Have we built machines that think like people?

Luca M. Schulze Buschoff[1,+,*], Elif Akata[1,2,+], Matthias Bethge[2], and Eric Schulz[1]

[1]Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany
[2]University of Tübingen, 72076 Tübingen, Germany
[*]lucaschulzebuschoff@gmail.com
[+]these authors contributed equally to this work

ABSTRACT

A chief goal of artificial intelligence is to build machines that think like people. Yet it has been argued that deep neural network architectures fail to accomplish this. Researchers have asserted these models' limitations in the domains of causal reasoning, intuitive physics, and intuitive psychology. Yet recent advancements, namely the rise of large language models, particularly those designed for visual processing, have rekindled interest in the potential to emulate human-like cognitive abilities. This paper evaluates the current state of vision-based large language models in the domains of intuitive physics, causal reasoning, and intuitive psychology. Through a series of controlled experiments, we investigate the extent to which these modern models grasp complex physical interactions, causal relationships, and intuitive understanding of others' preferences. Our findings reveal that, while these models demonstrate a notable proficiency in processing and interpreting visual data, they still fall short of human capabilities in these areas. The models exhibit a rudimentary understanding of physical laws and causal relationships, but their performance is hindered by a lack of deeper insights—a key aspect of human cognition. Furthermore, in tasks requiring an intuitive theory of mind, the models fail altogether. Our results emphasize the need for integrating more robust mechanisms for understanding causality, physical dynamics, and social cognition into modern-day, vision-based language models, and point out the importance of cognitively-inspired benchmarks.

Introduction

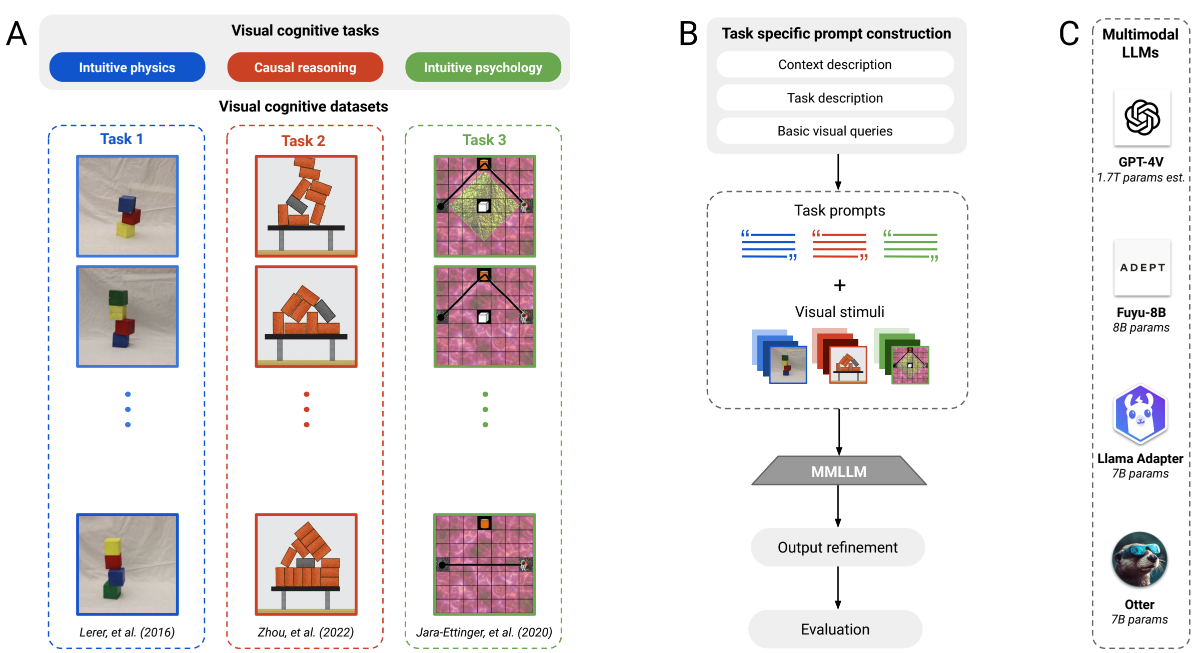


Figure 1. Overview of domains tasks, approach, and models. **A:** Example images for the different experiments. Each experiment was taken from one of three cognitive domains: intuitive physics, causal reasoning, and intuitive psychology. **B:** General approach. For every query, an image was submitted to the model, and different questions were asked about the image, i.e. we performed visual question answering.**C:** Used multi-modal large language models and their size.

People are quick to anthropomorphize, attributing human characteristics to non-human agents[1]. This tendency, far from being a modern quirk, is deeply rooted in our collective psyche, as evidenced through our literary traditions. Consider E.T.A. Hoffmann's "Der Sandmann"[2], where Nathaniel's infatuation with Olimpia, a mere mechanical automaton with a limited repertoire of expressions – her most notable being a plaintive 'Ach, Ach!' – serves as a poignant example. With further technological advances, this phenomenon, once confined to the realm of literature, has now manifested in reality. Already simple chatbots like ELIZA[3], which used simple pattern matching and substitution methodologies, gave users the illusion of human-like understanding and engagement[4]. The tendency to anthropomorphize has only intensified with the advent of Large Language Models (LLMs)[5]. These advanced AI systems represent a big leap from ELIZA both in complexity and capability. Unlike ELIZA's simplistic pattern matching, LLMs apply deep learning techniques to generate text[6], learning from vast datasets to produce responses that can be startlingly human-like[7]. Astonishingly, these models cannot only generate text. When scaled up to bigger training data and architectures, other, so-called "emergent abilities" appear[8,9]. The current models can, for example, pass the bar exam[10], write poems[11], compose music[12], and assist in programming and data analysis tasks[13]. As a result, the line between human and machine capabilities is increasingly blurred[14,15]. People not only interact with these systems as if they were humans[16], but they also start to rely on them for complex decision-making[17], artistic creation[18], and personal interactions[19]. It is, therefore, natural to ask: Have we built machines that think like people? Or are we, just like Nathaniel, projecting human qualities into rudimentary entities that are fundamentally different from us?

Judging whether or not artificial agents can mimic human thought is at the core of cognitive science[20,21]. Therein, researchers have long debated the capabilities of artificially intelligent agents[22–24]. In a seminal paper, Lake et al.[25] proposed core domains to consider when making such judgments. Published during the height of the deep learning revolution[26], the authors focused on domains that were easy for people but difficult for deep learning models: intuitive physics, causal reasoning, and intuitive psychology.

Research on intuitive physics has studied how people perceive and interpret physical phenomena[27–29]. Past work on this topic has emphasized that humans possess an innate ability to predict and understand the physical properties of objects and their interactions[30], even from a young age[31], a notion sometimes summarized as a "physics engine" in people's heads[32]. This understanding includes concepts such as gravity[33], inertia[34], and momentum[35]. Some of the most canonical tasks in this domain involve testing people's judgments about the stability of block towers[36,37]. These tasks have made their way into machine learning benchmarks[38,39], where they are used to test the intuitive physical understanding of neural networks (see[40] for an overview of previous work on building models with human-like physical knowledge).

Research on causal reasoning has studied how individuals infer and think about cause-effect relationships[41–43]. Past work on this topic has proposed that humans possess an intuitive capacity to infer, understand, and predict causal relationships in their environment[44–47], oftentimes described using Bayesian models of causal learning[48,49]. This cognitive ability encompasses recognizing patterns[50,51], inferring causes from interventions[52,53], and predicting future events based on hypothetical events[54]. Canonical tasks in this domain often involve assessing individuals' ability to infer causal relationships, for example, when judging the responsibility of one object causing other objects' movement[55,56]. Causal reasoning remains a challenge, even for current machine learning approaches[57,58].

Research on intuitive psychology has explored how individuals infer, understand, and interpret social phenomena and mental states of other agents[59,60]. Past work on this topic has emphasized the concept that humans possess an inherent ability to infer and reason about the mental states[61,62], intentions, and emotions of others, often referred to as a "theory of mind"[63,64]. This ability has been modeled as a Bayesian inference problem[65–67]. Canonical tasks in this domain often involve assessing individuals' capacity to predict actions based on understanding others' perspectives or intentions, such as determining agents' utility functions based on their actions in a given environment[68,69]. It is the subject of ongoing debates if modern algorithms show any form of intuitive psychology[70–72].

Lake et al. argued that some of these abilities act as "start-up software," because they constitute cognitive capabilities present early in development. Moreover, they proposed that these so-called "intuitive theories"[73,74] need to be expressed explicitly using the calculus of Bayesian inference[75], as opposed to being learned from scratch, for example, via gradient descent. However, with the increase in abilities of current neural networks, in particular LLMs, we pondered: Can LLMs, in particular vision LLMs, sufficiently solve problems from these core domains?

