Thinking Fast and Slow in Large Language Models:

a Review of the Decision-Making Capabilities of Generative AI Agents

Oliver Brady[1], Paul Nulty[2], Lili Zhang[3,4], Tomás E. Ward[3,4]

& David P. McGovern[1]

[1] School of Psychology, Dublin City University, Dublin 9, Ireland

[2] School of Computing and Mathematical Science, Birbeck, University of London, London, United Kingdom

[3] School of Computing, Dublin City University, Dublin 9, Ireland

[4] Insight SFI Research Centre for Data Analytics, Dublin City University, Dublin 9, Ireland

Correspondence concerning this article should be addressed to David P. McGovern, School of Psychology, Dublin City University, Glasnevin Campus, Dublin 9, Ireland. Email: david.p.mcgovern@dcu.ie

Thinking Fast and Slow in Large Language Models:

a Review of the Decision-making Capabilities of Generative AI Agents

Abstract

Large language models (LLMs) are increasingly being used in a wide range of everyday decision-making scenarios, transforming the way people make choices and interact with technology. However, despite their seemingly 'superhuman' capabilities, LLMs are not infallible and can exhibit pitfalls in their decision-making abilities if not deployed with caution. This review aims to analyse the decision-making capabilities of LLMs by comparing their abilities to humans through the lens of dual process theory. Guided by this framework, it is clear that LLMs can mimic both human-like System 1 thinking – exhibiting cognitive biases and relying on heuristics to support decision-making processes – and slower System 2 thinking through prompting methods like chain-of-thought reasoning. As LLMs have advanced, they have become more adept at comprehending tasks; however, they can still exhibit biases and make errors, some of which appear similar to human cognitive biases. What remains unclear, however, is the extent to which the processes in AI systems that lead to decision-making biases are truly analogous to those in human cognition, or if they are primarily a byproduct of the human-produced data and algorithms used to train the models. Moreover, LLMs can exhibit their own unique, nonhuman biases, such as hallucinations and overconfidence, that currently limit their application to real-world decision-making applications. Nonetheless, these models hold significant potential to revolutionise the way we make decisions across a diverse range of sectors. Thus, we conclude the review by offering recommendations for future research and practical suggestions on how to leverage LLMs to augment human decision-making.

Background

In February of 2024, Air Canada was obliged to issue a refund to a passenger in a departure from their usual policy due to their AI chatbot (Garcia, 2024). Their chatbot incorrectly told the passenger that the airline provided partial retroactive refunds for flights booked due to bereavement, a service not covered by Air Canada's policy. This led to the airline being taken to court and paying the passenger $812.02 in damages and court fees due to a mistake of their chatbot. Such instances of chatbots providing inaccurate information are not isolated occurrences; these errors represent a significant risk associated with offloading decision-making responsibilities onto large language models (LLMs). If left unchecked by human oversight, these inaccuracies have the potential to propagate misinformation and to lead to harmful outcomes (Au Yeung et al., 2023; Azaria et al., 2023; Kim et al., 2023; Meyrowitsch et al., 2023). If used correctly, however, LLMs have the potential to augment human decision-making to enhance efficiency, productivity, and innovation.

As LLMs gain popularity and wider use, their applications and capabilities continue to expand. These AI systems use deep learning techniques and are trained on vast datasets to predict and generate human-like text based on given prompts (Sartori & Orrù, 2023). With each new iteration, LLMs often demonstrate enhanced performance across a wide range of tasks, including more advanced language comprehension, improved knowledge retrieval, and greater generation capabilities. The GPT (Generative Pre-trained Transformer) series of LLMs are at the forefront of this revolution, with claims that these models possess the capability to surpass humans at certain tasks (Dillion et al., 2023; Kaddour et al., 2023). As a result, their practical applications have become increasingly widespread across a variety of sectors, with LLMs now being used in businesses, research, software development, and medicine to automate routine tasks and assist in making decisions (Griewing et al., 2024;

Jimenez et al., 2024; Jusman et al., 2023; Lee et al., 2023). For example, LLMs have the potential to assist in making high-impact strategic decisions by gathering market information, offering advice, and analysing alternative perspectives or scenarios in an effort to increase the quality of a decision to benefit a company (Basir et al., 2023; Gloria et al., 2024; Jusman et al., 2023).

In this review we focus on the kind of generative LLMs that have seen widespread adoption, investment and publicity since the introduction of ChatGPT in 2022 (OpenAI, 2022). Following the rapid advances in performance of deep neural systems in the early 2010s, a series of different architectures have at different times been at the forefront for different tasks: Convolutional Neural Networks for image recognition, recurrent architectures such as Long-Short Memory Networks (Hochreiter and Schmidhuber 1997) for sequence data with context layers, and most recently the transformer architecture (Vaswani et al 2017) for encoding sequence data with weighted attention connections among all elements of the sequence. The basic transformer design itself has been used in a variety of different architectures for solving problems in natural language processing. The transformer was first introduced as part of a system which embeds an input sequence into a contextual vector representation, and then generates a sequence from this embedding to conduct a task such as question-answering, translation, or text classification. This is known as an encoder-decoder architecture: the input sequence (e.g. a question, or a sentence to be translated) is encoded into an embedding, and the embedding is decoded to produce the output (e.g. an answer to a question, or a translated sentence). Such systems have an auto-regressive language modelling component (pre-trained by predicting the next word), but must also be fine-tuned with example input-output pairs to complete specific tasks. Radford et al (2019) showed that autoregressive language models alone can achieve good performance at multiple tasks without further fine tuning, and it is these generative, pretrained, “decoder-only” models that

we primarily have in mind in this review. However, with the recent turn towards the use of chain-of-thought tokens (Wei et al 2022), 'hidden' reasoning tokens (OpenAI, 2024), or even meaningless filler tokens used only to aid quantification (Pfau et al 2024), the line between these different classes of architecture is becoming blurred.

As the use and capabilities of LLMs continue to expand, research in this field has intensified to better understand their inner workings and limitations. One promising approach to better understand LLMs is through psychological testing (Binz & Schulz, 2023; Hagendorff, 2023; Sartori & Orrù, 2023). By treating LLMs as participants in psychological experiments, researchers can gain insight into their underlying processes and behaviours, and how they differ from humans. Cognitive testing, in particular, has been a fruitful avenue of research, where the cognitive performance of LLMs has been compared to humans to investigate the similarities and differences in cognition between these models and humans (Abbate, 2023; Binz & Schulz, 2023; Rich & Gureckis, 2019; Sartori & Orrù, 2023; Shiffrin & Mitchell, 2023; Suri et al., 2024). This comparative approach not only highlights potential shortcomings of AI cognition relative to human performance but also unveils a symbiotic relationship between cognitive science and LLM research (Qu et al., 2024). For example, insights from cognitive science can guide the development of more human-like reasoning in LLMs, enhancing their problem-solving capabilities and reducing unwanted biases. Conversely, LLMs serve as powerful tools for cognitive modelling, offering new avenues for testing hypotheses about human cognition and potentially uncovering novel perspectives on human behaviour. This bidirectional exchange promises to accelerate progress in both fields, ultimately informing where LLMs require improvement for effective deployment in real-world decision-making scenarios.

In this review, we investigate the decision-making skills of LLMs through the lens of the dual process theory of decision-making (Kahneman, 2003). We adapt insights from the field of cognitive psychology that have shed light on the limitations that can undermine human decision-making if left unchecked. By applying these insights to nonhuman LLMs, we can gain a better understanding of the strengths, limitations, and potential pitfalls of these AI systems. After introducing the intersection between decision-making and LLMs, we examine their rapid, System 1-like decision-making skills of LLMs, highlighting how they can mimic human cognitive biases and heuristics, as well as exhibit nonhuman biases such as "hallucinations". Next, we examine the slow, deliberate System 2 decision-making skills of LLMs and contemplate whether these models can attain a human-like understanding of their decision-making processes. We then discuss the importance of prompting in communicating with LLMs and how – when used appropriately – it can produce more accurate decisions in these models. Finally, we evaluate the effectiveness of LLMs as real-world decision-makers and put forth recommendations for the advancement of this rapidly evolving field, emphasising strategies to enhance the quality of future research.

Decision-making and LLMs

As LLMs become integrated into daily life and are increasingly used to support human decision-making, understanding both their capabilities and limitations is of utmost importance. In trying to understand the decision-making processes of LLMs, a useful starting point is to compare their capabilities to those of human decision-making (Binz & Schulz, 2023; Hagendorff et al., 2023; Sartori & Orrù, 2023). The field of behavioural economics has helped to identify a number of shortcomings in human decision-making (see Thaler, 2016 for a review), which can offer valuable lessons on how to identify and mitigate similar

shortcomings in the performance of AI systems. Collectively, this research has helped reveal two distinct systems that underlie human decision-making: System 1 and System 2 (Kahneman, 2003). These two systems operate in fundamentally different ways, with System 1 relying on fast, instinctive and often unconscious processes that allow us to make decisions with little or no deliberation (Evans, 2008; Kahneman & Frederick, 2002). The slower and more thought-out decisions denote System 2 thinking, which requires reflection and more effort to reach a decision (Evans, 2008; Kahneman, 2003; Kahneman & Frederick, 2002). System 1 decisions allow humans to make rapid, "quickfire" choices in our daily lives, such as deciding what food to eat or what to wear, constituting the majority of decisions we face on a regular basis (Evans, 2008; Kahneman & Frederick, 2002). In contrast, System 2 decisions tend to be rational and deductive, such as deciding what college course to attend or whether one should accept a job offer. However, while System 2 thinking is generally associated with more deliberate and analytical decision-making that is more resistant to errors (Evans, 2008; Kahneman, 2003), the relationship between the two systems is more complex; for example, decisions such as a choosing whether to accept a job offer likely involve a combination of System 1 and System 2 thinking, and mistakes can occur if the rapid, intuitive judgments of System 1 are not adequately regulated by the more deliberate reasoning of System 2 (Kahneman, 2003).

