nature computational science

Brief Communication

https://doi.org/10.1038/s43588-023-00527-x

Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT

Received: 17 February 2023

Accepted: 5 September 2023

Published online: 5 October 2023

Check for updates

Thilo Hagendorff[1], **Sarah Fabi**[2] & **Michal Kosinski**[3] ✉

We design a battery of semantic illusions and cognitive reflection tests, aimed to elicit intuitive yet erroneous responses. We administer these tasks, traditionally used to study reasoning and decision-making in humans, to OpenAI's generative pre-trained transformer model family. The results show that as the models expand in size and linguistic proficiency they increasingly display human-like intuitive system 1 thinking and associated cognitive errors. This pattern shifts notably with the introduction of ChatGPT models, which tend to respond correctly, avoiding the traps embedded in the tasks. Both ChatGPT-3.5 and 4 utilize the input–output context window to engage in chain-of-thought reasoning, reminiscent of how people use notepads to support their system 2 thinking. Yet, they remain accurate even when prevented from engaging in chain-of-thought reasoning, indicating that their system-1-like next-word generation processes are more accurate than those of older models. Our findings highlight the value of applying psychological methodologies to study large language models, as this can uncover previously undetected emergent characteristics.

As the range of applications for large language models (LLMs) rapidly expands, it is of paramount importance to understand the mechanisms through which LLMs reason and make decisions. Recent research has revealed that with the increasing complexity of LLMs they exhibit a multitude of skills and properties, some of which were not anticipated or intended by their creators[1,2]. Among these newfound abilities are the capacity to generate computer code, tackle mathematical problems, learn from examples, engage in introspection, carry out multistep reasoning, solve theory of mind tasks, deceive other agents and a plethora of other skills[3–6]. In this work, we aim to explore reasoning capabilities in the family of generative pre-trained transformer (GPT) models by OpenAI, while shedding light on the intricacies of their cognitive processes.

Research on humans often distinguishes between two broad categories of reasoning or—more broadly—cognitive processes: systems 1 and 2[7,8]. System 1 processes are fast, automatic and instinctual. They often involve heuristics, or mental shortcuts, which enable quick judgments and decisions without conscious effort. System 1 is essential for everyday functioning, as it allows humans to navigate their environments and make rapid decisions with minimal effort. System 2 processes, on the other hand, are deliberate and require conscious effort. This system is employed in logical reasoning, critical thinking and problem-solving. System 2 processes are slower and more resource intensive, but they are also more accurate and less susceptible to bias.

On the surface, current-day LLMs seem to be system 1 thinkers: the input text is processed by consecutive layers of neurons to produce a distribution of probabilities of all possible single-token (word) completions. This process is automatic and unidirectional, and involves a single wave of propagation through the neural network for each consecutive predicted word. Yet, past research and the results presented here suggest that, like humans, LLMs can also engage in system-2-like cognitive processes[4]. While generating each consecutive word, LLMs re-read their context window, including the task provided by a user, as well as the words they have thus far generated. As a result, LLMs can employ their context window as a form of an external short-term memory to engage

---

[1]University of Stuttgart, Stuttgart, Germany. [2]University of California San Diego, San Diego, CA, USA. [3]Stanford University, Stanford, CA, USA.
✉e-mail: michalk@stanford.edu

in chain-of-thought reasoning, re-examine the starting assumptions, estimate partial solutions or test alternative approaches. This is akin to how people use notepads to solve mathematical problems or write essays to sharpen and develop their arguments.