To address this question, we took canonical tasks from the domains of intuitive physics, causal reasoning, and intuitive psychology that could be studied by providing images and language-based questions. We submitted them to some of the currently most advanced LLMs. Our results showed that these models can, at least partially, solve these tasks. In particular, the largest currently available model, OpenAI's Generative Pre-trained Transformer (GPT-4) managed to perform robustly above chance in two of the three domains. Yet crucial differences emerged. First, none of the models matched human-level performance in any of the domains. Secondly, none of the models fully captured human behavior, leaving room for domain-specific models of cognition such as the Bayesian models originally proposed for the tasks.

Results

We tested four different models on three core components for human-like intelligence as outlined by Lake et al. ([25]; see Fig. 1A). The models we used are vision large language models, which are multimodal models that integrate image processing capabilities into large language models ([76,77]; see Fig. 1B). These models allow users to perform visual question answering[78,79]: users can upload an image and ask questions about it, which the model interprets and responds to accordingly.

The first model was GPT-4 with Vision (GPT-4V), developed by OpenAI[80]. This multimodal model extends the abilities of GPT-4 to analyze and interpret images, although the details of how the model accomplished this have not been made public. The second model is ADEPT's Fuyu-8B[81], which is a decoder-only multimodal model based on the classic transformer architecture. Fuyu-8B stands out due to its simpler architecture and training procedure compared to other multimodal models. The third model is LLaMA-Adapter V2[82], or short Adapter, which offers an alternative approach to enhance the capabilities of vision-language models by increasing learnable parameters, introducing an early fusion strategy for better visual integration, and employing a joint training approach for image-text and instruction-following data. The final model is Otter[83], which is based on OpenFlamingo[84], an open-sourced version of DeepMind's Flamingo[85], and specifically designed for multimodal in-context instruction following.

For testing the three core components, we used three tasks taken from the cognitive science literature that could be studied in vision LLMs via visual question answering. For intuitive physics, we asked models to judge the stability of different block towers, using stimuli originally proposed by Lerer and colleagues[86] and based on previous research by Battaglia et al.[36]. For causal reasoning, we again used block towers and asked how many blocks would fall if certain blocks were removed or to judge the responsibility of certain blocks for the stability of the tower, adopting a design that Zhou and colleagues had previously used in human subjects[87]. Finally, for intuitive psychology, we used a task where the models saw a picture of an agent's path on a grid and then had to make inferences about the costs and rewards associated with the environment, taken from Jara-Ettinger and colleagues[88].

For every task, we queried the visual reasoning abilities of the LLMs with tasks of increasing complexity (see Fig. 1C). First, we asked about simple features of the shown images such as the background color or the number of objects shown. Afterward, we submitted questions taken from the cognitive science experiments. We report results based on comparisons to the ground truth as well as the different models' matches to human data.

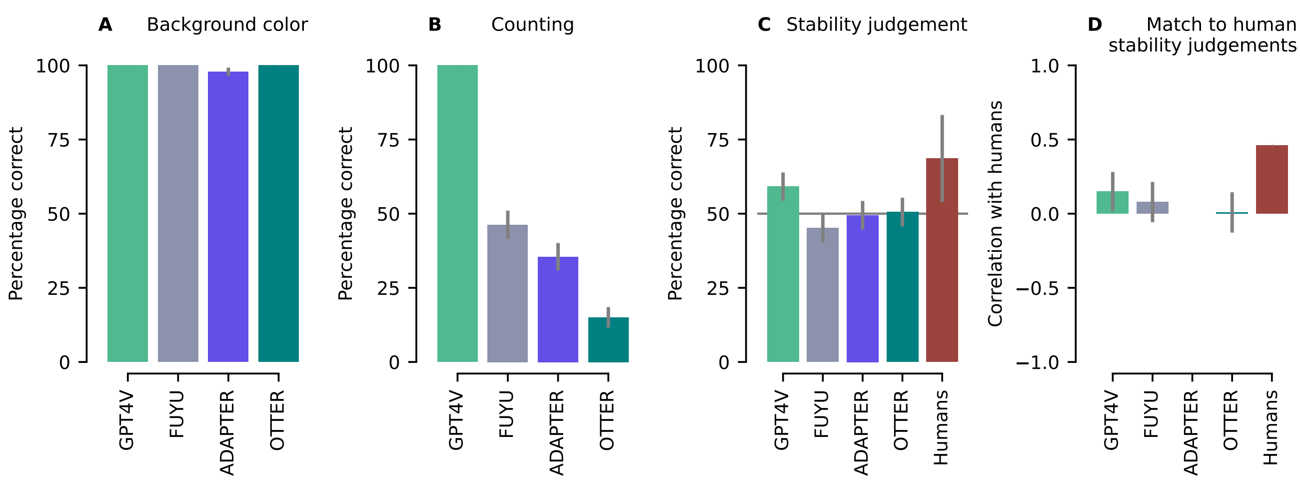


Figure 2. Results for four different large language models for tasks of increasing complexity given images of real block towers from[86]. We first ask for the background color in the image (A), then the number of blocks in the image (B), and finally a binary stability rating for the block towers (C). The last plot shows the match to human stability ratings as determined by the Pearson correlation (D). Errors bars in plots A - C are given by the standard deviation of a binomial distribution. The error bars in plot D are given by the 95% confidence interval for the correlation coefficient.

Intuitive physics

To test the intuitive physics capabilities of the different LLMs, we used photographs depicting wooden block towers from Lerer et al.[86] (see Fig. 5 in the Appendix for an example). These images mirror stimuli that developmental psychologists use to study the development of intuitive physics in infants. We used these images to test the models in increasingly complex tasks, starting with determining the background color of a given image (1), counting the number of colored blocks in the image (2), and giving a binary stability judgment of the depicted block tower (3).

All four models were able to correctly perform the first and easiest task: they all achieved almost perfect accuracy in determining the background color of the images (see Fig. 2A). In the second task, the performance of most models deteriorated, with only GPT-4V correctly determining the number of blocks in all images (see Fig. 2B). It is important to note that the first two tasks are fairly trivial for humans and we would expect human performance to be at 100% (the background color is always white and images featured either 2, 3, or 4 blocks).

While the first two tasks provided insight into the models' performance on high-level descriptive tasks, the third task directly tested their physical reasoning abilities. Performance for most models was at chance, with only GPT-4V performing slightly above chance in determining the stability of the block towers (see Fig. 2C, Fisher's exact test yielded an odds ratio of 2.38 with a one-sided p-value of 0.049). None of the other models performed significantly above chance (all p > 0.05). Human subjects were also not perfect but showed an average accuracy of 63.28%, which was also larger than chance with $p < .001$. Interestingly, there was no statistically significant difference between people's and GPT-4V's accuracy ($z = 0.684$, $p = .49$), likely because the task was also hard for humans and because Lerer and colleagues only collected 10 human subjects.

Finally, we determined the similarity between models' and humans' stability judgments as determined by the Pearson correlation (see Fig. 2D). We found that GPT-4V was the only model that showed a significant correlation with human judgments, $r = 0.155$, $t(205) = 2.241$, $p = .01$, while none of the other models showed any meaningful match to human judgments (all $p > .05$). However, the average correlation between humans ($r = 0.46$, $t(910) = 14.127$, $p < .001$) was larger than the correlation between GPT-4V and humans ($z = 3.825$, $p < .001$).

Causal reasoning

To test the causal reasoning capabilities, we used synthetic images again depicting block towers from Zhou et al.[87,89] (see Fig. 6 in the Appendix for an example). Here, the images showed static scenes of block towers that were stable but might collapse if one of the blocks was removed.