The dual process theory of decision-making to some extent reflects the longstanding debate between symbolic and connectionist approaches in AI and cognitive science (Bellini-Leite, 2022; Goel, 2022; Smolensky, 1987). Connectionist systems, including the transformer-based models that are at the heart of the most successful recent LLMs, learn patterns in large amounts of data by modifying weights between relatively uniform units. They are typically associative, fast at inference time, robust to imperfections in input, and less transparent and interpretable than symbolic systems. Although the parallel is not exact,

many of these same properties are also characteristic of System 1 thinking (Bellini-Leite, 2023; Clark, 2013). On the other hand, symbolic AI systems use built-in knowledge and rules about reasoning and inference to arrive at their answers, with the result that the processes that give rise to their answers or behaviours are more amenable to human interpretation (Goel, 2022). This is similar to how human System 2 thinking is monitored by operator and control systems such as working memory to achieve a slower and more refined form of reasoning (Bellini-Leite, 2022; Newell, 1980). Although the connectionist architecture underpinning modern LLMs is not typically associated with System 2-like reasoning abilities, recent research has shown promising results through iterative prompting approaches that reflect the dual-process decision-making system of humans, producing both quick System-1-like decisions (Hagendorff et al., 2023; Ma et al., 2023), as well as methodical System 2 thinking when prompted to do so (Hagendorff et al., 2023; Wei, Wang, et al., 2022). Some evidence suggests that LLMs can outperform humans in certain tasks with this combination of a deep neural architecture and explicit prompting for reasoning. For instance, LLMs have been shown to surpass human performance in the multiarmed bandit task which tests decision-making under uncertainty (Binz & Schulz, 2023) and they are also more adept at distinguishing between relevant and irrelevant information in reasoning tasks (Du, 2023). However, recent findings by Zhang et al. (submitted) challenge this notion, revealing that both GPT-3.5 and GPT-4 exhibit a strong tendency toward exploitation in two-armed bandit experiments, with GPT-4 displaying even greater rigidity and less adaptability to changing rewards than GPT-3.5. This suggests that LLMs may be vulnerable to manipulation in decision-making tasks, potentially due to biases in their training or decision-making processes. Moreover, these models also make errors consistent with human heuristics stemming from System 1 thinking, as well as "hallucinating" – a phenomenon where LLMs confidently generate false or fabricated information, such as citing non-existent sources or

facts – and producing other types of errors (Bellini-Leite, 2023; Hagendorff et al., 2023). Consequently, the dual process theory may offer insights into the decision-making mechanisms of both humans and LLMs, while also highlighting significant differences in their adaptability and susceptibility to manipulation.

System 1 decision-making in LLMs

Research has indicated that LLMs are capable of exhibiting human-like fast System 1 and slow System 2 decisions. Hagendorff and colleagues (2023) showed that, similar to humans, the pre-GPT-3.5 models of GPT can fall for the tricks employed by cognitive reflection tests and semantic illusions that are purposefully designed to elicit System 1 thinking. Cognitive reflection tests assess an individual's ability to resist intuitive but incorrect responses and engage in more analytical thinking. For example, the "bat and ball" problem (Frederick, 2005) illustrates this: “A bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?” The intuitive but incorrect answer is $0.10, while the correct answer, derived through careful consideration, is $0.05. Semantic illusions work similarly. One such illusion asks, “How many animals of each kind did Moses take on the Ark?” The common, incorrect System 1 response is "two", even though it was Noah, not Moses, who built the Ark (Erickson & Mattson, 1981). LLMs have also been observed to respond to cognitive reflection tests and semantic illusions in a similar intuitive but incorrect manner as humans, with this tendency becoming more pronounced with increasing model complexity until the GPT-3 models (see Figure 1; Hagendorff et al., 2023). GPT models have been shown to exhibit human-like irrational decision-making by reproducing cognitive biases and heuristics in their decisions, such as the tendency to believe a woman is more likely to be a bank teller and an active feminist, rather than just a bank teller (Binz & Schulz, 2023; Y. Chen et al., 2023; Jones & Steinhardt, 2022; Ma et al., 2023; Suri

et al., 2024). Additionally, LLMs have been shown to be vulnerable to a variety of cognitive biases including content effects, certainty effects, reflection effects, overweighting effects, framing effects, magnitude effects, confirmation bias, and recency bias, all of which are commonly associated with human decision-making (Binz & Schulz, 2023; Y. Chen et al., 2023; Jones & Steinhardt, 2022). These biases in LLMs contribute to their human-like System 1 thinking, often skewing their decisions. The precise origins of these biases remain uncertain, though some studies suggest that they may arise from the unfiltered data used to train the models (Acerbi & Stubbersfield, 2023; Azaria, 2023; Schramowski et al., 2022). This raises the possibility that LLMs inadvertently learn and perpetuate the cognitive biases and prejudices present in the human-generated content used to train them.

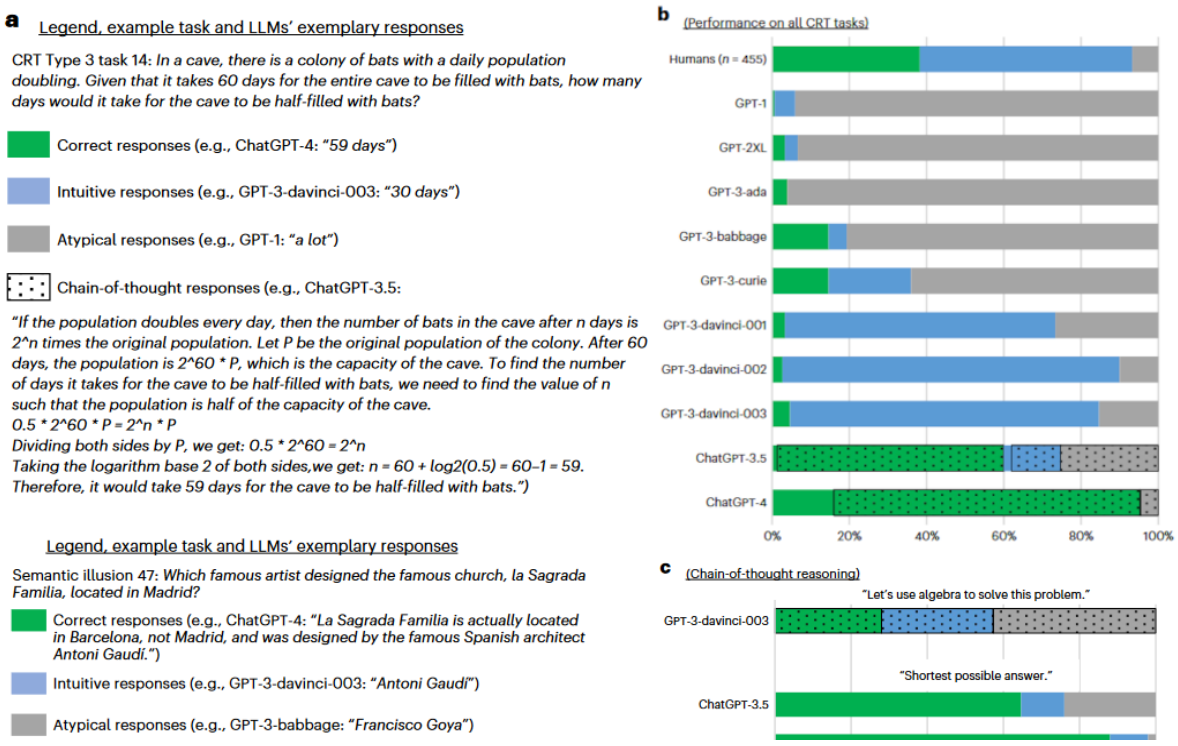

Figure 1. Performance of humans and the GPT series of LLMs on a series of cognitive reflection tasks (Hagendorff et al., 2023). a) The legend provides a sample cognitive reflection task and semantic illusion used to elicit System 1 thinking alongside potential responses. b) The results of both humans and LLMs on cognitive reflection tasks. c) The models' responses when prompted to think in different ways (System 1-like "shortest possible answer" and the System 2-like "Let's use algebra to solve this problem").

Research investigating decision-making in LLMs has shown that not only can they recreate human cognitive biases, but they also exhibit their own unique biases, such as hallucinations and overconfidence. Hallucinations involve LLMs making up a fact or source of information, resulting in an incorrect decision being made based on their 'hallucinated' source (Chakraborty et al., 2024; Dillion et al., 2023; Stella et al., 2023). Overconfidence refers to the tendency of the models to overestimate their intuition and judgments, even when answering questions incorrectly, due to their inability to extend beyond the provided prompt or express uncertainty (Borji, 2023; Y. Chen et al., 2023; Singh et al., 2023; Stella et al., 2023). These nonhuman biases in LLMs may stem from the training approaches employed, such as autoregression and reinforcement learning from human feedback (RLHF). The "base models" underlying systems such as ChatGPT are trained using autoregressive methods to predict the next word in a string by building a probabilistic model over a very large set of texts from a variety of sources. While effective at producing linguistically valid responses, in the absence of validation from external sources this method often produces factually inaccurate or fictional but plausible sounding answers (i.e. hallucinations; McCoy et al., 2023). RLHF (Ziegler et al., 2019) – a technique more commonly found in conversational models – uses human input to refine its conversational abilities. This can cause these models to employ a confident tone, which is preferred by human raters, even in cases where the model's responses are incorrect (Casper et al., 2023; Ouyang et al., 2022). Current LLMs also lack the capacity to instinctively reevaluate their decisions, rendering them unable to think critically about their responses or a given prompt. As a result, they are susceptible to producing false information or overconfident yet incorrect responses. However, when the models are explicitly trained to reflect on and evaluate their outputs, their tendency to produce hallucinated outputs decreases (L. Chen, Wang, et al., 2023; Ji et al., 2023). Research in this area is scarce, and future studies could investigate how LLMs' hallucinations

and confidence are affected by other types of reasoning, such as chain-of-thought reasoning (where LLMs are prompted to solve problems through slow, deliberate thinking akin to System 2 thinking in humans). Additionally, there is room for a review focusing solely on cognitive biases in LLMs, which would help provide a clearer perspective on which biases appear in LLMs and which ones do not.