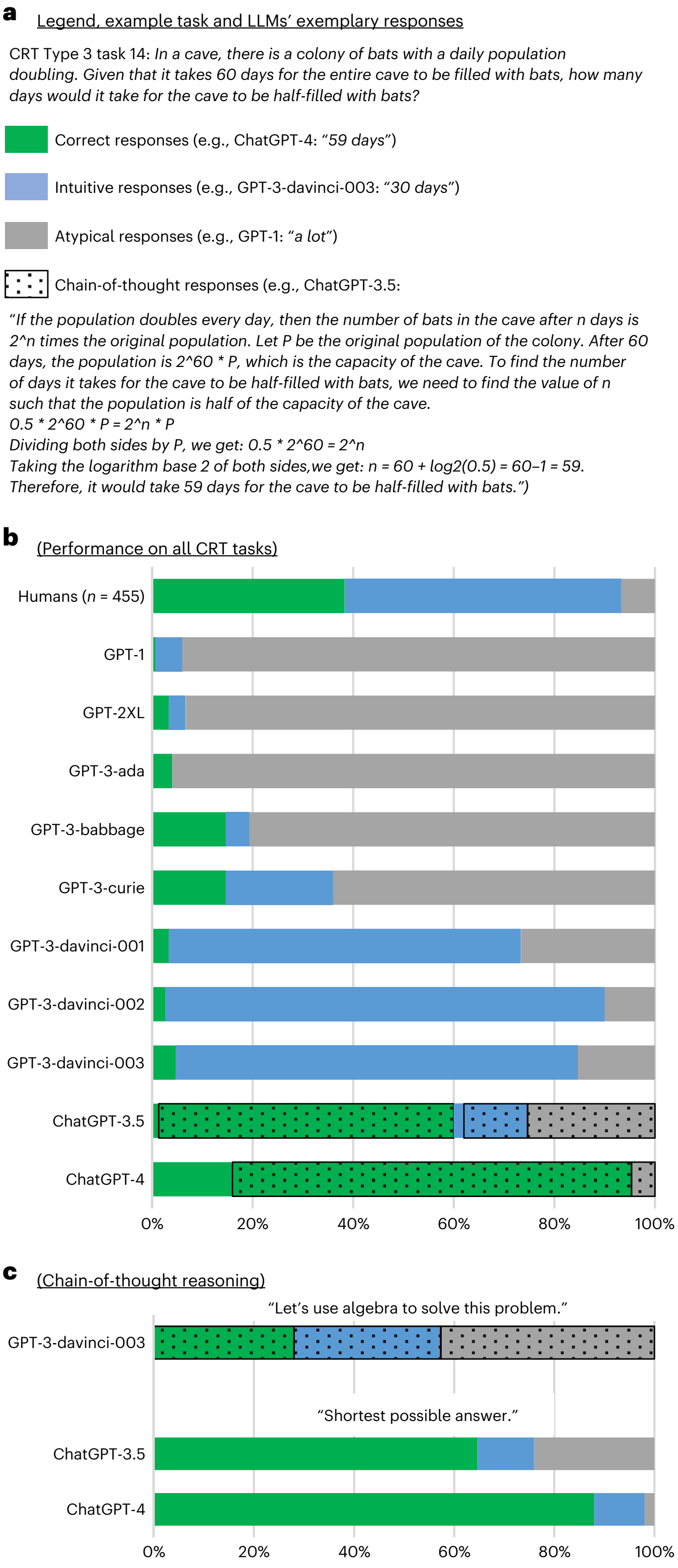
In this work, we build on psychological research on human reasoning and decision-making to explore system 1 and 2 processes in LLMs. We examine the performance of humans ($n = 455$) and ten OpenAI LLMs (ranging from GPT-1 to ChatGPT-4[3,9–11]) using tasks typically employed to test reasoning and decision-making in humans: cognitive reflection test (CRT)[12] tasks and semantic illusions[13] (see Supplementary Information and Supplementary Table 1 for more details). The CRT comprises three types of mathematical tasks that appear to be simpler than they really are, thus triggering an intuitive but incorrect system 1 response. CRT type 1 tasks, such as the widely known 'A bat and a ball' task, use a 'more than' phrase to trick participants into subtracting two of the values rather than solving a somewhat more complex equation. Type 2 tasks exploit people's tendency to complete a numerical triplet series, such as five machines making five widgets in five minutes because two machines make two widgets in two minutes. Type 3 tasks describe an exponential process but trick the participants into treating it as linear. Solving CRT tasks correctly requires engaging in deliberate system 2 reasoning or possessing well developed system 1 intuitions. Semantic illusions are questions containing a disguised error aimed at triggering an intuitive but incorrect system 1 response. In the well known Moses Illusion[13], for example, participants tend to be tricked into claiming that Moses took two animals of each kind on the Ark (when in fact it was Noah).

We address some of the limitations of past studies. First, while past research focused on a single model (GPT-3), we study reasoning capabilities across a range of models of different sizes and complexities. Second, as solving the CRT tasks requires mathematical abilities, LLMs' performance could be limited by their mathematical skills. To address this issue, we complement the CRT tasks with semantic illusions that do not rely on mathematical skills. Third, past research relied on three CRT tasks copied verbatim from human studies[14]. This is problematic, as observing LLMs' performance on three CRT tasks does not allow for meaningful statistical comparisons. Moreover, these tasks (as well as their solutions) were likely present in the LLMs' training data. To circumvent these issues, we designed 50 bespoke versions of each type of task (200 in total).

For brevity and convenience, we use words such as 'behavior', 'intuition', 'deliberation' or 'ability' when referring to LLMs, yet we do not mean to equate artificial intelligence (AI) and human cognitive processes. While AI's outputs are often similar to ones produced by humans, it typically operates in fundamentally different ways.

First, we present the results of study 1: the cognitive reflection test. To help the reader interpret the results, we discuss them in the context of LLMs' exemplary responses to one of the CRT tasks (Fig. 1a). The correct response to this task is '59 days', but it was designed to appear easier than it really is, tempting participants to simply divide the total time by two, triggering an intuitive (but incorrect) response of '30 days'.

The performance of humans and LLMs across 150 CRT tasks is presented in Fig. 1b. There are four distinct trends. First, most of the responses of early and smaller LLMs (up until GPT-3-curie) were atypical. This category includes responses that were evasive (for example, GPT-1's response "a lot"), which indicated failure to comprehend the task (for example, GPT-2XL's response "The colony would take 60 days to double in size") or which were incorrect in ways different from one that the task was designed to trigger (for example, GPT-3-babbage's response: "It would take about 10 days for the cave to be half-filled with bats"). Moreover, while 15% of responses of both GPT-3-babbage and GPT-3-curie were categorized as correct, they seemed accidental: all but one were given CRT type 2 tasks, which can be solved by simply repeating the number mentioned most frequently in the prompt—which these models tended to naively do in this and other tasks.