Again, we started by asking the models to count the blocks in the image (1), we continued by querying the models for the number of blocks that would fall if a specific block was removed from the scene (2), and finally, we asked the models to rate the responsibility of a specific block for the stability of the other blocks (3). For the second task, we established a baseline

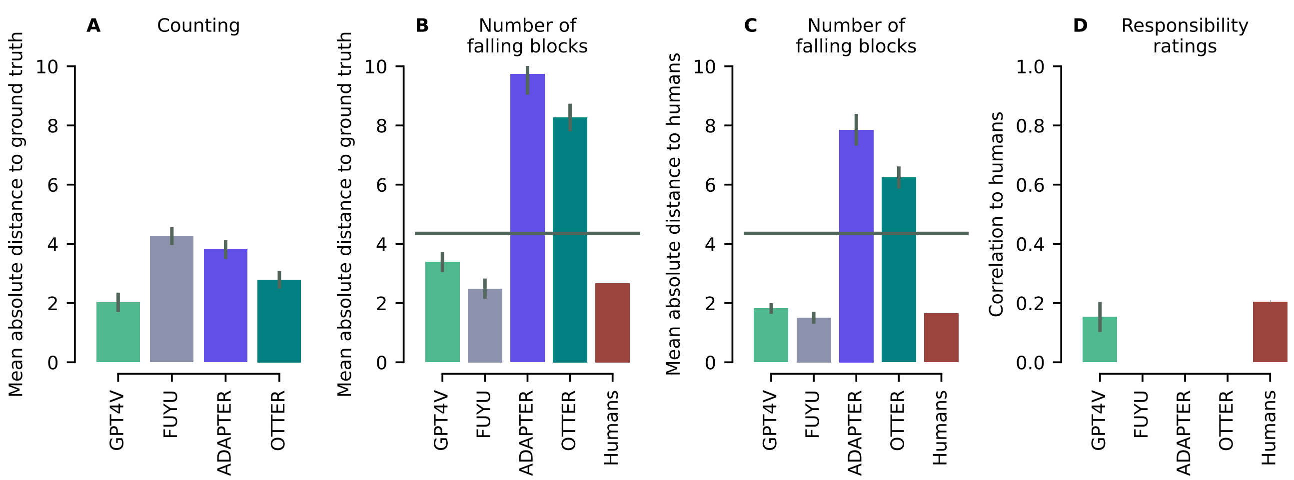


Figure 3. Results for analyses using images from causal reasoning experiments taken from[89]. We first ask for the number of blocks in the image, then the number of blocks that would fall if a specific block is removed, and finally a rating between 0 and 100 for how responsible a specific block is for the stability of the tower. For the responsibility ratings, all LLMs except for GPT-4V give constant ratings: Fuyu always responds with 100, while Otter and LLaMA-Adapter V2 always respond with 50. Error bars in plots A - C are given by the standard error of the mean, while the error bar plot D is given by the 95% confidence interval of the correlation coefficient.

performance represented by a horizontal line in Fig. 3A-C, which corresponds to a random agent. This random agent used a simple strategy: it gave the mean between 0 and the number of blocks in each image as its' prediction, essentially behaving like a uniform distribution over the possible number of blocks that could fall.

The images in this task displayed a larger number of blocks (ranging from 6 to 19), which made the basic counting task significantly more challenging than in the previous section. Models' responses approximated the ground truth, albeit rarely matching it exactly. Therefore, we report the mean absolute distance to the ground truth instead of the percentage of correct answers (see Fig. 3A). The models' performance highlighted the challenging nature of this task, with the best performing model (GPT4-V) still being on average 2 blocks off.

In Figures 3B and 3C, model performances for the second task are shown. Notably, both GPT-4V and Fuyu-8B surpassed the random baseline, their performance levels being close to the human results reported in[87], which is depicted by the rightmost bar in the plot. However, GPT4-V still diverges significantly from the average over human subjects ($t(42) = 2.59$, $p < .05$). In the second task, all four models exhibited mean correlations with human values ranging from 0.26 to 0.39, (all $p < .001$), with Adapter demonstrating the highest correlation. However, it is important to note that despite its relatively high correlation, Adapter also exhibited the highest mean absolute distance to ground truth values.

Next, we show the Pearson correlation of all models to human subjects for the third task in Figure 3D. Notably, all models except for GPT-4V gave constant ratings for this task (Fuyu always responds with 100, while Otter and LLaMA-Adapter V2 always respond with 50), making their correlation with human judgments undefined. GPT-4V, on the other hand, demonstrated a mean correlation of 0.15 with human values ($p < .001$). For the human-to-human correlation, we randomly paired human observers from the original study and calculated the correlation over their concatenated responses. We repeated this ten times and calculated the average of the Fisher z-transformed correlations ($r = 0.195$, $p < .001$). While the average correlation between humans was higher, it was not significantly higher than the correlation between GPT-4V and humans ($z = 0.919, p = 0.179$).

Intuitive psychology

To test the intuitive psychology of the different LLMs we used synthetic images depicting an astronaut on a colored background from Jara-Ettinger et al.[88] (see Fig. 5 in the Appendix for an example). The experiment consisted of three parts which differed slightly in their layout and components. Depending on the experiment, the images featured either one or two different terrains (indicated by different background colors and textures) as well as either one or two different care packages. The astronaut was shown with a path that led from a starting point to a base; the astronaut could collect care packages along the way. Depending on which terrain the astronaut crossed or which care package they chose to pick up or not, it was possible to infer the costs associated with the different terrains and the rewards associated with the different care packages.

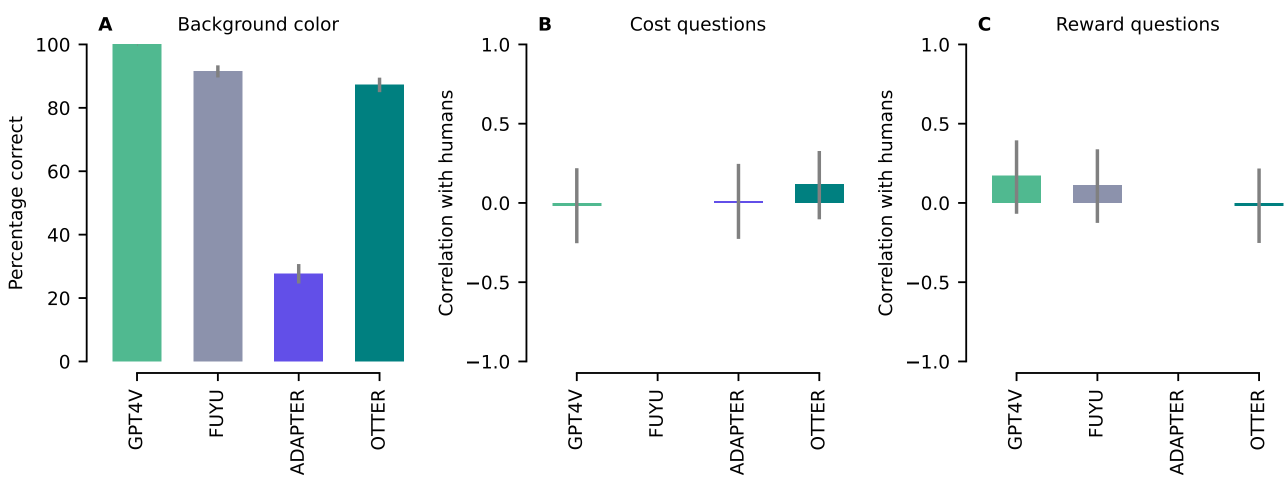


Figure 4. Results on tasks for intuitive psychology taken from[88]. Again, we first ask for the background color (A). Models are then asked to make inferences about the costs and rewards in an environment depending on the path an agent has taken (B and C). Correlation values for Fuyu and LLaMA-Adapter V2 are missing as they always responded with constant ratings for either cost or reward questions. Error bars in plot A are given by the standard deviation of a binomial distribution, while the error bars in plots B and C are given by the 95% confidence interval for the correlation coefficient.

Again, we first tasked models with determining the background color of the images (1), afterwards we asked them to infer the costs associated with the different terrains (2) and the rewards associated with the different care packages (3). The results for the first task are shown in Figure 4A. The performance of the models in determining the background color was worse compared to the intuitive physics data set, which might be due to the fact that the background color here was not uniform (see Fig. 8). For tasks 2 and 3, we pooled the answers for the three parts of the experiment, as there was only a small number of images in each individual experiment. As shown in Figures 4B and 4C, all models only showed no or very weak correlations with the average over human subjects in their judgments about the costs and rewards associated with the environment. Correlations with the average over human subjects ranged from -0.02 to 0.12 for cost questions and from -0.02 to 0.17 for reward questions (all $p > .05$).

Discussion

We started by asking whether, with the rise of modern large language models, researchers have created machines that – at least to some degree – think like people. To address this question, we took four recent multi-modal large language models and probed their abilities in three core cognitive domains: intuitive physics, causal reasoning, and intuitive psychology.

In intuitive physics, the models managed to solve some of the given tasks and showed a medium match with human data. Similarly, in a causal reasoning task, some models, in particular, GPT-4V, performed well and showed a medium match with human data. Finally, in an intuitive psychology task, none of the models performed well and none of them showed a reasonable match with human data. Thus, an appropriate answer to the question motivating our work would be "No.", or – perhaps more optimistically – "Not quite."

Limitations and Future Work

Although we have tried our best to give all models a fair chance and set up the experiments in a clean and replicable fashion, some shortcomings remain that should be addressed in future work.

First of all, we have only tested a handful of multi-modal models on just three cognitive domains. While we believe that the used models and tasks provide good insights into the state-of-the-science of LLMs' cognitive abilities, future studies should look at more domains and different models to further tease apart when and why LLMs can mimic human reasoning. For example, it would be interesting to see if scale is the only important feature influencing model performance[90,91]. Currently, our evidence suggests that even smaller models, for example Fuyu with its 8 billion parameters, can sometimes perform as well as GPT-4V in some tasks.