Heuristics – another hallmark of System 1 thinking – have also been observed in LLMs. In humans, heuristics present as mental shortcuts whereby people commonly overlook certain information to make quicker decisions, such as deciding what phone to purchase based on anecdotal information rather than researching different models (Gigerenzer & Gaissmaier, 2011). Research has indicated that the GPT series of models recreates commonly observed heuristics in humans including the representativeness heuristic (Binz & Schulz, 2023; Y. Chen et al., 2023), the availability heuristic (Azaria, 2023; Suri et al., 2024), and anchoring effects (Jones & Steinhardt, 2022; Ma et al., 2023). Each heuristic provides people with a distinct shortcut when making decisions. The representativeness heuristic involves stereotyping a situation based on readily available information while disregarding the broader context (Kahneman & Frederick, 2002). An example of this is the conjunction fallacy – a cognitive bias where people erroneously believe that the probability of two specific events occurring together is higher than the probability of a single, more general event – has been observed in the responses of LLMs such as GPT-3 (Binz & Schulz, 2023; Y. Chen et al., 2023; Suri et al., 2024). The availability heuristic involves relying on similar events that easily come to mind while making decisions, leading people to believe these events are more likely than they are in reality (Tversky & Kahneman, 1973). Despite LLMs' broader access to information through their training data, they remain susceptible to biases from anecdotal information (Suri et al., 2024). Anchoring effects involve how responses can be drawn by initial values, such as thinking an item of clothing is more expensive than it is because you

first saw a similar item for a cheaper price (Ariely et al., 2003; Tversky & Kahneman, 1974). Similar to cognitive biases observed in human decision-making, LLMs' estimates of different figures such as distances or prices can be biased by prior prompts (e.g. prepending a question about the length of the Mississippi with "the length of the Mississippi is greater than 1000 miles" leads the model estimates to be shifted towards this anchor value; Jones & Steinhardt, 2022; Ma et al., 2023).

Interestingly, these heuristics observed in the decision-making processes of LLMs appear to align more closely with the intuitive System 1 heuristics of humans, rather than the rule-based heuristics typically associated with traditional AI systems (Bellini-Leite, 2023; Newell & Simon, 2007). The presence of human-like heuristics and biases in LLMs was further corroborated by a recent study by Suri and colleagues' (2024), who developed a series of novel tests to elicit various heuristics and biases in both humans and GPT-3.5. They found that the anchoring, representativeness, and availability heuristics all appeared in GPT-3.5, providing further evidence that LLMs exhibit decision-making tendencies that align with the intuitive, System 1 heuristics characteristic of human cognition. However, despite developing novel prompts that were not present in the data used to train the LLMs, the study's findings may be somewhat constrained by the relatively small sample size of human participants and the fact that the research focused primarily on GPT-3.5, with some preliminary tests being performed on GPT-4 on anchoring effects.

In contrast to the findings of Suri et al., other studies have suggested that later versions of GPT tend to exhibit a reduced tendency to display heuristics (Binz & Schulz, 2023; Y. Chen et al., 2023; Du, 2023). For example, the representativeness heuristic is observed in GPT-3 and GPT-3.5 (Binz & Schulz, 2023; Suri et al., 2024) but GPT-4 can intuitively acknowledge this bias and can correct for it (Du, 2023). Similarly, more advanced

LLMs, such as GPT-3.5 and GPT-4, have demonstrated improved ability to mitigate anchoring effects compared to earlier models like GPT-3 (Du, 2023; Jones & Steinhardt, 2022). This improved performance in later models may be the result of these models rectifying these “errors” in reasoning through the use of reinforcement learning and human feedback in their training (Suri et al., 2024). However, this observation fails to explain why Suri et al. (2024) found evidence for the availability heuristic in ChatGPT-3.5, whereas Y. Chen et al. (2023) did not. These differences may arise from variations in their testing methods and prompts. Y. Chen et al. employed Tversky and Kahneman’s (1973) bus stop task, which assesses probabilistic reasoning, whereas Suri et al. used a task involving phone model selection influenced by anecdotal information. ChatGPT-3.5's performance may vary between these tasks due to their distinct cognitive demands, potentially handling structured reasoning tasks more effectively, while being more susceptible to anecdotal biases. Additionally, the bus stop task often includes visual aids, which were not applicable to ChatGPT-3.5’s text-based format (Y. Chen et al., 2023; Tversky & Kahneman, 1973). Importantly, frequent updates to LLMs mean that even studies using the same model version (e.g., ChatGPT-3.5) may be testing slightly different iterations, potentially contributing to divergent findings. To mitigate biases in heuristic testing, researchers should employ diverse tasks not present in the models' training data and clearly specify both the model version and testing timeframe (Hagendorff, 2023; Suri et al., 2024).

In computer science, a heuristic is a method used by an AI system to evaluate the current state and determine whether a branch in the tree of possible actions can be pruned in an effort to save time (Newell & Simon, 2007; Yao et al., 2023). This evaluation heuristic might be symbolic or connectionist in implementation; the key role of a heuristic is to improve the performance of the system by reducing the cost in time or computation to select an appropriate answer or behaviour, even at the cost of some inaccuracy or bias. In the

context of human cognition, heuristics are considered to be a type of System 1 thinking – they are shortcuts in reasoning or evaluation that help people to solve problems or make decisions quickly at the cost of a higher rate of error and bias. Evidence for this alignment between AI heuristics and System 1 thinking can be seen in the tendency of pre-ChatGPT models to exhibit more pronounced System 1-like biases and errors. In contrast, the ChatGPT models, which have been specifically fine-tuned for conversing with a human agent, appear to be more resistant to these System 1-driven errors (Hagendorff et al., 2023). Despite both types of heuristics having the goal of saving time in decision-making, their mechanisms may be different.

System 2 decision-making in LLMs

While the infrastructure of LLMs inherently favours System 1 decision-making, they are still capable of exhibiting System 2 reasoning when appropriately prompted. LLMs primarily operate by predicting the next word or phrase in a sequence, which aligns with the intuitive, fast, and automatic nature of System 1 thinking (Hagendorff et al., 2023). However, LLMs can also engage in more deliberate, analytical processes akin to System 2 thinking when appropriately prompted or when faced with tasks requiring structured reasoning (Wei, Wang, et al., 2022). This System 2-like behaviour is particularly evident in problem-solving scenarios that demand step-by-step approaches or algebraic reasoning (Kojima et al., 2022). For instance, when prompted to "think step-by-step" or to break down complex problems, LLMs can demonstrate a more methodical, explicit reasoning process reminiscent of human System 2 thinking (Weston & Sukhbaatar, 2023). Language models can perform better on tasks they struggle with by using these step-by-step approaches and simulating slower human System 2 thinking, such as mathematical problems and creative writing tasks (see Figure 2;

Yao et al., 2023; Zhang et al., 2023; Zhu et al., 2023). System 2 thinking in LLMs typically requires explicit prompting but from GPT-3.5 onwards, the intuitive System 1 responses from the GPT models have become more accurate, resembling human-like System 2 problem-solving (Y. Chen et al., 2023; Hagendorff et al., 2023). This is assumed to be a result of newer models automatically engaging in more deliberate chain-of-thought reasoning without being prompted to do so, or having a more developed intuition, both of which are likely consequences of reinforcement learning and human feedback (Ouyang et al., 2022). Alternatively, this difference could be considered a departure from human-like reasoning, suggesting that the model's default reasoning abilities may have evolved beyond typical human cognitive processes. Nevertheless, the emergence of System-2-like reasoning in LLMs – whose infrastructure favours System 1 thinking – brings them a step closer to human-like decision-making.

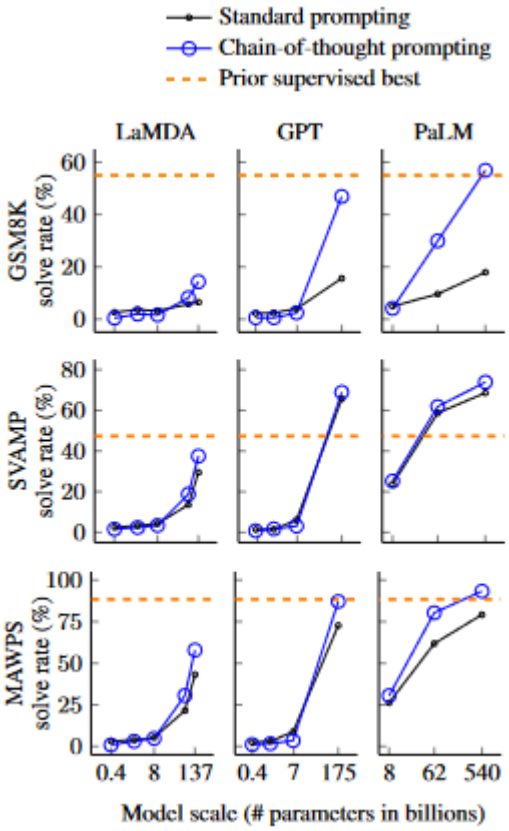

Figure 2. Comparison of the solve rates for three maths word problems tests using standard prompting and System 2-like chain-of-thought prompting across three LLM families (Wei, Wang, et al., 2022).

