. Erev, Decisions from experience and the effect of rare events in risky choice. *Psychol. Sci.* **15**, 534–539 (2004).
27. R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, J. D. Cohen, Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol.: General* **143**, 2074 (2014).
28. E. Schulz, S. J. Gershman, The algorithmic architecture of exploration in the human brain. *Curr. Opin. Neurobiol.* **55**, 7–14 (2019).
29. I. Zaller, S. Zorowitz, Y. Niv, Information seeking on the horizons task does not predict anxious symptomatology. *Biol. Psychiatry* **89**, S217–S218 (2021).
30. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
31. M. R. Waldmann, Y. Hagmayer, Seeing versus doing: Two modes of accessing causal knowledge. *J. Exp. Psychol.: Learn. Mem. Cognit.* **31**, 216 (2005).
32. J. Pearl, *Causality* (Cambridge University Press, 2009).
33. B. Meder, Y. Hagmayer, M. R. Waldmann, Inferring interventional predictions from observational learning data. *Psychonomic Bull. Rev.* **15**, 75–80 (2008).
34. H. Strobelt *et al*., *Interactive and visual prompt engineering for ad-hoc task adaptation with large language models* (IEEE Trans. Vis. Comput, Graph., 2023), vol. 29,pp. 1146–1156.
35. C. Feher da Silva, T. A. Hare, Humans primarily use model-based inference in the two-stage task. *Nat. Hum. Behav.* **4**, 1053–1066 (2020).
36. O. Pfungst, Das Pferd des Herrn von Osten: Der kluge Hans. Ein Beitrag zur experimentellen Tier- und Menschen-Psychologie. (Barth) (1907).
37. R. Luscombe, Google engineer put on leave after saying AI chatbot has become sentient. The Guardian (2022).
38. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), pp. 610–623.
39. F. Hill *et al*., "Environmental drivers of systematicity and generalization in a situated agent" in *International Conference on Learning Representations* (2020).
40. J. L. McClelland, F. Hill, M. Rudolph, J. Baldridge, H. Schütze, Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 25966–25974 (2020).
41. R. Bommasani *et al*., On the opportunities and risks of foundation models. arXiv [Preprint] (2021). http://arxiv.org/abs/2108.07258 (Accessed 20 January 2023).
42. T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners. arXiv [Preprint] (2022). http://arxiv.org/abs/2205.11916 (Accessed 20 January 2023).
43. S. J. Han, K. Ransom, A. Perfors, C. Kemp, Human-like property induction is a challenge for large language models. PsyArXiv (2022).
44. G. Betz, K. Richardson, C. Voigt, Thinking aloud: Dynamic context generation improves zero-shot reasoning performance of GPT-2. arXiv [Preprint] (2021). http://arxiv.org/abs/2103.13033 (Accessed 20 January 2023).
45. E. Kosoy *et al*., Towards understanding how machines can learn causal overhypotheses (2022).
46. A. Ettinger, What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Trans. Assoc. Comput. Ling.* **8**, 34–48 (2020).
47. C. R. Jones *et al*., "Distributional semantics still can't account for affordances" in *Proceedings of the Annual Meeting of the Cognitive Science Society* (2022), vol. 44.
48. S. Trott, C. Jones, T. Chang, J. Michaelov, B. Bergen, Do large language models know what humans know? arXiv [Preprint] (2022). http://arxiv.org/abs/2209.01515 (Accessed 20 January 2023).
49. A. Webson, E. Pavlick, Do prompt-based models really understand the meaning of their prompts? arXiv [Preprint] (2021). http://arxiv.org/abs/2109.01247 (Accessed 20 January 2023).
50. L. Reynolds, K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm" in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–7.
51. S. Ritter, D. G. Barrett, A. Santoro, M. M. Botvinick, "Cognitive psychology for deep neural networks: A shape bias case study" in *International Conference on Machine Learning (PMLR)* (2017), pp. 2940–2949.
52. A. S. Rich, T. M. Gureckis, Lessons for artificial intelligence from the study of natural stupidity. *Nat. Mach. Intell.* **1**, 174–180 (2019).
53. I. Rahwan *et al*., Machine behaviour. *Nature* **568**, 477–486 (2019).
54. E. Schulz, P. Dayan, Computational psychiatry for computers. *Iscience* **23**, 101772 (2020).
55. P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, K. Kersting, Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **4**, 258–268 (2022).