Another shortcoming of the current work is the simplicity of the used stimuli. While the block towers used in our first study were deliberately designed to be more realistic[86] than commonly used psychological stimuli[36], this was not true for the experiments in the other two domains. For the intuitive psychology experiments, in particular, we would expect the models to perform better if the stimuli contained more realistic images of people, which has been shown to work better in previous studies[92]. Interestingly, using more realistic stimuli can also change people's causal judgments[93]; how realistic stimuli used in cognitive experiments should be, remains an open question[94].

On a related point, we only used static images in our current experiments, which severely limits the breadth and level of detail of the questions we could ask. For example, some of the most canonical tasks investigating people's causal reasoning abilities involve videos of colliding billiard balls[55]. As future large language models will likely be able to answer questions about videos[95], these tasks represent the next frontier of cognitively-inspired benchmarks.

For the comparisons to human data, we currently used the participant data collected in the original studies and assessed the correspondence between models and this data via correlation coefficients. Future work could expand on this approach by collecting new data from human subjects choosing which of the model's judgments they prefer. This could lead to a more detailed comparison, similar to what has been proposed to discriminate among deep learning models for human vision[96] and language[97].

A crucial weakness of most studies using large language models is that they can be sensitive to specific prompts[98–100]. While we have attempted to use prompts that elicited good behavior, thereby giving LLMs a chance to perform well, future work could try to further optimize these prompts using available methods[101–103], while also assessing how the models respond to paraphrased versions of the same tasks. While it could be possible to further engineer the used prompts, as well as try out several other ways of phrasing the same prompt, we believe that our current approach was sufficient to showcase these models' abilities.

Finally, we applied all models out of the box and without additional fine-tuning. Future studies could attempt to fine-tune multi-modal LLMs to better align with cognitive data[104] and assess if this improves their reasoning abilities more generally.

Related work

We are not the first to assess LLMs' reasoning abilities[105–107]. Previous studies have focused, among others, on testing LLMs' cognitive abilities in model-based planning[108], analogical reasoning tests[109], exploration tasks[110], systematic reasoning tests[111,112], psycholinguistic completion studies[113], and affordance understanding problems[114]. In this sense, our contribution can be seen as a part of a larger movement in which researchers use methods from the behavioral sciences to understand black box machine learning models[115–117]. However, most of the previous studies did not investigate multi-modal LLMs but rather remained in the pure language domain. Although there are recent attempts to investigate vision LLMs cognitive features, including their reaction to visual illusions[118] as well as how they solve simple intelligence tasks[119], we are the first to investigate the proposed core components of cognition in these models.

Previous work has also looked at how LLMs solve cognitive tasks taken from the same domains that we have looked at. In intuitive physics, Zečević et al.[120] found that LLMs performed poorly in a task using language descriptions of physical scenarios. Zhang and colleagues[121] extracted programs from text produced by large language models to improve their physical reasoning abilities. Finally, Jassim and colleagues[122] proposed a novel benchmark for evaluating multimodal LLMs' understanding of situated physics. In causal reasoning, Binz and Schulz[123] showed that GPT-3 failed at simple causal reasoning experiments, while Kosoy et al.[124] showed that LLMs cannot learn human-like causal over-hypotheses. In research on intuitive psychology, Kosinsky argued that theory of mind might have emerged in LLMs[71] which has been criticized other researchers[72]. Akata et al. showed that GTP-4 plays repeated games very selfishly and could not pick up on simple conventions such as alternating between options[19]. Finally, Gandhi and colleagues[125] proposed a framework for procedurally generating Theory of Mind evaluations and found that GPT4's abilities mirror human inference patterns, though less reliable, while all other LLMs struggled.

Many of the past studies on LLMs have fallen risk of appearing in new models' training set. Recent work has recognized this issue and, in turn, evaluated language models on many problem variations to minimize training set effects[126]. Our work differs from these approaches as current models could not have just memorized solutions to the given problems because these problems require deep reasoning and are rarely ever published with the ground truth attached.

Conclusion

A major plot twist in E.T.A Hoffmann's "Der Sandmann" is that Nathaniel might have been a machine himself, which explains why he fell for Olimpia. This metaphor also relates to our anthropomorphization of LLMs: since LLMs are trained on human-generated data, their behaviors will always reflect our behaviors and biases. However, this reflection is sharpening, and modern neural network architectures have become more human-like. One of the other domains emphasized by Lake et al. was the ability to reason compositionally. Recent attempts have shown that neural networks can perform compositional reasoning if trained appropriately[127]. Similarly, our current work has shown that multimodal LLMs have come a long way, showing some correspondence to human behavior and often performing above chance. Moreover, machine learning researchers have put forward various ideas about how to close the remaining gap between humans and machines[128], including self-supervised learning[129], translating from natural into probabilistic languages[130], or grounding LLMs in realistic environments[131]. This continuous evolution in models' capabilities necessitates a reevaluation of the metaphors and tools we use to understand them. We believe that cognitive science can offer tools, theories, and benchmarks to evaluate how close we have come to building machines that think like people.

Methods

Code

The open-source models were installed per the instructions on their related Github or Huggingface repositories and evaluated on a Slurm-based cluster with a single A100. For the results reported as GPT-4V, we used the public ChatGPT interface and the OpenAI API, specifically the November 2023 release of `gpt4-vision-preview` model which is available via the completions endpoint. Code for replicating our results is available on GitHub (github.com/lsbuschoff/multimodal). All models were evaluated in Python using PyTorch[132]. Additional analyses were carried out using NumPy[133], Pandas[134], and SciPy[135]. Matplotlib[136] and Seaborn[137] were used for plotting.

Models

GPT4-V

We initially queried GPT-4V through the ChatGPT interface, since the OpenAI API was not publicly available at the outset of this project. The Intuitive Psychology task responses were collected using the `gpt4-vision-preview` model variant after its November 2023 release in the API. We set the maximum number of generated tokens for a given prompt to 1 to get single numerical responses. All other parameters were set to their default values. Note that this model does not currently feature an option for manually setting the temperature, and the provided documentation does not specify what the default temperature is.

FUYU

Fuyu is an 8B parameter multi-modal text and image decoder-only transformer. We used the Huggingface implementation with standard settings and without further finetuning (available here). The maximum number of generated tokens was set to 8 and responses were parsed by hand.

ADAPTER

We here use the multi-modal version of LLaMA-Adapter V2, which adds adapters into LLaMA's transformer in order to turn it into an instruction-following model. We used the GitHub implementation with standard settings and again without further finetuning (available here). The maximum number of generated tokens was left at 512 and responses were parsed by hand.

OTTER

Otter is a multi-modal LLM that supports in-context instruction tuning and it is based on the OpenFlamingo model. We used the Huggingface implementation (available here), again with standard settings and without finetuning. The maximum number of generated tokens was left at 512 and responses were parsed by hand.

Datasets

Intuitive physics

We tested the intuitive physical understanding of the models using images from Lerer et al.[86]. The photos depict a block tower consisting of colored wooden blocks in front of a white fabric (see Fig. 5 for an example). The images are of size 224 x 244. In the data set, there are a total of 516 images of block towers. We tested the models on 100 randomly drawn images. We first tested the models on their high-level visual understanding of the scenes: we tasked them with determining the background color and the number of blocks in the image. In order to test their physical understanding, we tested them on the same task as the original study: we asked them to give a binary rating on the stability of the depicted block towers. For the first two tasks, we calculated the percentage of correct answers for each of the models. For the third task, we used the Pearson correlation to determine the degree of similarity between the models and human subjects.

Causal reasoning

For the causality part we used images from Zhou et al.[87]. The images show artificial block stacks of red and gray blocks on a black table (see Fig. 6 for an example). The data set consists of 42 images on which we tested all models. We again first tested the models on their high-level visual understanding of the scene and therefore tasked them with determining the number of blocks in the scene. The ground truth number of blocks in the scenes ranged from 6 to 19. Since this task is rather challenging due to the increased number of blocks, we do not report the percentage correct as for the intuitive physics data set but the mean over the absolute distance between model predictions and the ground truth for each image (see Fig. 3A).

To test the causal reasoning of the models we adopted the tasks performed in the original study[89]. We asked models to infer how many red blocks would fall if the gray block was removed. We again report the absolute distance between model predictions and the ground truth for each image (see Fig. 3B). We calculate a random baseline which uses the mean between 0 and the number of blocks for each specific image as the prediction. We also ask the models for a rating between 0 and 100 for how responsible the gray block is for the stability of the tower. For both, the number of blocks that would fall if the gray block was removed, and its' responsibility for the stability of the tower, we calculate the mean Pearson correlation to human subjects from the original study (see Fig. 3C).