CRT Type 3 task 14: *In a cave, there is a colony of bats with a daily population doubling. Given that it takes 60 days for the entire cave to be filled with bats, how many days would it take for the cave to be half-filled with bats?*

🟩 Correct responses (e.g., ChatGPT-4: "*59 days*")

🟦 Intuitive responses (e.g., GPT-3-davinci-003: "*30 days*")

⬜ Atypical responses (e.g., GPT-1: "*a lot*")

▨ Chain-of-thought responses (e.g., ChatGPT-3.5:

*"If the population doubles every day, then the number of bats in the cave after n days is 2^n times the original population. Let P be the original population of the colony. After 60 days, the population is 2^60 * P, which is the capacity of the cave. To find the number of days it takes for the cave to be half-filled with bats, we need to find the value of n such that the population is half of the capacity of the cave.*
*0.5 * 2^60 * P = 2^n * P*
*Dividing both sides by P, we get: 0.5 * 2^60 = 2^n*
Taking the logarithm base 2 of both sides,we get: n = 60 + log2(0.5) = 60–1 = 59.
Therefore, it would take 59 days for the cave to be half-filled with bats.")

Fig. 1 | Human and LLM performance on the CRT tasks. a, Exemplary responses to one of the CRT tasks, categorized as correct, intuitive (but incorrect) and atypical (that is, all other incorrect responses). Within each category, the responses that were preceded by written chain-of-thought reasoning were additionally labeled as 'chain-of-thought responses'. **b**, Human and LLM performance on 150 CRT tasks. **c**, LLMs' responses when instructed to engage or prevented from engaging in chain-of-thought reasoning. The data source file includes 95% confidence intervals.

Second, as the models grew larger and their ability to comprehend the task increased, atypical responses were replaced by intuitive (but incorrect) responses, which the tasks were designed to trigger (for example, GPT-3-davinci-003: "30 days"). These constituted below 5% of responses of early models (up to GPT-3-babbage) and increased to 21% for GPT-3-curie (difference $(\frac{\text{total price}-\text{known price}}{2}) = 16\%$; $\chi^2(1) = 16.98$; $P < 0.001$) and to 70%–90% for the GPT-3-davinci family ($\delta \geq 49\%$; $\chi^2(1) \geq 69.64$; $P < 0.001$), a fraction much higher than observed in humans (55%; $\delta \geq 15\%$; $\chi^2(1) \geq 11.79$; $P < 0.001$).

In humans, intuitive but incorrect responses are interpreted as evidence of system 1 reasoning and failure to engage system 2, but they could also stem from deliberate—yet erroneous—system 2 reasoning. The generative process behind the LLMs' responses is less ambiguous. As we discuss in the introduction, current-day LLMs lack the built-in cognitive infrastructure necessary to internally engage in system 2 processes. Consequently, their intuitive responses can only stem from a system-1-like process.

Importantly, LLMs' intuitive responses are unlikely to be driven by insufficient mathematical ability. First, previous research has shown that LLMs can solve basic mathematical problems[1]. Second, intuitive responses to the CRT type 1 and 3 tasks also require solving a simple equation (for example, how much is 'half of 60'; responding intuitively to the CRT type 2 tasks requires no computation). Moreover, as we show in study 3, GPT-3-davinci-003's performance can be substantially improved by presenting it with training examples.

Third, LLMs' strong tendency to respond intuitively stops abruptly with the arrival of ChatGPT. The fraction of correct responses was equal to 59% for ChatGPT-3.5 and 96% for ChatGPT-4. This is much higher than the 5% of tasks solved correctly by GPT-3-davinci-003, an otherwise very apt model ($\delta \geq 54\%$; $\chi^2(1) \geq 102.44$; $P < 0.001$), or 38% achieved by humans ($\delta \geq 21\%$; $\chi^2(1) \geq 25.60$; $P < 0.001$). ChatGPT's tendency to respond correctly was accompanied by a substantial drop in its tendency to respond intuitively: 15% for ChatGPT-3.5 and 0% for ChatGPT-4 versus 80% for GPT-3-davinci-003 ($\delta \geq 65\%$; $\chi^2(1) \geq 125.81$; $P < 0.001$) and 55% for humans ($\delta \geq 40\%$; $\chi^2(1) \geq 86.30$; $P < 0.001$).

A closer look at ChatGPT models' responses reveals that the marked increase in their performance was accompanied by a novel response style. While pre-ChatGPT LLMs responded with brief phrases or single sentences (for example, GPT-3-davinci-003: "30 days"), 97% of ChatGPT-3.5's responses and 85% of ChatGPT-4.5's responses included some form of chain-of-thought reasoning (Fig. 1a).

As we discussed before, there is nothing deliberate in how ChatGPT-3.5 and other current-day LLMs generate the next word. Yet, each time the word is generated, an LLM re-reads the task and the response it has generated so far, combining system-1-like next-word generation into something resembling a system 2 process: generating a strategy needed to solve the task, dividing the task into more tractable subtasks and solving them one by one. This is akin to how humans use a notebook to solve mathematical tasks without the need to process them in their short-term memory.

Next, we show that chain-of-thought responses (study 2) not only resemble but also serve as system 2 processes, in line with previous studies showing that instructing LLMs to think step by step improves their ability to solve various tasks[4].