Prompting methods, such as chain-of-thought reasoning or tree-of-thoughts thinking, can also be used to promote human-like System 2 thinking in LLMs, enabling them to systematically break down complex problems (Bellini-Leite, 2023). Chain-of-thought reasoning is a novel form of reasoning observed in AI systems which involves solving problems when specifically prompted to think slowly and deliberately or when provided with a sample solution as guidance (Hagendorff et al., 2023; Kojima et al., 2022; Wei, Wang, et al., 2022). By prompting a LLM to think about a problem slowly or use algebra to solve it, researchers can help the models to avoid the pitfalls posed by reasoning tasks such as cognitive reflection tests and syllogisms, enabling them to provide accurate responses instead of defaulting to the instinctive System 1 answer (Hagendorff et al., 2023). Chain-of-thought

reasoning becomes increasingly effective as the number of parameters in a LLM grows, indicating a greater capacity for System 2 thinking as the model's complexity increases (Wei, Wang, et al., 2022). This reasoning process may parallel how humans use notebooks when solving problems (e.g. Zhang et al., 2023), as LLMs can use their output window to slowly reflect on a problem and enable this refined form of reasoning, similar to short-term memory (Hagendorff et al., 2023). Tree-of-thought prompting can simulate deliberate human System 2 thinking by prompting a language model to consider a range of possibilities (Yao et al., 2023). Unlike the linear progression facilitated by chain-of-thought prompting, tree-of-thought thinking simultaneously considers numerous potential thought-paths originating from a given input and resembles the branching structure of a tree when visually represented (e.g., if A, then B1, B2, or B3, followed by C1, C2, or C3 until a suitable output is reached). Heuristics are then used to evaluate these thoughts and select which one is best suited for the problem at hand, similar to control processes of human cognition (Newell & Simon, 2007). This prompting style is seen to outperform chain-of-thought prompting in a myriad of tasks, such as in creative writing tasks or when completing mini crosswords (Yao et al., 2023). However, tree-of-thought reasoning is seen as more costly in terms of effort and processing power than simpler prompting strategies, such as chain-of-thought prompting (e.g. Yao et al., 2023). The potential promise of tree-of-thought reasoning lies in its combination with reinforcement learning. DeepMind's AlphaZero system can learn to play expert-level chess and Go, given no domain knowledge other than the rules of the game. This is achieved through reinforcement learning, with the reward signal for the evaluation function ultimately deriving from the results of simulated games. Something like this approach underlies the latest best-performing model release from OpenAI, o1:

Similar to how a human may think for a long time before responding to a difficult question, o1 uses a chain of thought when attempting to solve a problem. Through reinforcement

learning, o1 learns to hone its chain of thought and refine the strategies it uses. (OpenAI, 2024)

In the case of Go or chess, the reward signal for the evaluation heuristic ultimately derives from the outcome of the simulated games. In the context of reasoning, coding, or question-answering tasks, the reward signal may be derived from automated tests, or human evaluation of responses – in the latter case effectively learning the function applied by human annotators in RLHF, and automatically applying it to each branch of the tree-of-thought. The parallel to dual process theory is then quite direct: the pretrained model generates a tree of possibilities – a comparatively slow process in which reasoning tokens are used to represent a tree of possible intermediate steps in a task – and then a heuristic evaluation function prunes this tree and selects possible paths, just as proposed in the search system of Newell and Simon (1980).

Even though LLMs are capable of System 2-like thinking, the extent to which they genuinely comprehend the judgments they make remains unclear. Since these models focus on summarising and predicting responses based on algorithms and computer heuristics, it may be that LLMs respond based on learned examples, rather than an inherent understanding of the question being asked (McCoy et al., 2019; Stella et al., 2023; Webson & Pavlick, 2022). They rely on the examples of similar prompts within their training data, struggling to respond correctly to an entirely novel problem or question (Webson & Pavlick, 2022; Zheng & Zhan, 2023). The reliance of LLMs on examples is evident in their 'few-shot learning' capability, where they can reason effectively after being trained in a given scenario or when provided with an example (Brown et al., 2020; Dasgupta et al., 2022). LLMs leverage their ability to solve a given problem by drawing upon relevant examples in their training data or literature searches, but their reasoning capability falters when confronted with unfamiliar questions, resulting in hallucinated responses (e.g., a question about a scientific discovery

that occurred after their training data cut-off; Brown et al., 2020; Chakraborty et al., 2024; Zheng & Zhan, 2023) or nonsense reasoning problems (e.g., a nonsense syllogism that poses the problem "if all zoet is spuff and all spuff are thrund, are all zoet thrund;" Dasgupta et al., 2022, p. 6). Moreover, upon prompting ChatGPT with a question based on a medical discovery made after its training dataset, Zheng and Zhan (2023) found that it creates responses without considering their logic or accuracy, leading to instances of misinformation and unwarranted confidence in its outputs. In contrast, humans can manipulate information in their working memory allowing them to fully explore a problem, whereas LLMs cannot expand beyond the context of their chatbox and are limited by their training data (Chakraborty et al., 2024; Mitchell & Krakauer, 2023; Nelson & Shiffrin, 2013; Shiffrin & Mitchell, 2023; Zheng & Zhan, 2023). For example, a language model would never truly understand the sensation of being tickled as they lack physical bodies, which may result in a narrower and potentially less accurate representation of reality relative to human embodied cognition (Mitchell & Krakauer, 2023). Hence, while a LLM may rival human benchmarks in certain comparisons, this does not necessarily imply that they are entirely human in their decision-making processes. Accordingly, their skills may not generalise to real-world decision-making scenarios, with the potential of being "right for the wrong reasons," and having no justification behind their decisions other than "parroting" what they were trained on (Bender et al., 2021; McCoy et al., 2019; Shiffrin & Mitchell, 2023). Furthermore, GPT-3's performance on semantic inference tasks was not affected by the type of instructions it was given (instructive vs misleading or irrelevant), while human performance is impaired when provided with misleading instructions (Webson & Pavlick, 2022). In sum, it appears that LLMs likely depend on heuristics for their decision-making, raising the question of whether future AI systems will achieve genuine human-like understanding and ability to respond to questions.

Explanations in favour of LLMs' ability to understand a prompt surrounds their tendency to 'zero-shot' a problem or how they relate a prompt to their training data. Solving a problem zero-shot refers to the ability of LLMs to solve problems without any previous explicit training, either when solely presented with a question ("Is this a cat?"), or when a question is followed by a prompt such as "Let's think step by step" (zero-shot reasoning; Kojima et al., 2022; Webb et al., 2023). Their ability to navigate and solve unknown problems – both with and without chain-of-thought reasoning – suggests that they do understand their judgments since they can competently respond to problems without a sample solution or recognise a problem and correctly apply it to the correct part of their training data (Kojima et al., 2022). One observed phenomenon of this type is referred to as "grokking", a term coined by the novelist Robert Heinlein in 1961. Power et al. (2021) show that when trained far beyond the point of overfitting, neural networks can progress from merely interpolating the training data in high-dimensional space to actually learning the algorithm underlying the data generating process on certain simple tasks. In their study, Power et al. focused on modular arithmetic problems, observing that the system eventually developed a geometric representation of the data that enabled full generalisation. This shift from memorisation to comprehension occurred suddenly, after an extended period of seeming stagnation in performance on the test set.

Another explanation could be that LLMs use their training data similarly to how humans use their past experiences to inform present decision-making (Seligman et al., 2013), with their training data allowing them to specialise in a certain field and potentially overcome the limitations in their current training data, such as its cut-off point (Jusman et al., 2023; X. Liu et al., 2023). Fine-tuning pretrained models on more specialised text corpora also makes it possible to provide LLMs with the ability to have domain-specific expertise and increase their understanding in a given topic (Gu et al., 2021). Additionally, even if LLMs do not

understand their prompt in a human sense, they may have a nonhuman form of understanding that emerges through statistical correlations (Mitchell & Krakauer, 2023). The pattern-matching capabilities of LLMs – which allows them to match current prompts with past prompts – could be interpreted as a form of nonhuman understanding, similar to the nonhuman cognitive biases and heuristics they inherit. To help clarify this debate on understanding, future research should investigate how the responses of LLMs are influenced by different training data and past chat history, while also being mindful of the reliance of these models on prompting.

Prompting

The importance of the prompts given to LLMs should not be understated. Prompts are how people feed information to and ask LLMs questions, and serve as the primary means for how psychological tests are administered to these models (Hagendorff, 2023). Slight variations in prompts can lead to vastly different interpretations and responses from LLMs (Binz & Schulz, 2023; Jones & Steinhardt, 2022; Kojima et al., 2022; Shiffrin & Mitchell, 2023). The phrasing and information provided in a prompt can significantly influence the performance of LLMs, enabling them to achieve a deeper level of understanding and unlocking new potential, or conversely leading to inaccurate and misleading responses. Prompting strategies, such as chain-of-thought and metacognitive prompting – a prompting strategy designed to stimulate introspective reasoning in LLMs with the goal of improving their response accuracy – can cause LLMs to reevaluate their decision and become more accurate in their responses, simulating human-like System 2 reasoning (Wang & Zhao, 2023; Wei, Wang, et al., 2022). Hence, it is critical for research in this area to use a range of paraphrased prompts to enable a comparison of responses and to ensure the robustness of

results against content biases that might influence model performance (Dasgupta et al., 2022; Hagendorff, 2023; Shiffrin & Mitchell, 2023).