Intuitive psychology

To test the intuitive psychology of the different LLMs, we used stimuli from Jara-Ettinger et al.[88]. This part consisted of three different experiments each consisting of 16, 17, and 14 images showing a 2D depiction of an astronaut and care packages in different terrains (see Fig. 5 for an example). In order to check their high-level understanding of the images, we again asked the models to determine the background color of the images. Since this background color is not uniform, we counted both "Pink" and "Purple" as correct answers. We report the percentage of correct answers for the background color in Figure 2A.

In accordance with the original study, analyses for the intuitive psychological capabilities of the models are split into cost questions (passing through a terrain is associated with a cost for the agent) and reward questions (collecting a care package yields some sort of reward for the agent). We pooled cost and reward questions over all three experiments and reported the mean Pearson correlation with both humans and a heuristic outlined in[88] (see Figs. 2B and 2C). This heuristic calculates the costs and rewards associated with the environment from the amount of time an agent spends in each terrain and which care package it collects.

References

1. Mitchell, M. *Artificial intelligence: A guide for thinking humans* (Penguin UK, 2019).
2. Hoffmann, E. T. A. *ETA Hoffmann:" Der Sandmann"* (Walter de Gruyter, 1988).
3. Weizenbaum, J. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**, 36–45 (1966).
4. Hofstadter, D. The ineradicable eliza effect and its dangers. *Fluid concepts creative analogies: Comput. models fundamental mechanisms thought* (1995).
5. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
6. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).
7. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).
8. Bubeck, S. *et al.* Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
9. Wei, J. *et al.* Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).
10. Katz, D. M., Bommarito, M. J., Gao, S. & Arredondo, P. Gpt-4 passes the bar exam. *Available at SSRN 4389233* (2023).
11. Sawicki, P. *et al.* On the power of special-purpose gpt models to create and evaluate new poetry in old styles. *preprint* (2023).
12. Borsos, Z. *et al.* Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, Lang. Process.* (2023).
13. Poldrack, R. A., Lu, T. & Beguš, G. Ai-assisted coding: Experiments with gpt-4. *arXiv preprint arXiv:2304.13187* (2023).
14. Kasneci, E. *et al.* Chatgpt for good? on opportunities and challenges of large language models for education. *Learn. individual differences* **103**, 102274 (2023).
15. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
16. Elkins, K. & Chun, J. Can gpt-3 pass a writer's turing test? *J. Cult. Anal.* **5** (2020).
17. Dell'Acqua, F. *et al.* Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harv. Bus. Sch. Technol. & Oper. Mgt. Unit Work. Pap.* (2023).
18. Bašić, Ž., Banovac, A., Kružić, I. & Jerković, I. Better by you, better than me? chatgpt-3 as writing assistance in students' essays. *preprint* (2023).
19. Akata, E. *et al.* Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867* (2023).
20. Simon, H. A. Cognitive science: The newest science of the artificial. *Cogn. science* **4**, 33–46 (1980).
21. Rumelhart, D. E., McClelland, J. L., Group, P. R. *et al.* Parallel distributed processing. *Foundations* **1** (1988).
22. Wichmann, F. A. & Geirhos, R. Are deep neural networks adequate behavioral models of human visual perception? *Annu. Rev. Vis. Sci.* **9** (2023).
23. Bowers, J. S. *et al.* On the importance of severely testing deep learning models of cognition. *Cogn. Syst. Res.* **82**, 101158 (2023).
24. Marcus, G. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631* (2018).
25. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40** (2017).
26. Sejnowski, T. J. *The deep learning revolution* (MIT press, 2018).
27. Smith, K. A., Battaglia, P. W. & Vul, E. Different physical intuitions exist between tasks, not domains. *Comput. Brain & Behav.* **1**, 101–118 (2018).
28. Bates, C. J., Yildirim, I., Tenenbaum, J. B. & Battaglia, P. Modeling human intuitions about liquid flow with particle-based simulation. *PLoS computational biology* **15**, e1007210 (2019).

29. Battaglia, P. *et al.* Computational models of intuitive physics. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34 (2012).

30. Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B. & Gureckis, T. M. Intuitive experimentation in the physical world. *Cogn. psychology* **105**, 9–38 (2018).

31. Ullman, T. D. & Tenenbaum, J. B. Bayesian models of conceptual development: Learning as building models of the world. *Annu. Rev. Dev. Psychol.* **2**, 533–558 (2020).

32. Ullman, T. D., Spelke, E., Battaglia, P. & Tenenbaum, J. B. Mind games: Game engines as an architecture for intuitive physics. *Trends cognitive sciences* **21**, 649–665 (2017).

33. Hamrick, J., Battaglia, P. & Tenenbaum, J. B. Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd annual conference of the cognitive science society*, vol. 2 (Cognitive Science Society Austin, TX, 2011).

34. Mildenhall, P. & Williams, J. Instability in students' use of intuitive and newtonian models to predict motion: the critical effect of the parameters involved. *Int. J. Sci. Educ.* **23**, 643–660 (2001).

35. Todd, J. T. & Warren Jr, W. H. Visual perception of relative mass in dynamic events. *Perception* **11**, 325–335 (1982).

36. Battaglia, P. W., Hamrick, J. B. & Tenenbaum, J. B. Simulation as an engine of physical scene understanding. *Proc. Natl. Acad. Sci.* **110**, 18327–18332 (2013).

37. Hamrick, J. B., Battaglia, P. W., Griffiths, T. L. & Tenenbaum, J. B. Inferring mass in complex scenes by mental simulation. *Cognition* **157**, 61–76 (2016).

38. Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L. & Girshick, R. Phyre: A new benchmark for physical reasoning. *Adv. Neural Inf. Process. Syst.* **32** (2019).

39. Riochet, R. *et al.* Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616* (2018).

40. Schulze Buschoff, L. M., Schulz, E. & Binz, M. The acquisition of physical knowledge in generative neural networks. *Proc. 40th Int. Conf. on Mach. Learn.* (2023).

41. Waldmann, M. *The Oxford handbook of causal reasoning* (Oxford University Press, 2017).

42. Cheng, P. W. From covariation to causation: A causal power theory. *Psychol. review* **104**, 367 (1997).

43. Holyoak, K. J. & Cheng, P. W. Causal learning and inference as a rational process: The new synthesis. *Annu. review psychology* **62**, 135–163 (2011).

44. Pearl, J. *Causality* (Cambridge university press, 2009).

45. Griffiths, T. L. & Tenenbaum, J. B. Theory-based causal induction. *Psychol. review* **116**, 661 (2009).

46. Lagnado, D. A., Waldmann, M. R., Hagmayer, Y. & Sloman, S. A. Beyond covariation. *Causal learning: Psychol. philosophy, computation* 154–172 (2007).

47. Carey, S. On the origin of causal understanding. (1995).

48. Gopnik, A. *et al.* A theory of causal learning in children: causal maps and bayes nets. *Psychol. review* **111**, 3 (2004).

49. Lucas, C. G. & Griffiths, T. L. Learning the form of causal relationships using hierarchical bayesian models. *Cogn. Sci.* **34**, 113–147 (2010).

50. Bramley, N. R., Gerstenberg, T., Mayrhofer, R. & Lagnado, D. A. Time in causal structure learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 1880 (2018).

51. Griffiths, T. L. & Tenenbaum, J. B. Structure and strength in causal induction. *Cogn. psychology* **51**, 334–384 (2005).

52. Schulz, L., Kushnir, T. & Gopnik, A. Learning from doing: Intervention and causal inference. *Causal learning: Psychol. philosophy, computation* 67–85 (2007).

53. Bramley, N. R., Dayan, P., Griffiths, T. L. & Lagnado, D. A. Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychol. review* **124**, 301 (2017).

54. Gerstenberg, T. What would have happened? counterfactuals, hypotheticals and causal judgements. *Philos. Transactions Royal Soc. B* **377**, 20210339 (2022).

55. Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A. & Tenenbaum, J. B. Eye-tracking causality. *Psychol. science* **28**, 1731–1744 (2017).

56. Gerstenberg, T., Goodman, N. D., Lagnado, D. A. & Tenenbaum, J. B. A counterfactual simulation model of causal judgments for physical events. *Psychol. review* **128**, 936 (2021).

57. Jin, Z. *et al.* Cladder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems* (2023).

58. Dasgupta, I. *et al.* Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162* (2019).

59. Baker, C. L. & Tenenbaum, J. B. Modeling human plan recognition using bayesian theory of mind. *Plan, activity, intent recognition: Theory practice* **7**, 177–204 (2014).

60. Jern, A. & Kemp, C. A decision network account of reasoning about other people's choices. *Cognition* **142**, 12–38 (2015).

61. Vélez, N. & Gweon, H. Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Curr. opinion behavioral sciences* **38**, 110–115 (2021).

62. Spelke, E. S., Bernier, E. P. & Skerry, A. Core social cognition (2013).

63. Baker, C., Saxe, R. & Tenenbaum, J. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, vol. 33 (2011).

64. Frith, C. & Frith, U. Theory of mind. *Curr. biology* **15**, R644–R645 (2005).

65. Saxe, R. & Houlihan, S. D. Formalizing emotion concepts within a bayesian model of theory of mind. *Curr. opinion Psychol.* **17**, 15–21 (2017).