We first show that GPT-3-davinci-003's accuracy increases when it is instructed to engage in chain-of-thought reasoning. We present it with the CRT tasks suffixed with "Let's use algebra to solve this problem". The results presented in Fig. 1c show that our manipulation was successful: the fraction of chain-of-thought responses increased from 0% in study 1 to 100% ($\delta = 100\%$; $\chi^2(1) = 147.01$; $P < 0.001$). The model seemed to design and execute a task-solving strategy. Most of the time, this strategy was poorly conceived or executed, leading to the increase of atypical responses from 15% to 43% ($\delta = 28\%$; $\chi^2(1) = 14.72$; $P < 0.001$). Yet, in other cases, the strategy was sound, boosting the fraction of correct responses from 5% to 28% ($\delta = 23\%$; $\chi^2(1) = 28.20$; $P < 0.001$) and reducing the model's tendency to fall for the trap embedded in the task: intuitive responses dropped from 80% to 29% ($\delta = 51\%$; $\chi^2(1) = 75.66$; $P < 0.001$).

Next, we show that preventing the model from engaging in chain-of-thought reasoning can decrease its ability to solve the tasks. We presented ChatGPT models with the CRT tasks suffixed with "Provide the shortest possible answer (for example, '$2' or '1 week'), do not explain your reasoning". The results presented in Fig. 1 show that our manipulation was again successful: the fraction of chain-of-thought responses fell from 97% to 0% for ChatGPT-3.5 ($\delta = 97\%$; $\chi^2(1) = 276.79$; $P < 0.001$) and from 84% to 0% for ChatGPT-4 ($\delta = 84\%$; $\chi^2(1) = 213.81$; $P < 0.001$). The fraction of correct responses did not change for ChatGPT-3.5 ($\delta = 4\%$; $\chi^2(1) = 0.47$; $P = 0.49$). For ChatGPT-4, it fell from 95% to 88% ($\delta = 7\%$; $\chi^2(1) = 4.36$; $P < 0.05$), accompanied by an increase in intuitive responses from 0% to 10% ($\delta = 10\%$; $\chi^2(1) = 13.75$; $P < 0.001$).

The results of study 2 suggest that chain-of-thought reasoning helps LLMs to avoid falling for the traps embedded in the CRT tasks and improves their ability to solve them correctly. Yet, they also reveal that ChatGPT models could solve the great majority of the CRT tasks even when forced to provide a system-1-like response. This is consistent with ChatGPT-4's performance in study 1, where it solved 24% of the CRT task without using chain-of-thought reasoning.

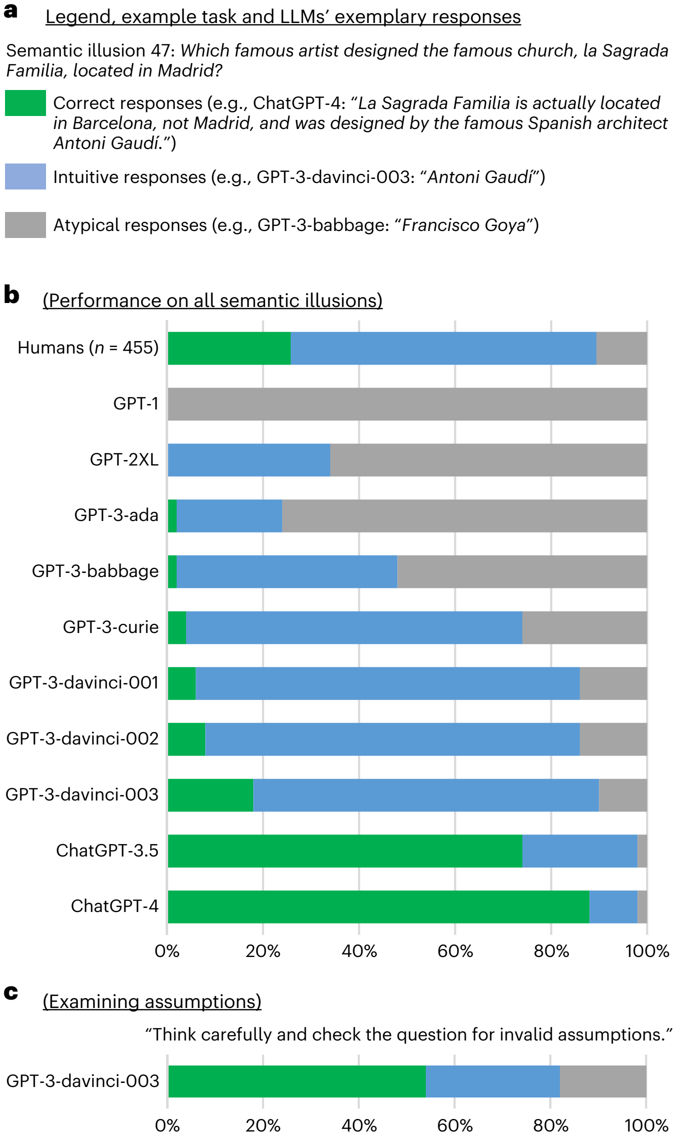
In humans, this would be taken as evidence of a well developed intuition stemming from previous exposure to similar tasks[15] (although the persistence and size of this effect is disputed[16]). Here we show results suggesting that the same applies to LLMs. This is in line with past results showing that LLMs can learn, even from a single example[3].