While it is important to consider prompting strategies when engaging with LLMs, there is also some evidence to suggest that GPT-4 is becoming more resistant to prompting, with instances of it ignoring requests to use chain-of-thought reasoning. For example, L. Chen, Zaharia, et al. (2023) found that GPT-4's adherence to prompted instructions drifted across a period of three months, including failures in chain-of-thought reasoning, formatting compliance in code or text generation, and responding to OpinionQA items (a dataset made to identify the opinions of LLMs; Santurkar et al., 2023). This unwillingness to follow instructions resulted in a drop in accuracy in various tasks, such as solving maths problems and generating working code. This decrease in prompt-following was only observed in GPT-4, with the GPT-3.5 model becoming more accurate after three months. Upon further testing, it was observed that when GPT-4 was only given one instruction – such as to capitalise every letter in a sentence – its adherence to prompts remained consistent over time. However, when tasked with multiple instructions simultaneously – such as to capitalise every letter and add a comma after each word – GPT-4 became less accurate and more resistant to instructions over three months (L. Chen, Zaharia, et al., 2023). Chen et al. suggest that these shifts in responding could be unintended side effects caused by minor updates regularly made to each model aimed at improving their capabilities but have negative consequences to other behaviours. Hence, these issues may be resolved with a future minor update to the AI system. However, since OpenAI does not disclose the complete changelog of their updates publicly, the source of these behaviour shifts remains uncertain as it stands. More longitudinal research should be performed on LLMs to determine how their adherence to instructions changes over time, as well as investigating how cognitive load affects prompt adherence.

To date, there has been no research comparing decision-making performance in LLMs to humans when prompted in the same way. For example, prompts designed to provoke chain-of-thought reasoning have been exclusively applied to AI agents, without parallel testing in human participants (Hagendorff et al., 2023; Kojima et al., 2022; Wei, Wang, et al., 2022). Similarly, Wang and Zhao's (2023) metacognitive prompting was only applied to LLMs, without the inclusion of a human sample for comparison, with the model's performance instead compared to a metacognition framework derived from previous studies in human participants. Studies surrounding prompting strategies tend to be from a computer science perspective, where the goal of these studies is based on computer behaviour, not how it relates to that of humans. When viewed through a psychological lens, these methods of prompting LLMs may be considered 'unusual' if applied to humans; for example, asking a person to solve the lily pad and the pond cognitive reflection test using algebra when the trick lies in the wording of the question (Hagendorff et al., 2023). Moreover, the overreliance of LLMs on their prompts could pose a challenge for research in this domain, as they lack the ability to contextualise a given prompt within the broader context of the world in a similar vein to humans (Shiffrin & Mitchell, 2023; Stella et al., 2023; Webson & Pavlick, 2022). Thus, a recommendation for the field is that study authors should document and disclose the exact wording for the prompts and any prompt variations used in their studies, while being mindful of the phrasing used to ensure that the model responses are fully contextualised (Hagendorff, 2023).

Could LLMs replace or augment human decision-makers?

As the abilities of LLMs improve, it is important to consider how they can assist with human decision-making. LLMs surpass humans in decision-making across various domains

and particularly excel in intuitive System 1 decisions (Hagendorff et al., 2023). Through reinforcement learning and advanced algorithms, LLMs are capable of reaching "superhuman" levels of proficiency, allowing them to outperform human System 2 thinking in activities such as chess (Mitchell & Krakauer, 2023; Silver et al., 2017). As a result of their superior capabilities, LLMs are often viewed as ideal candidates to aid in decision-making scenarios; however, the promise of these models in assisting with decision-making should be tempered by the concerns raised in the previous sections of the review and it is crucial to approach their use with caution. Nonetheless, LLMs have achieved some notable successes in a range of real-world decision-making applications, such as in the medical domain. Although not specifically trained for medical purposes, LLMs hold significant potential to enhance medical decision-making due to their training based on openly available medical databases (Lee et al., 2023; Thirunavukarasu et al., 2023). As such, GPT-4 has demonstrated impressive performance on sample questions from the United States Medical Licensing Examination and has shown promising results in medical question-answering tasks (Nori et al., 2023). In business, LLMs can help summarise information, generate reports, and come up with plans based on real-time data to help people make informed decisions regarding management, finance, and other corporate domains (Chuma & de Oliveira, 2023; Jusman et al., 2023). LLMs can also be used as a research tool, aiding in tasks such as literature reviews and hypothesis generation (Aydın & Karaarslan, 2022; Demszky et al., 2023; Ke et al., 2024). These applications point to the significant promise that LLMs hold, with their potential to improve decision-making practices across a variety of fields if used correctly.

Despite LLMs' apparent competency in making decisions, the flaws of these models should be reiterated. Information provided by AI chatbots may be susceptible to inaccuracies and raises concerns of potential plagiarism from existing sources, further complicating the reliability and authenticity of the content (Aydın & Karaarslan, 2022; Meyrowitsch et al.,

2023). Moreover, while LLMs show promise in augmenting human decision-making, the capabilities of these models do not equate to that of an expert in that field. For example, Chuma and de Oliveira (2023) identified that the input provided by ChatGPT on business decisions was not close to an expert level, providing simple and "generic" overviews and explanations for the situations it was presented with. Even if they can outperform medical benchmarks, these skills may not translate into real-world medical scenarios and LLMs are still prone to errors and biases, including gender and racial prejudices (An et al., 2024; Au Yeung et al., 2023; Borji, 2023; Liang et al., 2021; Nori et al., 2023). The adoption of LLMs and other AI technologies also presents the risk of degrading job quality and raising unemployment rates, as they can automate jobs previously carried out by human workers (Eloundou et al., 2023; Weidinger et al., 2021). Thus, at present, attempts to integrate LLMs into the workplace should focus on how they can be utilised to work in tandem with people to arrive at conclusions as opposed to replacing humans in decision-making tasks (Jusman et al., 2023; S. Liu et al., 2023; Nori et al., 2023). Humans can consider a multitude of different factors in making decisions, such as their long-term consequences (Jusman et al., 2023; Shiffrin & Mitchell, 2023) while LLMs are over-reliant on their prompted context and cannot expand beyond their training data (Jusman et al., 2023; S. Liu et al., 2023; Shiffrin & Mitchell, 2023; Weidinger et al., 2021; Zheng & Zhan, 2023). Instead of replacing humans, language models can help inform human decision-making as a support system, being supervised by experts to provide information and suggestions to help people reach conclusions (Au Yeung et al., 2023; Azaria et al., 2023; J. Chen, Liu, et al., 2023; S. Liu et al., 2023). This thought process also aligns with OpenAI's recommended uses of ChatGPT, being used as a chatbot to assist with a range of tasks such as making informed decisions through conversation (OpenAI, 2022). This leads to the conclusion that ChatGPT and other chatbots should aid in human decision-making under expert supervision or rigorous fact-

checking, rather than autonomously making decisions on behalf of humans (Azaria et al., 2023; Jusman et al., 2023; Nori et al., 2023). However, the limitations of these models still need to be accounted for when using them as decision-making assistants. Risks such as misinformation, hallucinations, and data security all pose problems while using LLMs, even with supervision (Weidinger et al., 2021). To mitigate these risks, it will be important to develop effective strategies to highlight the potential pitfalls of LLMs, along with measures to detect misinformation and prevent overreliance on these models (Heersmink, 2024; Jusman et al., 2023; Nori et al., 2023).

The ever-changing nature of LLMs

With each update, LLMs naturally evolve to become more accurate and intelligent, even developing new cognitive processes that were not present in earlier iterations (Hagendorff et al., 2023; Kosinski, 2023; Webb et al., 2023; Wei, Tay, et al., 2022). In Hagendorff et al.'s (2023) investigation of the System 1 and 2 decision-making abilities across various iterations of GPT, the models demonstrated significant improvements in their System 1 decision-making accuracy with each successive release of GPT. Notably, in this study ChatGPT-3.5 and ChatGPT-4 were able to intuitively engage in chain-of-thought reasoning and answer correctly without being prompted to do so, something which older models such as GPT-3 were not able to do. Furthermore, when explicitly asked to answer as quickly as possible, ChatGPT-3.5 and ChatGPT-4 maintained higher accuracy compared to their predecessors, suggesting a more developed intuition (Hagendorff et al., 2023). While these findings are compelling for the GPT family of models, it is important to consider the generalisability of these results to other LLMs and contexts. For example, there have also been reports of GPT-4 being more rigid, conservative in its decision-making relative to GPT-

3 on a two-armed bandit task (Zhang et al., submitted). However, the broader trend in LLM development is in keeping with the findings of Hagendorff et al., with each new state-of-the-art LLM tending to outperform the benchmarks set by previous models. For instance, the newer Claude 3 Opus model reportedly surpasses GPT-4's abilities across all tested domains, such as general reasoning, maths skills, and common knowledge (Anthropic, 2024). It is also noteworthy how differences in responses have been observed within the same model version of GPT to the same prompts at different times due to the minor updates made to each model (see Figure 3; L. Chen, Zaharia, et al., 2023). Thus, it is important to assess the various iterations of LLMs not only against human performance but also against their own cognitive abilities – both in the present and over time – as well as against other families of LLMs such as comparing Claude and GPT models (for examples of these comparisons, see L. Chen, Zaharia, et al. [2023], Dasgupta et al. [2022], and Hagendorff et al. [2023]). Moreover, it is necessary for research in this field to include the timeframe within which assessments were conducted on LLMs within their methodologies in acknowledgement of the constant refinements and adjustments made to these models over time.