66. Goodman, N. D. *et al.* Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the twenty-eighth annual conference of the cognitive science society*, vol. 6 (Cognitive Science Society Vancouver, 2006).

67. Shum, M., Kleiman-Weiner, M., Littman, M. L. & Tenenbaum, J. B. Theory of minds: Understanding behavior in groups through inverse planning. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 6163–6170 (2019).

68. Baker, C. L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1**, 0064 (2017).

69. Zhi-Xuan, T. *et al.* Solving the baby intuitions benchmark with a hierarchically bayesian theory of mind. *arXiv preprint arXiv:2208.02914* (2022).

70. Rabinowitz, N. *et al.* Machine theory of mind. In *International conference on machine learning*, 4218–4227 (PMLR, 2018).

71. Kosinski, M. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083* (2023).

72. Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399* (2023).

73. Schulz, L. The origins of inquiry: Inductive inference and exploration in early childhood. *Trends cognitive sciences* **16**, 382–389 (2012).

74. Ullman, T. D. *On the nature and origin of intuitive theories: learning, physics and psychology*. Ph.D. thesis, Massachusetts Institute of Technology (2015).

75. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *science* **331**, 1279–1285 (2011).

76. Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis machine intelligence* **41**, 423–443 (2018).

77. Reed, S. *et al.* Generative adversarial text to image synthesis. In *International conference on machine learning*, 1060–1069 (PMLR, 2016).

78. Wu, Q. *et al.* Visual question answering: A survey of methods and datasets. *Comput. Vis. Image Underst.* **163**, 21–40 (2017).

79. Manmadhan, S. & Kovoor, B. C. Visual question answering: a state-of-the-art review. *Artif. Intell. Rev.* **53**, 5705–5745 (2020).

80. OpenAI. Gpt-4 technical report (2023). 2303.08774.

81. Bavishi, R. *et al.* Introducing our multimodal models (2023).

82. Gao, P. *et al.* Llama-adapter v2: Parameter-efficient visual instruction model (2023). 2304.15010.

83. Li, B. *et al.* Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726* (2023).

84. Awadalla, A. *et al.* Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390* (2023).

85. Alayrac, J.-B. *et al.* Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **35**, 23716–23736 (2022).

86. Lerer, A., Gross, S. & Fergus, R. Learning physical intuition of block towers by example. In *International Conference on Machine Learning*, 430–438 (PMLR, 2016).

87. Zhou, L., Smith, K. A., Tenenbaum, J. B. & Gerstenberg, T. Mental jenga: A counterfactual simulation model of causal judgments about physical support. *J. Exp. Psychol. Gen.* (2023).

88. Jara-Ettinger, J., Schulz, L. E. & Tenenbaum, J. B. The naïve utility calculus as a unified, quantitative framework for action understanding. *Cogn. Psychol.* **123**, 101334, DOI: https://doi.org/10.1016/j.cogpsych.2020.101334 (2020).

89. Gerstenberg, T., Zhou, L., Smith, K. A. & Tenenbaum, J. B. Faulty towers: A hypothetical simulation model of physical support. In *CogSci* (2017).

90. Sutton, R. The bitter lesson. *Incomplete Ideas (blog)* **13** (2019).

91. Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

92. Ju, C., Han, T., Zheng, K., Zhang, Y. & Xie, W. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, 105–124 (Springer, 2022).

93. Meding, K., Bruijns, S. A., Schölkopf, B., Berens, P. & Wichmann, F. A. Phenomenal causality and sensory realism. *i-Perception* **11**, 2041669520927038 (2020).

94. Allen, K. R. *et al.* Using games to understand the mind. *preprint* (2023).

95. Maaz, M., Rasheed, H., Khan, S. & Khan, F. S. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424* (2023).

96. Golan, T., Raju, P. C. & Kriegeskorte, N. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proc. Natl. Acad. Sci.* **117**, 29330–29337 (2020).

97. Golan, T., Siegelman, M., Kriegeskorte, N. & Baldassano, C. Testing the limits of natural language models for predicting human language judgements. *Nat. Mach. Intell.* 1–13 (2023).

98. Reynolds, L. & McDonell, K. 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*, 1–7 (2021).

99. Strobelt, H. *et al.* Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *arXiv preprint arXiv:2208.07852* (2022).

100. Webson, A. & Pavlick, E. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247* (2021).

101. Liu, P. *et al.* Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, 1–35 (2023).

102. Gu, J. *et al.* A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980* (2023).

103. Coda-Forno, J. *et al.* Meta-in-context learning in large language models. *arXiv preprint arXiv:2305.12907* (2023).

104. Binz, M. & Schulz, E. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917* (2023).

105. Huang, J. & Chang, K. C.-C. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403* (2022).

106. Sawada, T. *et al.* Arb: Advanced reasoning benchmark for large language models. *arXiv preprint arXiv:2307.13692* (2023).

107. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).

108. Binz, M. & Schulz, E. Using cognitive psychology to understand gpt-3. *Proc. Natl. Acad. Sci.* **120**, e2218523120 (2023).

109. Webb, T., Holyoak, K. J. & Lu, H. Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541 (2023).

110. Coda-Forno, J. *et al.* Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111* (2023).

111. Eisape, T. *et al.* A systematic comparison of syllogistic reasoning in humans and language models. *arXiv preprint arXiv:2311.00445* (2023).

112. Hagendorff, T., Fabi, S. & Kosinski, M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nat. Comput. Sci.* 1–6 (2023).

113. Ettinger, A. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions Assoc. for Comput. Linguist.* **8**, 34–48 (2020).

114. Jones, C. R. *et al.* Distrubutional semantics still can't account for affordances. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 44 (2022).

115. Rahwan, I. *et al.* Machine behaviour. *Nature* **568**, 477–486 (2019).

116. Schulz, E. & Dayan, P. Computational psychiatry for computers. *Iscience* **23** (2020).

117. Rich, A. S. & Gureckis, T. M. Lessons for artificial intelligence from the study of natural stupidity. *Nat. Mach. Intell.* **1**, 174–180 (2019).

118. Zhang, Y., Pan, J., Zhou, Y., Pan, R. & Chai, J. Grounding visual illusions in language: Do vision-language models perceive illusions like humans? *arXiv preprint arXiv:2311.00047* (2023).

119. Mitchell, M., Palmarini, A. B. & Moskvichev, A. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks (2023). 2311.09247.

120. Zečević, M., Willig, M., Dhami, D. S. & Kersting, K. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067* (2023).

121. Zhang, C., Wong, L., Grand, G. & Tenenbaum, J. Grounded physical language understanding with probabilistic programs and simulated worlds. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45 (2023).

122. Jassim, S. *et al.* Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models (2023). 2311.09048.

123. Binz, M. & Schulz, E. Using cognitive psychology to understand gpt-3. *arXiv preprint arXiv:2206.14576* (2022).

124. Kosoy, E. *et al.* Towards understanding how machines can learn causal overhypotheses. *arXiv preprint arXiv:2206.08353* (2022).

125. Gandhi, K., Fränken, J.-P., Gerstenberg, T. & Goodman, N. D. Understanding social reasoning in language models with language models. *arXiv preprint arXiv:2306.15448* (2023).

126. Srivastava, A. *et al.* Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

127. Lake, B. M. & Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature* 1–7 (2023).

128. Geirhos, R. *et al.* Partial success in closing the gap between human and machine vision. *Adv. Neural Inf. Process. Syst.* **34**, 23885–23899 (2021).

129. Balestriero, R. *et al.* A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210* (2023).

130. Wong, L. *et al.* From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672* (2023).

131. Carta, T. *et al.* Grounding large language models in interactive environments with online reinforcement learning. *arXiv preprint arXiv:2302.02662* (2023).

132. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019).

133. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362, DOI: 10.1038/s41586-020-2649-2 (2020).

134. pandas development team, T. pandas-dev/pandas: Pandas, DOI: 10.5281/zenodo.3509134 (2020).

135. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272, DOI: 10.1038/s41592-019-0686-2 (2020).

136. Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. Sci. & Eng.* **9**, 90–95, DOI: 10.1109/MCSE.2007.55 (2007).

137. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

Acknowledgements

This work was supported by the Max Planck Society, the Volkswagen Foundation, the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2064/1 – 390727645.

Author contributions statement

All authors conceived the experiments. L.B.S. and E.A. conducted the experiments. All authors analysed the results. All authors wrote the manuscript.