As ChatGPT models already seem to possess well developed intuition, we attempt to improve the system-1-like responses of GPT-3-davinci-003 (study 3). We present it with each of the CRT tasks, each time preceding this with 0 to 49 remaining tasks of the same type, accompanied by the correct solution. The CRT tasks of the same type are semantically very similar, enabling the model to develop system-1 intuitions akin to that expressed by the ChatGPT model family.

The results presented in Extended Data Fig. 1 show that GPT-3-davinci-003's ability to answer correctly (rather than intuitively) increased with each additional example. The fastest gains were observed for the CRT type 2 tasks, where the accuracy increased from 2% to 92% after two examples ($\delta = 90\%$; $\chi^2(1) = 77.72$; $P < 0.001$). This is to be expected, as they can be solved correctly by simply repeating the duration listed in the task. The CRT type 3 tasks, solvable by reporting the total time minus one unit, proved to be somewhat more complex: the accuracy increased from 12% to 92% after seven training examples ($\delta = 80\%$; $\chi^2(1) = 60.94$; $P < 0.001$). It took most examples to develop the model's intuition to solve the CRT type 1 tasks, where the correct answer is equal to $\frac{\text{total price}-\text{known price}}{2}$. However, even here, the model's accuracy increased from 0% to 78% after 30 examples ($\delta = 78\%$; $\chi^2(1) = 60.70$; $P < 0.001$).

The CRT tasks employed in studies 1–3 rely heavily on mathematical skills and are highly semantically uniform. To ensure that the results generalize beyond the CRT tasks, we replicate studies 1–3 using much more semantically diverse semantic illusions (study 4). Similarly to the CRT tasks, semantic illusions contain a disguised error aimed at triggering an intuitive but incorrect system 1 response. Unlike the CRT tasks, semantic illusions do not require mathematical skills, instead relying on participants' general knowledge.

To help the reader interpret the results, we discuss them in the context of LLMs' exemplary responses to semantic illusion 47 (Fig. 2a). As in the context of the CRT tasks, responses were divided into three categories: intuitive, correct and atypical. The question was designed to trigger an intuitive system-1 response 'Antoni Gaudí' while overlooking the embedded invalid assumption (la Sagrada Família is in Barcelona). Importantly, responding 'Antoni Gaudí' can be treated



a Legend, example task and LLMs' exemplary responses

Semantic illusion 47: *Which famous artist designed the famous church, la Sagrada Familia, located in Madrid?*

🟩 Correct responses (e.g., ChatGPT-4: "*La Sagrada Familia is actually located in Barcelona, not Madrid, and was designed by the famous Spanish architect Antoni Gaudí.*")

🟦 Intuitive responses (e.g., GPT-3-davinci-003: "*Antoni Gaudí*")

⬜ Atypical responses (e.g., GPT-3-babbage: "*Francisco Goya*")

b (Performance on all semantic illusions)

c (Examining assumptions)

"Think carefully and check the question for invalid assumptions."

Fig. 2 | Human and LLM performance on semantic illusions. a, Exemplary responses to one of the semantic illusions, categorized as correct, intuitive and atypical. **b**, Human and LLM performance on 50 semantic illusions. **c**, GPT-3-davinci-003's responses when instructed to examine the task's assumptions. The data source file includes 95% confidence intervals.

as indicative of system 1 processing only if the respondent has the knowledge necessary to recognize the error. Thus, given an intuitive response, the model was reset, and its underlying knowledge was tested using an additional question (here "Where is the famous church, la Sagrada Família, located?"; see Supplementary Information for the list of knowledge questions). Intuitive responses given by LLMs that failed this post hoc test were recategorized as atypical, along with responses revealing a further lack of necessary knowledge (for example, GPT-3-babbage: "Francisco Goya") and nonsensical responses (for example, GPT-1: "the church of san francisco"). Responses recognizing the invalid assumption were categorized as correct.

The results presented in Fig. 2b show a pattern similar to one observed in study 1. Most of the responses of early and smaller LLMs (up to GPT-3-babbage) were atypical (gray bars), as they struggled to comprehend the question or lacked the necessary knowledge. As LLMs grew in size and overall ability, the fraction of atypical responses fell from 52% for GPT-3-babbage to 10% for GPT-3-davinci-003 ($\delta = 42\%$; $\chi^2(1) = 18.70$; $P < 0.001$). They were replaced by intuitive responses (blue bars): GPT-3-davinci-003 fell for the semantic illusion 72% of the time. As in the CRT tasks, this trend changes markedly with the introduction of ChatGPT. The fraction of correct responses increased from 18% for GPT-3-davinci-003 to 74% and 88% for ChatGPT-3.5 and ChatGPT-4, respectively (green bars; $\delta \geq 56\%$; $\chi^2(1) = 29.35$; $P < 0.001$). As we discussed before, there is nothing deliberate in LLMs' next-word generation process, yet this system-1-like process proved to be very apt at detecting invalid assumptions embedded in semantic illusions.