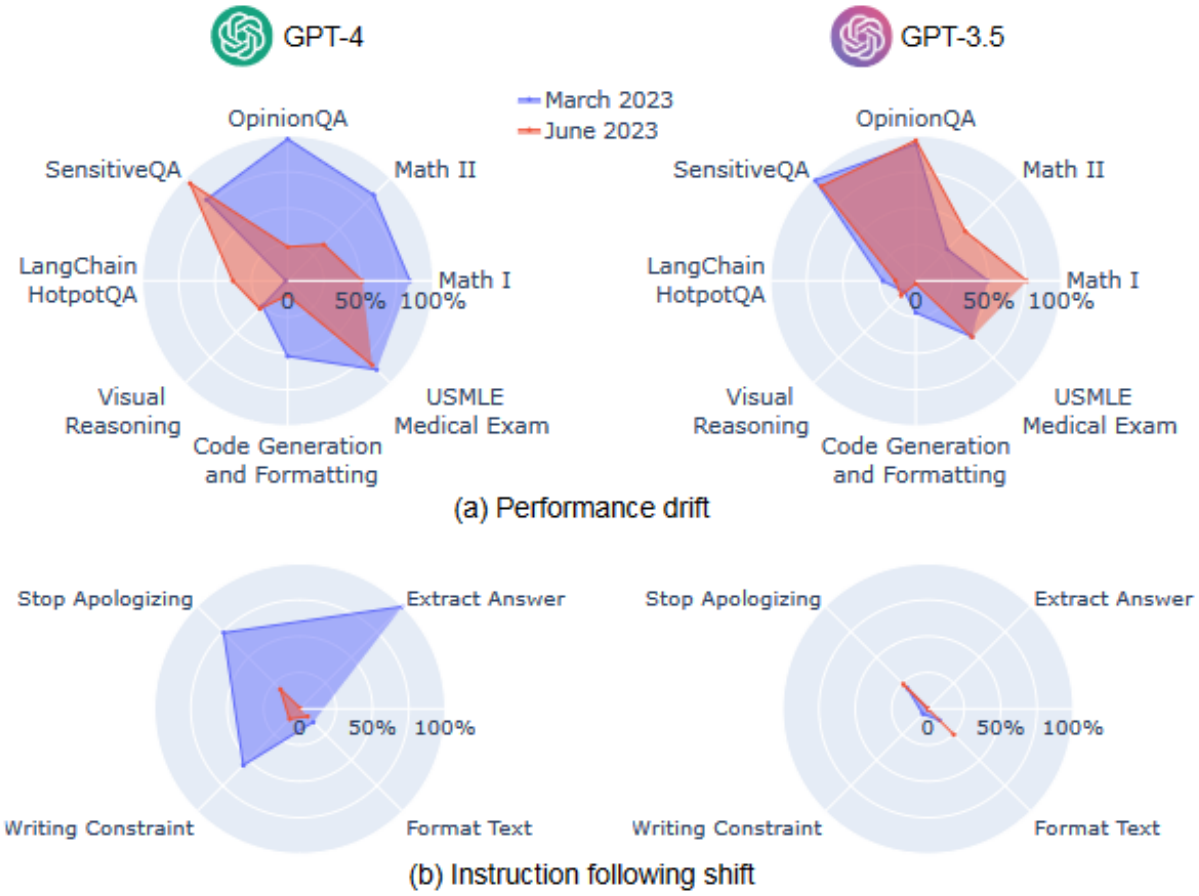

Figure 3. The change in (a) performance and (b) instruction-following of GPT-4 and GPT-3.5 from March to June of 2023 (L. Chen, Zaharia, et al., 2023).

Integrating agent-based approaches with LLMs presents a promising pathway to enhance decision-making capabilities in AI systems. Agent-based systems, which are designed to autonomously perceive, reason, and act within an environment, can provide a structured framework for guiding LLMs through more complex decision-making processes. This approach helps overcome LLMs' limitations in making goal-oriented, context-sensitive decisions (Luo et al., 2023). By embedding LLMs within agent architectures, models can be augmented with features such as continuous learning, situational awareness, and adaptability. These agents can simulate different environments, test out scenarios, and interact with other agents or systems, allowing for better-informed, contextually-grounded decisions (Zhang et

al., 2022). For example, in clinical settings or real-time customer service applications, agent-based approaches can continuously monitor evolving data, adjust the LLM's decision-making strategies, and refine its output to meet specific, time-sensitive goals.This synergy enhances the practicality of LLMs in real-world decision-making tasks by ensuring that the model's responses align with broader, agent-driven objectives and constraints. Recent research has demonstrated the potential of this integration in various domains, including robotics, financial forecasting, and personalised education (Chen et al., 2024).

Future research

Since 2022, LLMs have been investigated as a participant in psychological tests to uncover their cognitive capabilities (Hagendorff, 2023). As new models are released and older models are updated, their abilities have steadily improved with each iteration. They have shown promising skills in a variety of cognitive domains, being able to match or outperform humans in some areas (Binz & Schulz, 2023; Sartori & Orrù, 2023). Despite the progress made in understanding the decision-making capabilities of these models, this research is in its early stages, with many arguing that the internal processes of LLMs are akin to a “black box” from a cognitive perspective (Ke et al., 2024). Much remains unknown about the decision-making skills of LLMs and current research in the field can provide methodological recommendations for future work.

One such recommendation is that researchers conducting studies on LLMs should incorporate a wide range of prompt variations and psychological tests in order to provide a comprehensive understanding of these models. Previous studies have indicated that slight alterations in prompt wording can produce markedly different responses due to a specific prompt or test being present in the model’s training data, or biases caused by the wording of

prompts (Binz & Schulz, 2023; Hagendorff, 2023; Loya et al., 2023; Shiffrin & Mitchell, 2023). Using diverse prompts can help account for this sensitivity to prompt phrasing and enable research to become more robust and resistant to these confounding variables on a variety of tests. Furthermore, incorporating diverse testing scenarios allows readers to identify specific contexts where behaviours are exhibited or absent (e.g. the availability heuristic; Y. Chen et al., 2023; Suri et al., 2024). However, when creating prompt variations for research, it is important to validate them first, ensuring that they accurately target the desired variable (Hagendorff, 2023). Additionally, the exact wording of the prompts used in studies should be provided by the authors to make readers aware of any potential flaws in prompt design and help with future replications.

One limitation of current research investigating the decision-making abilities of LLMs is that comparisons between these models and human participants are at a surface level. Studies that investigate the performance of language models on benchmarks tend not to compare their results with humans, and when they do, they tend to be underpowered with small sample sizes and forgo reporting statistics such as standard deviations (Sartori & Orrù, 2023). Future studies that aim to compare LLMs to humans should address these limitations while also extending beyond merely using humans as benchmarks for accuracy. For instance, comparisons should explore human and AI decision-making across a variety of everyday tasks where accuracy is not the only factor. Just because LLMs can perform to a benchmark, does not mean they can equate to human performance in the real world (McCoy et al., 2019; Nori et al., 2023); simply comparing outcomes to human performance on a given test does not provide full insight into the factors at play contributing to a decision, such as how context-dependent LLMs are or their understanding of a question and response (Shiffrin & Mitchell, 2023). Probing LLMs for justifications about their decisions could provide insight into these decisions. Furthermore, when comparing the performance of language models, it is

important to broaden the scope beyond the GPT family to include other LLMs such as the Mixtral (Mistral AI, 2023) or the Claude (Anthropic, 2024) families of models, as well as open source and transparent models such as the Hugging Face ecosystem (Hussain et al., 2023).

Proper guidelines for using LLMs in assisting with decision-making need to be established. Due to their ever-increasing uses and applications, it is recommended that future research investigate methods and regulations for using LLMs to help avoid misuse (Nori et al., 2023; Weidinger et al., 2021). Developing methods and guidelines to use LLMs in different settings, such as Azaria and colleagues' (2023) flowcharts for academic writing with AI, can help counter risks such as misinformation and data security, and enable an easier application of LLMs in these areas. To enhance the quality and consistency of research that involves cognitive testing in LLMs, it will be important to establish standardised guidelines for administering psychological tests to AI agents. These guidelines should cover the appropriate parameters and settings to use when testing these models, which will serve to reduce the variability in the results that these investigations produce (Azaria et al., 2023; Hagendorff, 2023).

While the focus of this review has been on the decision-making capabilities of LLMs, it is also worth considering how the use of these models will impact human decision-making. If used correctly, LLMs can be invaluable assets in enhancing decision-making processes (Martínez et al., 2010; Perlis et al., 2024) and extending human decision-making capabilities by acting as a cognitive artefact or external decision-making agent (Qu et al., 2024; Rasmequan & Russ, 2000). LLMs may indeed enhance and extend our decision-making processes by taking on part of the cognitive load, thereby enabling us to focus on more complex and strategic aspects of decision-making. However, it is also important to consider

potential indirect negative consequences on human cognition that may arise from excessively outsourcing cognitive tasks to LLMs. This concern is not unfounded; for instance, Treiman et al. (2024) recently found that humans modify their responses when informed that their decisions will be used to train AIs, with this effect persisting even after they are told their answers will no longer be used for AI training. Such findings reflect a broader historical pattern of apprehension regarding the impact of new technologies on cognitive abilities. Indeed, these concerns can be traced back to ancient Greek philosophers, with Socrates and Plato believing that the advent of writing would negatively affect human memory (Frentz, 2006). This scepticism has persisted through technological advancements, with claims that modern technologies have significantly influenced cognition; for instance, widespread use of smartphones and the internet are believed to have altered memory processes and attention spans (Carr, 2020; Tanil & Yong, 2020), while GPS systems have impacted navigation and spatial reasoning skills (Clemenson et al., 2021; Dahmani & Bohbot, 2020). In this context, there are growing concerns that overreliance on LLMs for decision-making could potentially diminish critical thinking and independent problem-solving abilities, leading to a form of cognitive offloading where complex reasoning is increasingly delegated to AI systems. However, it is also important to consider potential positive effects. LLMs might enhance critical thinking through feedback and supportive learning, serving as cognitive scaffolding that helps users develop more sophisticated thinking strategies (Bai et al., 2023). Furthermore, interaction with AI chatbots has been shown to reduce belief in conspiracy theories, suggesting a potential role in promoting critical evaluation of information (Costello et al., 2024). Given the uncertain long-term impact of LLMs on human decision-making and cognitive processes, Heersmink (2024) emphasises the need for longitudinal studies to inform policies and safeguards. These studies should aim to understand how sustained interaction

with AI systems affects various cognitive domains, from problem-solving and creativity to memory and attention.