A Example trials

A.1 Intuitive physics

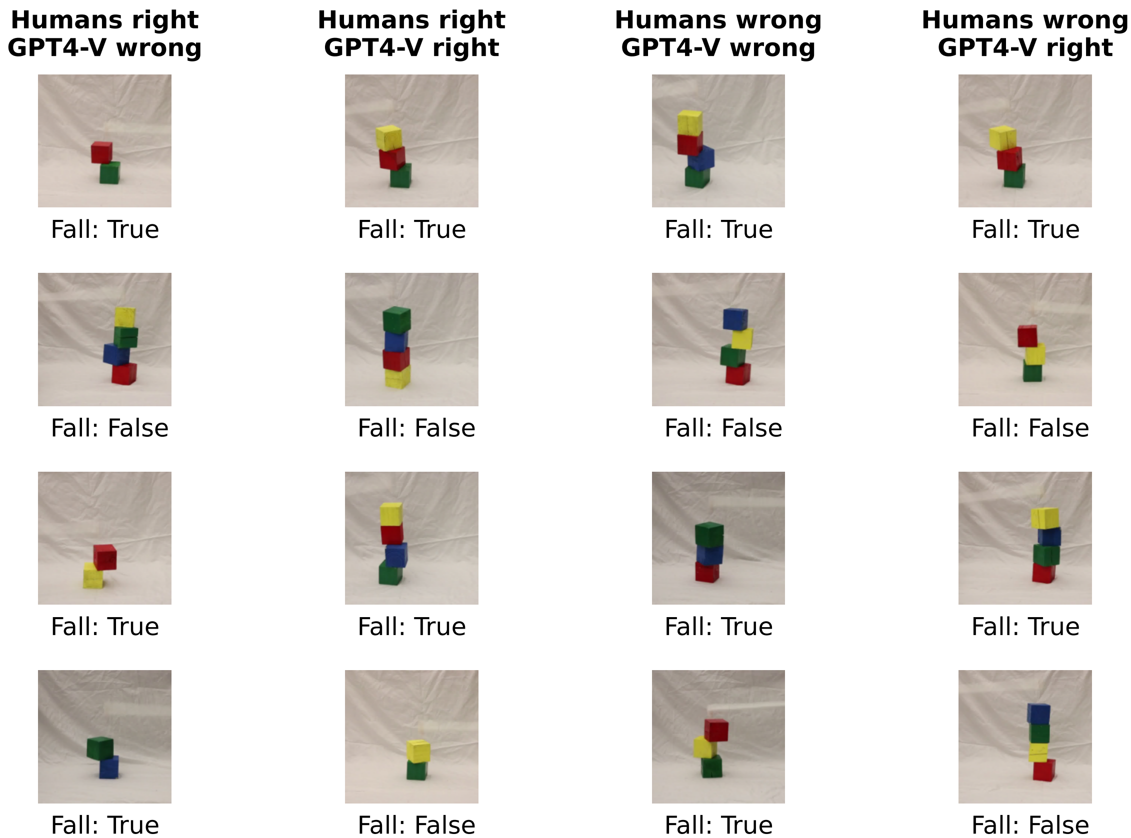


Figure 5. Example images from[86] which are either challenging or easy for Humans, GPT4-V, or both in the binary tower stability judgement task for physical intuition.

A.2 Causal reasoning

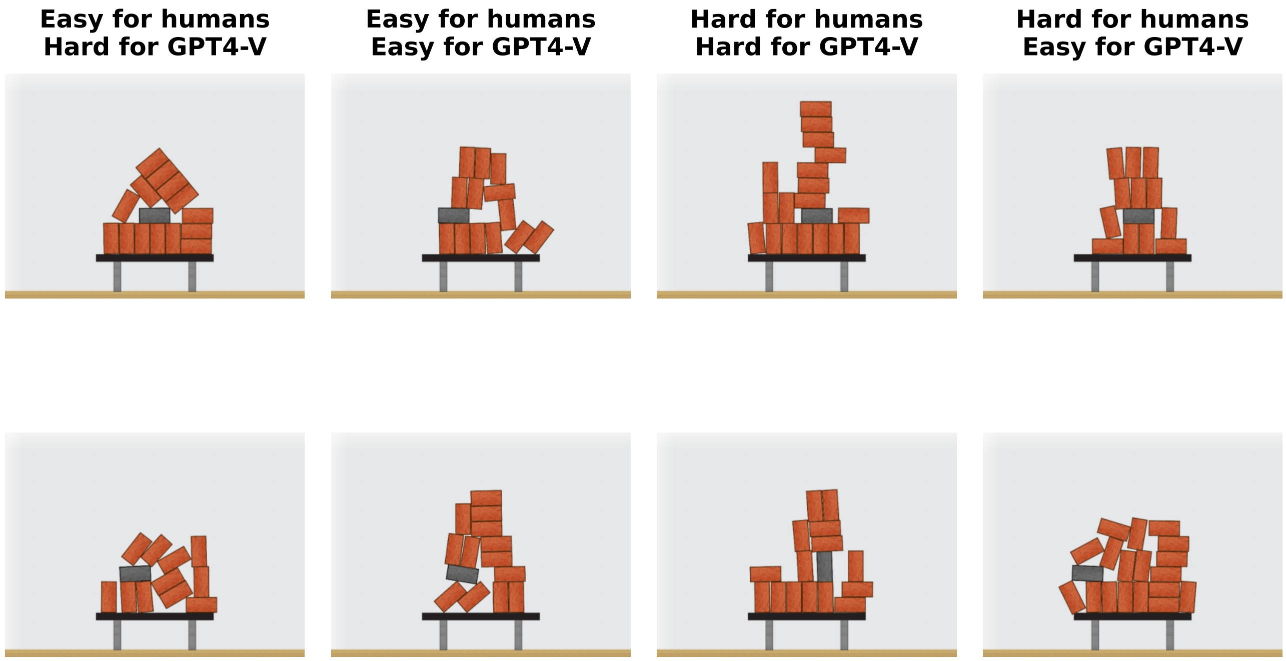
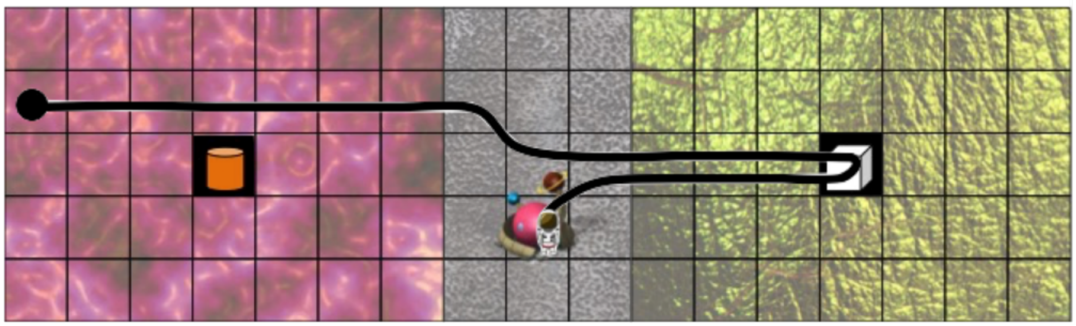


Figure 6. Example images from[87] which are either challenging or easy for Humans, GPT4-V, or both in the number of blocks that will fall task for causal reasoning. Easy here refers to images where judgements have a low average distance to the that the ground truth. Hard refers to images where judgements have a high average distance to the ground truth.

A.3 Intuitive psychology

A



B

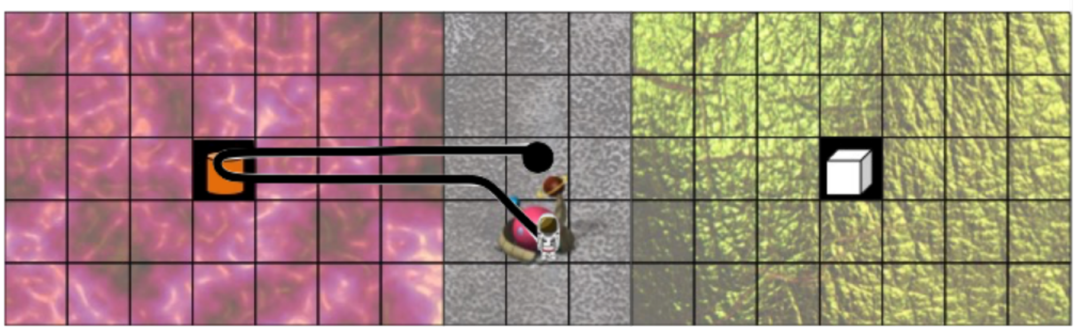
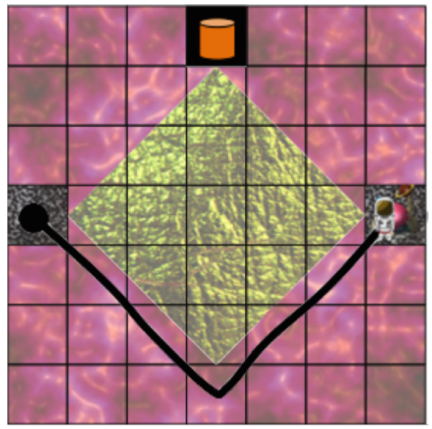


Figure 7. Example images from experiment 1B where the ratings given by GPT4-V match or diverge from human psychological intuition. **A**: GPT4-V answers 7, 5, 2, 7 for left cost, right cost, left reward, right reward. These values make sense given the path of the agent. It crosses straight through the left terrain, indicating that it might have a high cost associated with it and that the agent does not find the left care package rewarding. It then crosses through the yellow terrain to pick up the white care package, indicating that this package has a high reward associated with it. **B**: GPT4-V answers 5, 4, 2, 7 for left cost, right cost, left reward, right reward. These values are counterintuitive. The agent is seen crossing into the left terrain and picking up the left reward, which indicates that the left care package has a significant reward associated with it and that the left terrain should not incurr a large cost upon the agent. GPT4-V however assigns the left terrain a higher cost than the right terrain and also the left care package a lower reward than the right care package.

A



B

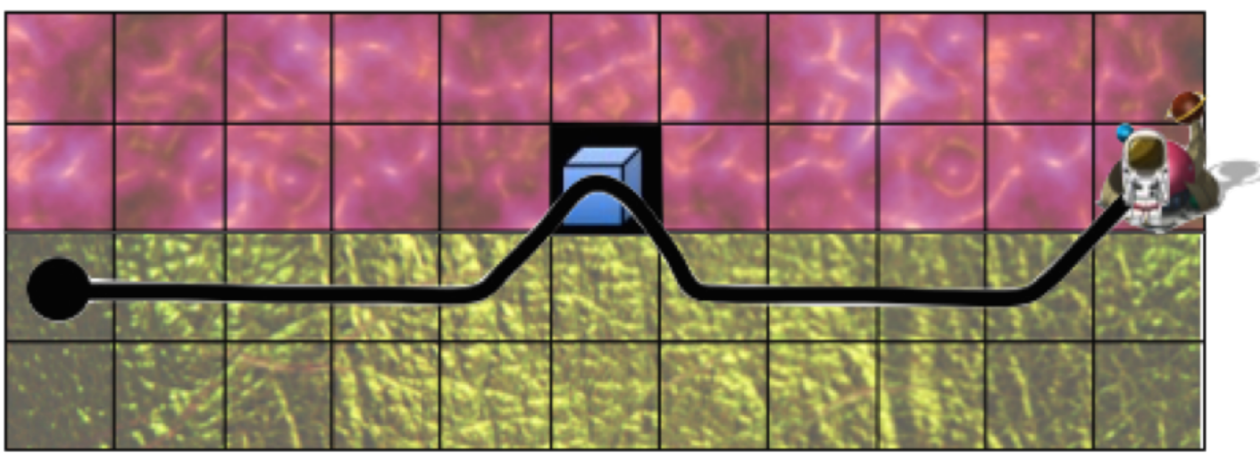


Figure 8. Example images from experiments 1A and 1B. **A**: GPT4-V answers 7, 7, 0 for inside cost, outside cost, top reward. That the care package has no value for the agent is sensible given the image, however we would expect the costs for inside and outside terrains to diverge, as the agent seemingly takes a detour to avoid crossing the inside terrain. **B**: GPT4-V answers 5, 5, 7 for bottom cost, top cost, top reward. The reward here should have a high value seeing as the agent takes a detour to collect the care package. However, the bottom terrain should be associated with a lower cost compared to the top terrain, as the agent is seen taking a detour to cross the bottom terrain instead of the top terrain after picking up the care package.

Prompts

Intuitive physics

For the intuitive physics experiment, we used a task from Lerer et al.[86]. For each trial, we asked the models three questions:

Question 1	What is the background color?
Question 2	How many blocks are in the image?
Question 3	Will this block tower fall? Give a boolean answer.

Causal reasoning

For the causality experiment, we used a task from Zhou et al.[87,89]. For each trial, we again asked the models three questions:

Question 1	How many blocks are there in the image? You are only allowed to answer with a single number. No words allowed!
Question 2	How many of the red bricks would fall off the table if the dark grey brick wasn't there? You are only allowed to answer with a single number between 0 and {num_blocks} corresponding to how many blocks would fall. No words allowed!
Question 3	How responsible is the dark grey brick for the red bricks staying on the table? You are only allowed to answer with a number on a scale from 0% (not at all responsible) to 100% (fully responsible). No words allowed!

Naive Utility Calculus

For the intuitive psychology experiment, we used three tasks from Jara-Ettinger et al.[88]. We again asked the model two descriptive questions in order to assess their basic comprehension of the scene:

Question 1	Please answer the following question with a single color name only: What color is the background of the central part of the image? You are only allowed to answer with a single color name!
Question 2	Please answer the following question with a number only: How many orange or white containers are in the image? You are only allowed to answer with a number!

For experiment 1A we first gave the models the following basic prompt, which was combined with different trial specific questions:

> "This task is about astronauts. The astronauts are exploring planets with alien terrains depicted with different colours and textures. Each astronaut has different skills, making each terrain more or less exhausting or easy for them to cross. All astronauts can ultimately cross all terrains, even if it's exhausting. The astronauts land far from the base and have to walk there. In each image, the black circle on the left indicates where the astronaut landed. The base is on the middle right part of the image. Sometimes care packages are dropped from above and the astronauts can pick them up. There are two kinds of care packages depicted with an orange cylinder and a white cube. Each astronaut has different preferences and likes each kind of care package in different amounts. The astronauts don't actually need the care packages. They can go straight to the base, or they can pick one up. You will see images of different astronauts with different skills and preferences travelling from their landing location to the home base. The astronauts always have a map. So they know all about the terrains and the care packages. Please answer the following question with a number only:"

For experiment 1B we first gave the models the following basic prompt, which was again combined with different trial-specific questions:

> "This task is about astronauts. The astronauts are exploring planets with alien terrains depicted with different colours and textures. Each astronaut has different skills, making each terrain more or less exhausting or easy for them to cross. All astronauts can ultimately cross all terrains, even if it's exhausting. Sometimes, the astronauts land far from the base and have to walk there. In each image, the black circle indicates where the astronaut landed. The base is in the center of the image. Sometimes care packages are dropped from above and the astronauts can pick them up. There are two kinds of care packages depicted with an orange cylinder and a white cube. Sometimes both care packages are identical. The astronauts cannot pick both care packages. Each astronaut has different preferences and likes each kind of care package in different amounts. The astronauts don't actually need the care packages. They can go straight to the base, or they can pick one up. You will see images of different astronauts with different skills and preferences travelling from their landing location to the home base. The astronauts always have a map. So they know all about the terrains and the care packages. Please answer the following question with a number only:"

Finally, for experiment 1C we first gave the models the following basic prompt:

> "This task is about astronauts. The astronauts are exploring planets with alien terrains depicted with different colours and textures. Each astronaut has different skills, making each terrain more or less exhausting or easy for them to cross. All astronauts can ultimately cross all terrains, even if it's exhausting. The astronauts land far from the base and have to walk there. In each image, the black circle on the left indicates where the astronaut landed. The base is on the right part of the image. The path astronauts take from where they land to their base is indicated by a thick black line between the black circle on the left and the astronaut on the right. Sometimes care packages depicted by a blue cube on a black background are dropped from above and the astronauts can pick them up. Each astronaut has different preferences and likes each care package in different amounts. The astronauts don't actually need the care packages. They can go straight to the base, or they can pick one up. You will see images of different astronauts with different skills and preferences travelling from their landing location to the home base. Your task is to judge how easy/exhausting it is for the astronaut in each image to cross each terrain, and how much they like each care package. The astronauts always have a map. So they know all about the terrains and the care packages. Please answer the following question with a number only:"

Experiment 1A

Images	Questions
0, 1	1. How easy is it for the astronaut in this image to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How much does the astronaut in this image like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!
2, 3, 4, 5	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How much does the astronaut like the white care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number! 3. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!"
6, 7, 8, 9, 10	1. How easy is it for the astronaut to cross the purple terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 3. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!
11, 12, 13, 14, 15	1. How easy is it for the astronaut to cross the purple terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 3. How much does the astronaut like the white care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number! 4. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!"

Experiment 1B

Images	Questions
1, 2, 3, 4, 5, 6, 7	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How easy is it for the astronaut to cross the yellow terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 3. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number! 4. How much does the astronaut like the white care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!
8, 9, 10	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How easy is it for the astronaut to cross the yellow terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 3. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!
11, 12, 13, 14, 15, 16, 17	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number! 3. How much does the astronaut like the white care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!

Experiment 1C

Images	Questions
All images	1. How easy is it for the astronaut to cross the yellow terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 3. How much does the astronaut like the blue care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!