The results of studies 2 and 3 suggest that LLMs' propensity to commit reasoning errors in the CRT tasks can be reduced by instructing them to examine the task more carefully and providing them with examples of correct solutions to similar tasks. Study 5 replicates these results in the context of semantic illusions.

We first add the suffix "Think carefully and check the question for invalid assumptions" to each semantic illusion and administer them to GPT-3-davinci-003. The results presented in Fig. 2c show that the fraction of correct responses increased threefold, from 18% in study 4 to 54% ($\delta = 36\%$; $\chi^2(1) = 12.54$; $P < 0.001$), while the fraction of intuitive responses decreased from 72% to 28% ($\delta = 44\%$; $\chi^2(1) = 17.64$; $P < 0.001$).

Next, as in study 3, we precede each semantic illusion with 0 to 49 other semantic illusions, accompanied by the correct solution. The results presented in Extended Data Fig. 1 show that GPT-3-davinci-003's ability to answer correctly increased from 18% for zero examples to over 64% for ten and more examples ($\delta \geq 46\%$; $\chi^2(1) = 20.01$; $P < 0.001$).

Discussion

Our results reveal an interesting pattern. As LLMs' ability to comprehend the tasks increases, they tend to fall for the traps embedded in the tasks. This, in humans, would be interpreted as evidence of fast, automatic and instinctual system 1 processing. The most able of the pre-ChatGPT models, GPT-3-davinci-003, decisively outpaces humans in its tendency to respond intuitively rather than correctly. Yet, this changes abruptly with the arrival of ChatGPT models. They responded correctly to a great majority of tasks, decisively outperforming humans in their ability to avoid traps embedded in the tasks.

How would we explain pre-ChatGPT models' tendency to respond intuitively, despite their sufficient mathematical abilities and factual knowledge demonstrated in studies 3–5? As we discuss in the introduction, LLMs lack the cognitive infrastructure necessary to engage in system 2 processes, which humans may employ when answering such questions. Thus, in the absence of well developed intuition or explicit chain-of-thought reasoning, they are particularly prone to fall for the traps embedded in the tasks.

Furthermore, how would we explain the steep shift in accuracy between GPT-3 and ChatGPT? ChatGPT models tend to engage in chain-of-thought reasoning: the models use their input–output context window to develop strategies needed to solve the task, examine the starting assumptions, estimate partial solutions or test alternative approaches—in a way akin to how people use notepads to solve mathematical problems or write essays to develop their arguments. Instructing an older model (that is, GPT-3-davinci-003) to engage in chain-of-thought reasoning substantially boosts its performance.

Yet, chain-of-thought reasoning cannot be the sole explanation. ChatGPT models' accuracy barely drops when they are prevented from engaging in chain-of-thought reasoning. This suggests that they have well developed intuitions enabling them to solve tasks without engaging system-2-like processes. This is confirmed by results showing that GPT-3-davinci-003's performance can be substantially increased by presenting it with example tasks and their correct solutions.

Some progress is to be expected. In humans, the CRT and semantic illusions are good predictors of an ability to engage in unbiased, reflective and rational decision-making[17], as well as overall cognitive ability[12]. Thus, LLMs' ability to solve the CRT and semantic illusions should increase as their overall ability increases. Yet, the shift observed in this study seems to be steeper than the increase in LLMs' overall abilities.

We can only speculate on this, given that OpenAI does not provide their models in open access and only shares limited information on their technical specification and training process. First, it is unlikely that the shift was driven merely by larger model size. According to OpenAI, ChatGPT-3.5-turbo was derived from text-davinci-003 by fine-tuning it for chat. The two models are likely of similar sizes. Second, it could be that the shift was driven by the employment of reinforcement learning from human feedback[18,19]. In reinforcement learning from human feedback, human-written demonstrations on example prompts are used to train supervised learning baselines. Next, human 'AI trainers' rank model outputs on a larger set of prompts, and a reward model is trained to predict their preferences. This reward model is then used to fine-tune the models using Proximal Policy Optimization algorithms. While reinforcement learning from human feedback has been employed since GPT-3 text-davinci-002[19], this procedure was enhanced in ChatGPT training: AI trainers played both sides: the user and an AI assistant[20]. Next, it is likely that ChatGPT models were exposed to sufficient CRT-like tasks in their training to be able to respond to them intuitively. Those tasks are highly semantically similar and, as illustrated by study 3, exposure to training examples can rapidly boost an LLM's accuracy. This explanation is less likely in the context of semantic illusions, which are much more irregular and diverse. This question will hopefully be addressed by further research or more transparency in LLM development.

Next to the analysis of LLM performance on reasoning tasks, one can approach the issue from a normative perspective, asking whether phenomena of intuitive decision-making are desirable in LLMs. In the cognitive science literature, researchers stress that the notion of intuitive errors relies on a normative concept of logics and statistics, which can be inappropriate for real-world situations. Instead, decision-making processes should be evaluated in the sense of 'ecological rationality', meaning on the basis of how well they fit the structure of the environment in which they occur[21]. In this vein, the CRT tasks as well as semantic illusions create a 'hostile' test environment, which intentionally aims to mislead humans. Should LLMs perhaps go with the conversational flow and just 'overlook' small mistakes instead of correcting factually incorrect questions, as humans tend to do? Or should they insist on correcting mistakes (as ChatGPT models often did in our study), so as to minimize the inaccuracies, 'hallucinated' outputs, factual incorrectness and misinformation—a major problem in LLM use?