Conclusion

LLMs hold huge potential to assist or augment human decision-making; however, they also possess the capacity to introduce biases, errors and misinformation if their output is not properly screened or monitored. Similar to humans, these models are subject to cognitive biases and heuristics previously identified in the fields of cognitive psychology and behavioural economics, necessitating that their responses are treated with caution. Additionally, LLMs can exhibit model-specific biases and heuristics posing a new set of problems for industries seeking to use AI systems to replace or augment human decision-making.

Methods of prompting, such as those that promote chain-of-thought reasoning, and agentic AI approaches can allow LLMs to become more accurate in their decision-making, simulating System 2-like thought processes. LLMs can be human-like in their decision-making, making the same human System 1 errors to a prompted question, as well as being able to reflect on a question and think methodically about it, akin to System 2 thinking. However, there are some limitations to these abilities; for instance, their overreliance on prompting, their tendency to generate misinformation, and a general brittleness in their responses that casts doubt on the level of understanding that they possess. Thus, in their current state, LLMs cannot substitute for human decision-makers in applied settings, but they do have the potential to assist in human decision-making – offering suggestions and summarising research to allow people to make an informed decision in a given situation – if appropriate guardrails are in place. Even if the System 1 reasoning of LLMs becomes more

accurate compared to humans, they can still unconsciously show cognitive biases or hallucinate responses, and explicit methods to elicit System 2 thinking should be employed to mitigate the risks of intuitive System 1 thinking.

This review provides a summary of the current research in the field and indicates where it could be improved upon by considering the lessons we have learned in the study of human decision-making. However, despite mimicking human decision-making skills, their underlying decision-making processes may be distinct from those of humans. Regardless of how 'human' the decision-making of a LLM is, principles from the cognitive psychology of decision-making can still be applied to these AIs to guide our understanding of their overall decision-making skills (Gronchi & Perini, 2024; Rich & Gureckis, 2019). Even if the inner workings of LLM decision-making remain a black box, cognitive psychology can help inspire and identify limits of their capabilities and indicate where they need to be improved upon. As LLMs become increasingly integrated into our daily lives, understanding their cognitive processes through the lens of human psychology will be crucial for designing reliable and ethically-aligned AI systems that can effectively complement human decision-making in a range of real-world scenarios.

Acknowledgements

The authors would like to thank Min Wu for feedback and comments on the review.

References

Abbate, F. (2023). Natural and artificial intelligence: A comparative analysis of cognitive aspects. *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science*. https://doi.org/10.1007/s11023-023-09646-w

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *120*(44), 1–5. https://doi.org/10.1073/pnas.2313790120

An, J., Huang, D., Lin, C., & Tai, M. (2024). *Measuring Gender and Racial Biases in Large Language Models* (arXiv:2403.15281). arXiv. http://arxiv.org/abs/2403.15281

Anthropic. (2024, March 4). *Introducing the next generation of Claude*. https://www.anthropic.com/news/claude-3-family

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106. https://doi.org/10.1162/00335530360535153

Au Yeung, J., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., Dobson, R. J., & Teo, J. T. (2023). AI chatbots not yet ready for clinical use. *Frontiers in Digital Health*, *5*. https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2023.1161098

Aydın, Ö., & Karaarslan, E. (2022). *OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare* (SSRN Scholarly Paper 4308687). https://doi.org/10.2139/ssrn.4308687

Azaria, A. (2023). *ChatGPT: More Human-Like Than Computer-Like, but Not Necessarily in a Good Way*.

Azaria, A., Azoulay, R., & Reches, S. (2023). *ChatGPT is a Remarkable Tool—For Experts*

(arXiv:2306.03102). arXiv. https://doi.org/10.48550/arXiv.2306.03102

Bai, L., Liu, X., & Su, J. (2023). ChatGPT: The cognitive effects on learning and memory. *Brain-X*, *1*(3), e30. https://doi.org/10.1002/brx2.30

Basir, A., Puspitasari, E. D., Aristarini, C. C., Sulastri, P. D., & Ausat, A. M. A. (2023). Ethical Use of ChatGPT in the Context of Leadership and Strategic Decisions. Jurnal Minfo Polgan, 12(1), Article 1. https://doi.org/10.33395/jmp.v12i1.12693

Bellini-Leite, S. C. (2022). Dual Process Theory: Embodied and Predictive; Symbolic and Classical. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.805386

Bellini-Leite, S. C. (2023). Dual Process Theory for Large Language Models: An overview of using Psychology to address hallucination and reliability issues. *Adaptive Behavior*, 10597123231206604. https://doi.org/10.1177/10597123231206604

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Borji, A. (2023). *A Categorical Archive of ChatGPT Failures* (arXiv:2302.03494). arXiv. https://doi.org/10.48550/arXiv.2302.03494

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Carr, N. (2020). *The shallows: What the Internet is doing to our brains*. WW Norton & Company.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., … Hadfield-Menell, D. (2023). *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback* (arXiv:2307.15217). arXiv. http://arxiv.org/abs/2307.15217

Chakraborty, N., Ornik, M., & Driggs-Campbell, K. (2024). *Hallucination Detection in Foundation Models for Decision-Making: A Flexible Definition and Review of the State of the Art* (arXiv:2403.16527). arXiv. http://arxiv.org/abs/2403.16527

Chen, X., Li, Y., & Brown, A. (2024). Applications of integrated LLM-agent systems in diverse domains. *IEEE Transactions on Artificial Intelligence*, 5(2), 312-329.

Chen, J., Liu, L., Ruan, S., Li, M., & Yin, C. (2023). Are Different Versions of ChatGPT's Ability Comparable to the Clinical Diagnosis Presented in Case Reports? A Descriptive Study. *JOURNAL OF MULTIDISCIPLINARY HEALTHCARE*, *16*, 3825–3831. https://doi.org/10.2147/JMDH.S441790

Chen, L., Wang, L., Dong, H., Du, Y., Yan, J., Yang, F., Li, S., Zhao, P., Qin, S., Rajmohan, S., Lin, Q., & Zhang, D. (2023). *Introspective Tips: Large Language Model for In-Context Decision Making* (arXiv:2305.11598). arXiv. http://arxiv.org/abs/2305.11598

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* (arXiv:2307.09009). arXiv. https://doi.org/10.48550/arXiv.2307.09009

Chen, Y., Andiappan, M., Jenkin, T., & Ovchinnikov, A. (2023). *A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?* (SSRN Scholarly Paper 4380365). https://doi.org/10.2139/ssrn.4380365

Chuma, E. L., & de Oliveira, G. G. (2023). Generative AI for Business Decision-Making: A

Case of ChatGPT. *Management Science and Business Decisions*, *3*(1), 5–11. https://doi.org/10.52812/msbd.63

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clemenson, G. D., Maselli, A., Fiannaca, A. J., Miller, A., & Gonzalez-Franco, M. (2021). Rethinking GPS navigation: Creating cognitive maps through auditory clues. *Scientific Reports*, *11*(1), 7764. https://doi.org/10.1038/s41598-021-87148-4

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Dahmani, L., & Bohbot, V. D. (2020). Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Scientific Reports*, *10*(1), 6310. https://doi.org/10.1038/s41598-020-62877-0

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). *Language models show human-like content effects on reasoning tasks*. https://doi.org/10.48550/ARXIV.2207.07051

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701. https://doi.org/10.1038/s44159-023-00241-5

Dezfouli, A., Nock, R., & Dayan, P. (2020). Adversarial vulnerabilities of human decision-making. *Proceedings of the National Academy of Sciences*, *117*(46), 29221–29228. https://doi.org/10.1073/pnas.2016921117

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Du, M. (2023). Machine vs. human, who makes a better judgment on innovation? Take GPT-4 for example. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1206516

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models* (arXiv:2303.10130). arXiv. https://doi.org/10.48550/arXiv.2303.10130

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540–551. https://doi.org/10.1016/S0022-5371(81)90165-1

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Frentz, T. S. (2006). Memory, Myth, and Rhetoric in Plato's Phaedrus. *Rhetoric Society Quarterly*, *36*(3), 243–262. https://doi.org/10.1080/02773940500511546

Gallagher, S., & Crisafi, A. (2009). Mental Institutions. *Topoi*, *28*(1), 45–51. https://doi.org/10.1007/s11245-008-9045-0

Garcia, M. (2024, February 19). *What Air Canada Lost In 'Remarkable' Lying AI Chatbot Case*. Forbes. https://www.forbes.com/sites/marisagarcia/2024/02/19/what-air-canada-lost-in-remarkable-lying-ai-chatbot-case/

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of*

Psychology, *62*, 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gloria, B., Melsbach, J., Bienert, S., & Schoder, D. (2024). Real-GPT: Efficiently Tailoring LLMs for Informed Decision-Making in the Real Estate Industry. *Journal of Real Estate Portfolio Management*, 1-17.

Goel, A. (2022). Looking Back, Looking Ahead: Symbolic versus Connectionist AI. *AI Magazine*, *42*(4), Article 4. https://doi.org/10.1609/aaai.12026

Griewing, S., Knitza, J., Boekhoff, J., Hillen, C., Lechner, F., Wagner, U., ... & Kuhn, S. (2024). Evolution of publicly available large language models for complex decision-making in breast cancer care. Archives of Gynecology and Obstetrics, 1-14.

Gronchi, G., & Perini, A. (2024). Dual-process theories of thought as potential architectures for developing neuro-symbolic AI models. *Frontiers in Cognition*, *3*. https://doi.org/10.3389/fcogn.2024.1356941

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 2:1-2:23. https://doi.org/10.1145/3458754

Hagendorff, T. (2023). *Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods* (arXiv:2303.13988). arXiv. https://doi.org/10.48550/arXiv.2303.13988

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*(10), Article 10. https://doi.org/10.1038/s43588-023-00527-x

Heersmink, R. (2024). Use of large language models might affect our cognitive skills. *Nature Human Behaviour*, 1–2. https://doi.org/10.1038/s41562-024-01859-y

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation MIT-Press*.