We list a few of the limitations of our study. First, it is limited to OpenAI's GPT family of models. There are many other models, including non-English-language models, whose functioning should be studied. Second, our study was limited to just two types of tasks; future work should examine other tasks or real-world examples. Third, we focused on LLMs' observable behavior; it would be useful to study the patterns of their neural activations. Finally, many of our tasks—and particularly the CRT type 1 and 2 tasks—were highly schematic. It is possible that some models encountered enough examples in their training to solve them 'from memory'.

The progress in LLMs not only increased their capabilities, but also reduced our ability to anticipate their properties and behavior. It is increasingly difficult to study LLMs through the lenses of their architecture and hyperparameters. Instead, as we show in this work, LLMs can be studied using methods designed to investigate another capable and opaque structure, namely the human mind. Our approach falls within a quickly growing category of studies employing classic psychological tests and experiments to probe LLM 'psychological' processes, such as judgment, decision-making and cognitive biases[14,22–24].

Methods

Tasks

Hypothesis-blind research assistants recruited on Upwork, a freelancing platform, prepared 50 semantic illusions and 50 CRT type 3 tasks. The CRT type 1 and 2 tasks were generated automatically. All tasks can be found in Supplementary Information.

Testing LLM performance

The tasks were administered to the family of OpenAI GPT models ranging from GPT-1 to ChatGPT-4[3,9–11]. To minimize the variance in the models' responses and thus increase the replicability of our results, the 'temperature' parameter was set to 0. For ChatGPT models, the default ('You are a helpful assistant.') system message was used. The task was prefixed by 'Question:' and suffixed with '\nAnswer:' for other models. As specified in the main text, in some experiments, additional suffixes were added to the tasks, such as 'Let's use algebra to solve this problem'. The models' response length was set to 100 tokens but was extended if needed. The responses were trimmed once they started repeating themselves or stopped responding to the task. The LLMs' responses were reviewed and scored manually.

Testing human performance

The same tasks were also administered to 500 human participants recruited on Prolific.io on 10 June 2023 (50% female). Each participant was presented with a random set of four tasks (one of each kind) followed by a control question inquiring whether they used a language model or another external resource; 45 participants responded positively and were excluded from the analysis. Human respondents' performance suggests that our tasks were of similar difficulty to those used in past human studies. In the CRT, 38% of responses were correct, compared with 41% in the original study ($n = 3{,}428$)[9] ($\delta = 3\%$; $\chi^2(1) = 3.60$; $P = 0.06$). In semantic illusions, 64% of participants responded intuitively, compared with 52% in the original study ($n = 61$; they did not report the fraction of correct responses ($\delta = 12\%$; $\chi^2(1) = 2.41$; $P = 0.12$))[13].

Statistics and reproducibility

Proportions were compared using the prop.test() function in R[25]. All statistical tests were two sided. No statistical method was used to predetermine the number of tasks. The number of human respondents was chosen to enable the detection of small effects (Cohen's $h = 0.2$) with the power of 0.8 at the significance level of 0.05. The resulting desired number of total responses to each test ($n = 196$) was multiplied by 2.5 to account for potential dropouts.

Ethics

The study was executed in strict adherence to ethical guidelines and standards. Our procedures were reviewed and approved by Stanford University's institutional review board. All participants were made fully aware of the nature and objectives of the study and provided informed consent.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets are publicly available at https://osf.io/w5vhp. Source data for all figures are available with this Brief Communication.

Code availability

Our code is publicly available at https://osf.io/w5vhp (ref. 26).

References

1. Wei, J. et al. Emergent abilities of large language models. *Transactions on Machine Learning Research* (2022).
2. Schaeffer, R., Miranda, B. & Koyejo, S. Are emergent abilities of large language models a mirage? Preprint at https://arxiv.org/abs/2304.15004 (2023).
3. Brown, T. B. et al. Language models are few-shot learners. Preprint at https://arxiv.org/abs/2005.14165 (2020).