Hussain, Z., Binz, M., Mata, R., & Wulff, D. U. (2023). *A tutorial on open-source large language models for behavioral science*. OSF. https://doi.org/10.31234/osf.io/f7stn

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., & Fung, P. (2023, December 1). *Towards Mitigating LLM Hallucination via Self Reflection*. The 2023 Conference on Empirical Methods in Natural Language Processing. https://openreview.net/forum?id=up8EYzyrKV

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). *SWE-bench: Can Language Models Resolve Real-World GitHub Issues?* (arXiv:2310.06770). arXiv. https://doi.org/10.48550/arXiv.2310.06770

Jones, E., & Steinhardt, J. (2022). Capturing Failures of Large Language Models via Human Cognitive Biases. *Advances in Neural Information Processing Systems*, *35*, 11785–11799.

Jusman, I. A., Ausat, A. M. A., & Sumarna, A. (2023). Application of ChatGPT in Business Management and Strategic Decision Making. *Jurnal Minfo Polgan*, *12*(2), Article 2. https://doi.org/10.33395/jmp.v12i2.12956

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). *Challenges and Applications of Large Language Models* (arXiv:2307.10169). arXiv. http://arxiv.org/abs/2307.10169

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, *93*(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press.

https://doi.org/10.1017/CBO9780511808098.004

Ke, L., Tong, S., Cheng, P., & Peng, K. (2024). *Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review* (arXiv:2401.01519). arXiv. https://doi.org/10.48550/arXiv.2401.01519

Kim, J. H., Kim, J., Park, J., Kim, C., Jhang, J., & King, B. (2023). When ChatGPT Gives Incorrect Answers: The Impact of Inaccurate Information by Generative AI on Tourism Decision-Making. *Journal of Travel Research*, 00472875231212996. https://doi.org/10.1177/00472875231212996

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Kosinski, M. (2023). *Theory of Mind Might Have Spontaneously Emerged in Large Language Models* (arXiv:2302.02083). arXiv. https://doi.org/10.48550/arXiv.2302.02083

Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*, *388*(13), 1233–1239. https://doi.org/10.1056/NEJMsr2214184

Liang, P. P., Wu, C., Morency, L.-P., & Salakhutdinov, R. (2021). Towards Understanding and Mitigating Social Biases in Language Models. *Proceedings of the 38th International Conference on Machine Learning*, 6565–6576. https://proceedings.mlr.press/v139/liang21a.html

Liu, S., Wright, A. P., Patterson, B. L., Wanderer, J. P., Turer, R. W., Nelson, S. D., McCoy, A. B., Sittig, D. F., & Wright, A. (2023). Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association*, *30*(7), 1237–1245. https://doi.org/10.1093/jamia/ocad072

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2023). GPT understands, too. *AI Open*. https://doi.org/10.1016/j.aiopen.2023.08.012

Loya, M., Sinha, D. A., & Futrell, R. (2023). Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variation and Hyperparameters. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3711–3716. https://doi.org/10.18653/v1/2023.findings-emnlp.241

Luo, Y., Zhang, S., & Chen, H. (2023). Enhancing LLM decision-making through agent-based frameworks. *Journal of Artificial Intelligence Research*, 78, 1245-1278.

Ma, D., Zhang, T., & Saunders, M. (2023). *Is ChatGPT Humanly Irrational?* https://doi.org/10.21203/rs.3.rs-3220513/v1

Martínez, L., Ruan, D., & Herrera, F. (2010). Computing with Words in Decision support Systems: An overview on Models and Applications. *International Journal of Computational Intelligence Systems*, *3*(4), 382–395. https://doi.org/10.2991/ijcis.2010.3.4.2

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference* (arXiv:1902.01007). arXiv. https://doi.org/10.48550/arXiv.1902.01007

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). *Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve* (arXiv:2309.13638). arXiv. https://doi.org/10.48550/arXiv.2309.13638

Meyrowitsch, D. W., Jensen, A. K., Sørensen, J. B., & Varga, T. V. (2023). AI chatbots and (mis)information in public health: Impact on vulnerable communities. *Frontiers in Public Health*, *11*. https://doi.org/10.3389/fpubh.2023.1226776

Mistral AI. (2023, December 11). *Mixtral of experts*. https://mistral.ai/news/mixtral-of-

experts/

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*(2), 356–394. https://doi.org/10.1037/a0032020

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*(2), 135–183. https://doi.org/10.1016/S0364-0213(80)80015-2

Newell, A., & Simon, H. A. (2007). Computer science as empirical inquiry: Symbols and search. In *ACM Turing Award Lectures* (p. 1975). Association for Computing Machinery. https://doi-org.dcu.idm.oclc.org/10.1145/1283920.1283930

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). *Capabilities of GPT-4 on Medical Challenge Problems* (arXiv:2303.13375). arXiv. https://doi.org/10.48550/arXiv.2303.13375

OpenAI. (2022). *Introducing ChatGPT*. https://openai.com/blog/chatgpt

OpenAI. (2024, September 12). *Learning to Reason with LLMs.* Open AI Blog, https://openai.com/index/learning-to-reason-with-llms/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Perlis, R. H., Goldberg, J. F., Ostacher, M. J., & Schneck, C. D. (2024). Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology*, 1–5. https://doi.org/10.1038/s41386-024-01841-2

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2021). GROKKING: Generalization beyond overfitting on small algorithmic datasets. In Proceedings of the 1st Mathematical Reasoning in General Artificial Intelligence Workshop, ICLR 2021. OpenAI.

Qu, Y., Du, P., Che, W., Wei, C., Zhang, C., Ouyang, W., ... & Liu, Q. (2024). Promoting interactions between cognitive science and large language models. *The Innovation*, 5(2).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical report*, *1*(8), 9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rasmequan, S., & Russ, S. (2000). Cognitive artefacts for decision support. *Smc 2000 Conference Proceedings. 2000 Ieee International Conference on Systems, Man and Cybernetics. 'cybernetics Evolving to Systems, Humans, Organizations, and Their Complex Interactions' (Cat. No.0*, *1*, 651–656 vol.1. https://doi.org/10.1109/ICSMC.2000.885069

Rich, A. S., & Gureckis, T. M. (2019). Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence*, *1*(4), Article 4. https://doi.org/10.1038/s42256-019-0038-z

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). *Whose Opinions Do Language Models Reflect?* (arXiv:2303.17548). arXiv. https://doi.org/10.48550/arXiv.2303.17548

Sartori, G., & Orrù, G. (2023). Language models and psychological sciences. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1279317

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large

pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, *4*(3), Article 3. https://doi.org/10.1038/s42256-022-00458-8

Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. (2013). Navigating Into the Future or Driven by the Past. *Perspectives on Psychological Science*, *8*(2), 119–141. https://doi.org/10.1177/1745691612474317

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, *120*(10), e2300963120. https://doi.org/10.1073/pnas.2300963120

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2017). *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm* (arXiv:1712.01815). arXiv. https://doi.org/10.48550/arXiv.1712.01815

Singh, A. K., Devkota, S., Lamichhane, B., Dhakal, U., & Dhakal, C. (2023). *The Confidence-Competence Gap in Large Language Models: A Cognitive Study* (arXiv:2309.16145). arXiv. https://doi.org/10.48550/arXiv.2309.16145

Smolensky, P. (1987). Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review*, *1*(2), 95–109. https://doi.org/10.1007/BF00130011

Stella, M., Hills, T. T., & Kenett, Y. N. (2023). Using cognitive psychology to understand GPT-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences*, *120*(43), e2312911120. https://doi.org/10.1073/pnas.2312911120

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001547

Tanil, C. T., & Yong, M. H. (2020). Mobile phones: The effect of its presence on learning and memory. *PLOS ONE*, *15*(8), e0219233. https://doi.org/10.1371/journal.pone.0219233

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Treiman, L. S., Ho, C.-J., & Kool, W. (2024). The consequences of AI training on human decision-making. *Proceedings of the National Academy of Sciences*, *121*(33), e2408731121. https://doi.org/10.1073/pnas.2408731121

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Wang, Y., & Zhao, Y. (2023). *Metacognitive Prompting Improves Understanding in Large Language Models* (arXiv:2308.05342). arXiv. https://doi.org/10.48550/arXiv.2308.05342

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

Webson, A., & Pavlick, E. (2022). *Do Prompt-Based Models Really Understand the Meaning of their Prompts?* (arXiv:2109.01247). arXiv.

https://doi.org/10.48550/arXiv.2109.01247

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models* (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language Models* (arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems*, *36*, 11809–11822.

Zhang, H., Huang, J., Li, Z., Naik, M., & Xing, E. (2023). *Improved Logical Reasoning of Language Models via Differentiable Symbolic Programming* (arXiv:2305.03742). arXiv. https://doi.org/10.48550/arXiv.2305.03742

Zhang, L., Wang, J., & Liu, T. (2022). Agent-augmented language models for complex reasoning tasks. *Proceedings of the Conference on Neural Information Processing Systems*, 35, 8790-8802.

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., & Smola, A. (2023). *Multimodal chain-of-thought reasoning in language models.* (arXiv:2302.00923). arXiv

Zheng, H., & Zhan, H. (2023). ChatGPT in Scientific Writing: A Cautionary Tale. *The*

American Journal of Medicine, *136*(8), 725-726.e6. https://doi.org/10.1016/j.amjmed.2023.02.011

Zhu, X., Wang, J., Zhang, L., Zhang, Y., Gan, R., Zhang, J., & Yang, Y. (2023). Solving Math Word Problems via Cooperative Reasoning induced Language Models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4471–4485. https://doi.org/10.18653/v1/2023.acl-long.245

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv Preprint arXiv:1909.08593*.