4. Wei, J. et al. Chain of thought prompting elicits reasoning in large language models. *36th Conference on Neural Information Processing Systems* (2022).
5. Hagendorff, T. Deception abilities emerged in large language models. Preprint at https://arxiv.org/abs/2307.16513 (2023).
6. Kosinski, M. Theory of mind might have spontaneously emerged in large language models. Preprint at https://arxiv.org/abs/2302.02083 (2023).
7. Tversky, A. & Kahneman, D. *Science* **185**, 1124–1131 (1974).
8. Sloman, S. A. *Psychol. Bull.* **119**, 3–22 (1996).
9. OpenAI. GPT-4 Technical Report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
10. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. Preprint at https://paperswithcode.com/paper/improving-language-understanding-by (2018).
11. Radford, A. et al. Language models are unsupervised multitask learners. Preprint at https://paperswithcode.com/paper/language-models-are-unsupervised-multitask (2019).
12. Frederick, S. *J. Econ. Perspect.* **19**, 25–42 (2005).
13. Erickson, T. D. & Mattson, M. E. *J. Verbal Learn. Verbal Behav.* **20**, 540–551 (1981).
14. Binz, M. & Schulz, E. *Proc. Natl Acad. Sci. USA* **120**, e2218523120 (2023).
15. Stieger, S. & Reips, U.-D. *PeerJ* **4**, e2395 (2016).
16. Bialek, M. & Pennycook, G. *Behav. Res. Methods* **50**, 1953–1959 (2018).
17. Toplak, M. E., West, R. F. & Stanovich, K. E. *Mem. Cogn.* **39**, 1275–1289 (2011).
18. Christiano, P. et al. Deep reinforcement learning from human preferences. *31st Conference on Neural Information Processing Systems* (2017).
19. Ouyang, L. et al. Training language models to follow instructions with human feedback. *36th Conference on Neural Information Processing Systems* (2022).
20. OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt (2022).
21. Todd, P. M. & Gigerenzer, G. *Ecological Rationality: Intelligence in the World* (Oxford Univ. Press, 2012).
22. Hagendorff, T. Machine psychology: investigating emergent capabilities and behavior in large language models using psychological methods. Preprint at https://arxiv.org/abs/2303.13988 (2023).
23. Jones, E. & Steinhardt, J. Capturing failures of large language models via human cognitive biases. *36th Conference on Neural Information Processing Systems* (2022).
24. Nye, M., Tessler, M. H., Tenenbaum, J. B. & Lake, B. M. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *35th Conference on Neural Information Processing Systems* (2021).
25. R Core Team. R: a language and environment for statistical computing. https://www.R-project.org/ (2021).
26. Hagendorff, T., Fabi, S. & Kosinski, M. Code and author notes for: human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. https://doi.org/10.17605/OSF.IO/W5VHP (2023).

Acknowledgements

T.H. was supported by the Ministry of Science, Research and Arts Baden-Württemberg under Az. 33-7533-9-19/54/5 in Reflecting Intelligent Systems for Diversity, Demography, and Democracy (IRIS3D) as well as the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

Author contributions

T.H., M.K. and S.F. wrote the paper, collected and analyzed the data, and interpreted the results. T.H. and M.K. share first authorship.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43588-023-00527-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-023-00527-x.

Correspondence and requests for materials should be addressed to Michal Kosinski.

Peer review information *Nature Computational Science* thanks Ryan L. Boyd, Wim De Neys and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Fernando Chirigati, in collaboration with the *Nature Computational Science* team.

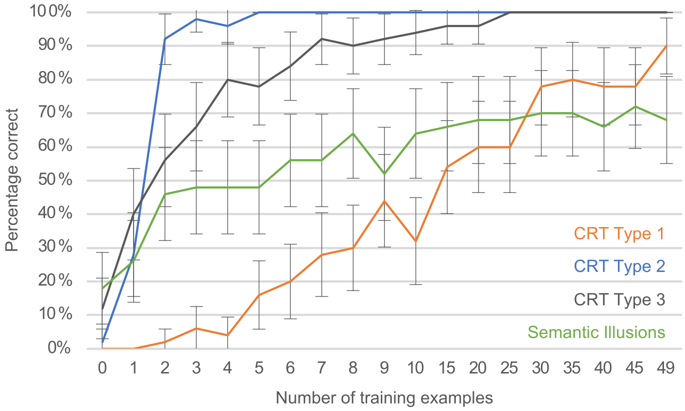
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

https://doi.org/10.1038/s43588-023-00527-x



Extended Data Fig. 1 | Learning curves. Change in the fraction of GPT-3-davinci-003's correct responses against the number of training examples that the task was prefixed with. Error bars represent 95% confidence intervals.

nature portfolio

Corresponding author(s): Michal Kosinski

Last updated by author(s): Sep 1, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected using OpenAI API, HuggingFace API. Human participants were recruited on Prolific.co
Data analysis	We used GPT-1, GPT-2XL, GPT-3-[ada-001/babbage-001/curie-001/davinci-00[1/2/3]], ChatGPT-[3.5-turbo/4] to generate responses. Data was collected using openai (v. 0.27.1) and requests (v 2.25.1) Python libraries. Data was analysed in Python (3.9.13) and R (4.3.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets are publicly available at https://osf.io/w5vhp/.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	We did not collect information about participants' sex or gender.
Population characteristics	50% of participants were female according to Prolific.co
Recruitment	Participants were recruited on Prolific.co
Ethics oversight	The study was reviewed and approved by Stanford University IRB.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A set of 200 reasoning tasks was administered to a range of LLMs and 500 human participants to study the accuracy of their responses.
Research sample	We recruited a sample of 500 anonymous workers on Prolific.co. The manuscript describes the power analysis.
Sampling strategy	Random subset of the US participants (50%/50% male and female) were recruited by the platform.
Data collection	Data was collected on Prolific.co platform. Participants were paid $.40.
Timing	The study was conducted on June 10th, 2023. Participants had max 19 minutes to answer 4 questions (most finished in less than 2 minutes).
Data exclusions	45 participants admitted to using language models to generate their responses and were excluded from the analysis
Non-participation	We only received data from the participants who completed the task and submitted their responses
Randomization	Participants were presented with a random set of 4 tasks.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☒	☐	Animals and other organisms
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging