

Dual-process theory and decision-making in large language models

Oliver Brady¹, Paul Nulty², Lili Zhang^{3,4}, Tomás E. Ward^{3,4}

& David P. McGovern¹

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

² School of Computing and Mathematical Science, Birkbeck, 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

Dual-process theory and decision-making in large language models

Abstract

Large language models (LLMs) are increasingly embedded in everyday decision-making scenarios, altering how people make choices. Despite the seemingly ‘superhuman’ capabilities of LLMs in some domains, there are pitfalls in the decision-making performance of LLMs and they should therefore be used with caution. In this Review, we examine LLM outputs through the lens of dual-process theory and against the backdrop of human decision-making. We detail how in decision-making scenarios, LLMs mimic both System-1-like responses – exhibiting cognitive biases and employing heuristics – and System-2-like responses – slow and carefully reasoned – through specific prompting methods. However, LLM reasoning is not fully analogous to human dual-process cognition. For instance, the ‘cognitive’ biases observed in LLMs often reflect patterns in their training data and LLMs exhibit specific non-human biases, such as hallucinations, that constrain their use in real-world decision-making. Despite these limitations, LLMs have the potential to augment human decision-making when deployed responsibly. Thus, we conclude with recommendations for mitigating biases and improving reliability to enable the deployment of LLMs as effective decision-support systems.

Keywords: Decision-making, Large Language models, Dual process theory, System 1, System 2

Introduction

In November 2023, a class-action lawsuit was filed against UnitedHealth Group, alleging that it used an artificial intelligence (AI)-driven system to systematically deny health insurance claims for elderly patients. Worryingly, the system was said to override physicians' recommendations by rejecting claims even when doctors deemed continued care to be necessary. These decisions allegedly led to premature patient discharges, potentially putting vulnerable elderly patients at risk of receiving inadequate medical treatment or facing the financial burden of covering their care independently¹. Similarly, private health insurers have deployed AI-driven decision-making systems to automate claims processing, allowing large language models (LLMs) to determine the validity of treatment requests². In some cases, these systems might have denied legitimate claims without adequate human oversight, prioritising cost-saving efficiencies over medical necessity. Together, these cases raise critical concerns about the role of AI in decision-making and suggest that without human oversight, AI use has the potential to lead to harmful outcomes³⁻⁶. However, if used correctly, LLMs have the potential to augment human decision-making and to enhance efficiency, productivity and innovation.

LLMs are gaining popularity and their applications and capabilities continue to expand. These AI systems are trained on vast datasets to predict and generate human-like text in response to prompts – textual inputs that provide instructions to guide the model's output⁷. With each new release, LLMs often demonstrate enhanced performance over their predecessors across a wide range of tasks, including more advanced language comprehension, improved knowledge retrieval and greater generation capabilities. The generative pre-trained transformer series of LLMs are at the forefront of these advances, with claims that these models possess the capability to surpass humans at certain tasks^{8,9} (Box 1). Generative pre-trained transformer (GPT) is both the generic name of an LLM architecture

and the name given by the company OpenAI to specific models of this type (such as GPT-3). Other well-known LLM models such as Gemini, Llama, Claude, DeepSeek and PaLM also use this generative pre-trained transformer architecture. The practical applications of these models have become increasingly widespread across a variety of sectors. LLMs are now used in businesses, research, software development and medicine to automate routine tasks and assist in making decisions^{10–13}. For example, LLMs have the potential to support high-impact strategic decision-making by gathering market information, offering advice and analysing alternative perspectives or scenarios, thereby enhancing the quality of decisions and driving positive outcomes^{12,14,15}.

As the use and capabilities of LLMs continue to expand, research aimed at a better understanding of their inner workings and limitations has intensified. One promising approach to understanding LLMs better is through psychological testing^{7,16–18}, in which prompts serve as the primary means of administering these tests. By treating LLMs as participants in psychological experiments, researchers can gain insight into their underlying processes and behaviours, and how they differ from humans. In particular, cognitive testing has been a fruitful avenue of research, used to compare LLM performance with human behaviour and to investigate their similarities and differences^{7,16,20–23}. This comparative approach not only highlights potential shortcomings in current LLM performance relative to human cognition but also unveils a symbiotic relationship between cognitive science and LLM research²⁴. 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 and indicate where LLMs require improvement for effective deployment in real-world decision-making scenarios.

One way to better understand the strengths and weaknesses of LLMs in decision-making is to apply insights from behavioural economics, a field that has identified several shortcomings in human decision-making (see²⁵). Among its key contributions is dual-process theory, which proposes that two distinct systems underlie human decision-making, referred to as System 1 and System 2²⁶. System 1 refers to fast, automatic and intuitive processes that enable rapid responses with minimal cognitive effort, whereas System 2 involves slower, more deliberate and effortful reasoning and is often employed for complex or unfamiliar problems. These labels do not necessarily refer to discrete neural systems, but rather to qualitatively different modes of processing that differ in speed, control and resource demands²⁷. Everyday decisions, such as choosing what to eat, are typically handled by System 1 processes, which are efficient and often accurate in familiar environments. More complex decision-making, such as selecting a university course, typically recruits System 2 to provide a more reflective analysis. Thus, System 2 thinking is generally associated with more deliberate, analytical decision-making and metacognitive self-monitoring that can help to reduce errors^{26,28}. However, the relationship between the two systems is nuanced. For example, many real-world decisions – such as deciding whether to accept a job offer – often involve an interplay of intuitive (System 1) and deliberative (System 2) processes. In such cases, suboptimal outcomes can arise when rapid System-1-based judgements are not sufficiently complemented by slower, more reflective System 2 reasoning^{26,29}. In other instances, fast and intuitive System 1 processes alone can lead to accurate and efficient judgements, particularly in familiar environments with stable patterns and clear feedback^{30,31}. Conversely, System 2 is not inherently rational; it can introduce its own biases through overthinking or motivated reasoning²⁷.

The dual-process theory of decision-making to some extent reflects the longstanding debate between symbolic and connectionist approaches in AI and cognitive science^{32–34}. On the one hand, connectionist systems – including the transformer-based models that are at the heart of the most successful LLMs of the past few years – learn patterns in large amounts of data by modifying the connection weights between relatively uniform units. These systems are typically associative, fast when required to generate responses at inference time (test time), robust to imperfections in the input and not particularly transparent to humans. Although the parallel is not exact, many of the same properties are also characteristic of System 1 thinking^{35,36}. 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 interpretation³³. These properties are 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 compared to the single-pass process of purely connectionist systems^{34,37}. Expert systems that combine statistical modules with symbolic rules or learned rulesets, such as random forests, are examples of this type of architecture³⁸. Although the core connectionist architecture that underpins modern LLMs is typically associated with System-1-like reasoning, the outputs of modern LLMs cannot be considered simply System-1-like. Indeed, research in the past few years has detailed LLM outputs that seem to reflect the dual-process decision-making system of humans, including both quick System-1-like decisions^{39,40} and slower System 2 decisions (when the model is prompted to do so)^{39,41}. Consequently, the dual-process theory might offer insights into the decision-making mechanisms of both humans and LLMs while also highlighting substantial differences in their adaptability and susceptibility to manipulation.

In this Review, we examine the decision-making skills of LLMs through the lens of the dual-process theory of decision-making²⁶. We adapt insights from the field of cognitive psychology that have shed light on the limitations that can undermine human decision-making. By applying these insights to non-human LLMs, we can gain a better understanding of the strengths, limitations and potential pitfalls of these AI systems. We first examine the rapid, System-1-like decision-making outputs of LLMs and highlight how they can mimic certain human cognitive biases, while also exhibiting non-human biases such as ‘hallucinations’. Next, we examine the slow, deliberate System 2 decision-making outputs of LLMs, the importance of prompting and how certain types of prompt can produce more accurate outputs. We next contemplate whether these models have the capacity to make decisions in a similar way to humans, the origins of their biases and the question of what an LLM can ‘understand’. Finally, we assess the strengths and limitations of using LLMs to assist with real-world decision-making and offer recommendations for the advancement of this rapidly evolving field. Although we draw comparisons throughout between human decision-making and LLM decision-making, these comparisons are intended to guide analysis and not to assert a one-to-one correspondence. In this sense, the dual-process framework is used as an organising structure to interpret outputs from LLMs and should not be interpreted as evidence that they share underlying cognitive architectures or conscious reasoning patterns with humans (see Box 2 for a discussion of the dangers of anthropomorphising LLMs). We consider the remit of ‘decision-making’ to include a wide range of tasks that involve making a choice. Although not all reasoning tasks discussed in the cognitive science literature (particularly mathematical reasoning) directly constitute decision problems, they illustrate the reasoning processes of LLMs in ways that can plausibly be extended to cases in which LLMs need to perform mathematical reasoning to inform a choice (and subsequently possibly an action). In keeping with many decision-making tests involving

humans, we also include scenarios in which models are asked to make a choice in a hypothetical scenario without a real action taking place.

System 1-like decision-making in LLMs

System-1-like responses are typically fast, intuitive and automatic, often relying on heuristics or mental shortcuts. Although heuristics can often be useful, enabling fast and efficient decisions⁴², they also increase the likelihood of systematic cognitive biases⁴³. Cognitive reflection tests assess an individual’s ability to resist these intuitive but incorrect responses and engage in more analytical thinking. For example, the bat-and-ball problem⁴⁴ illustrates this pattern: “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, whereas the correct answer 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 human response is ‘two’, even though it was Noah, not Moses, who is known for building the Ark⁴⁵. These tests are intentionally designed to elicit intuitive answers based on System 1 heuristics and correct answers require these initial responses to be overridden through more deliberate System 2 processing. In essence, they measure the tendency to resist a prepotent but incorrect answer and to engage in future reflection that produces the correct solution⁴⁶. However, performance on such tasks is not strictly diagnostic of System 1 thinking, as success can also depend on factors such as numeracy, inhibitory control and reading comprehension^{44,46,47}. Generative pre-trained transformer models before GPT-3.5 produce responses that align with the intuitive but incorrect answers that humans give on cognitive reflection tests and for semantic illusions, becoming increasingly human-like as the model complexity increases³⁹ (Fig. 1). The ChatGPT models, which have been specifically fine-tuned for conversing with humans, seem to be more resistant to these System-1-type errors

than other often-used LLMs³⁹. However, modern LLMs such as the GPT-4 and Llama models continue to reproduce several well-documented cognitive biases that are characteristic of System-1-like responding, such as the conjunction fallacy^{16,23,40,48–51}. They exhibit framing effects⁵², certainty effects, confirmation bias⁵³ and recency bias, all of which are usually associated with System 1 processing and can reliably bias the content and direction of the responses they generate^{16,23,48–50,54,55} (Table 1). The precise origins of these biases remain uncertain, although some studies suggest that they might arise from the unfiltered data used to train the models^{56–58}; because such data reflect the heuristics and biases of human reasoning, LLMs are likely to learn and reproduce these patterns in their outputs. This observation raises the possibility that LLMs inadvertently learn and perpetuate the cognitive biases and prejudices that are present in the human-generated content that is used to train them⁵⁹.

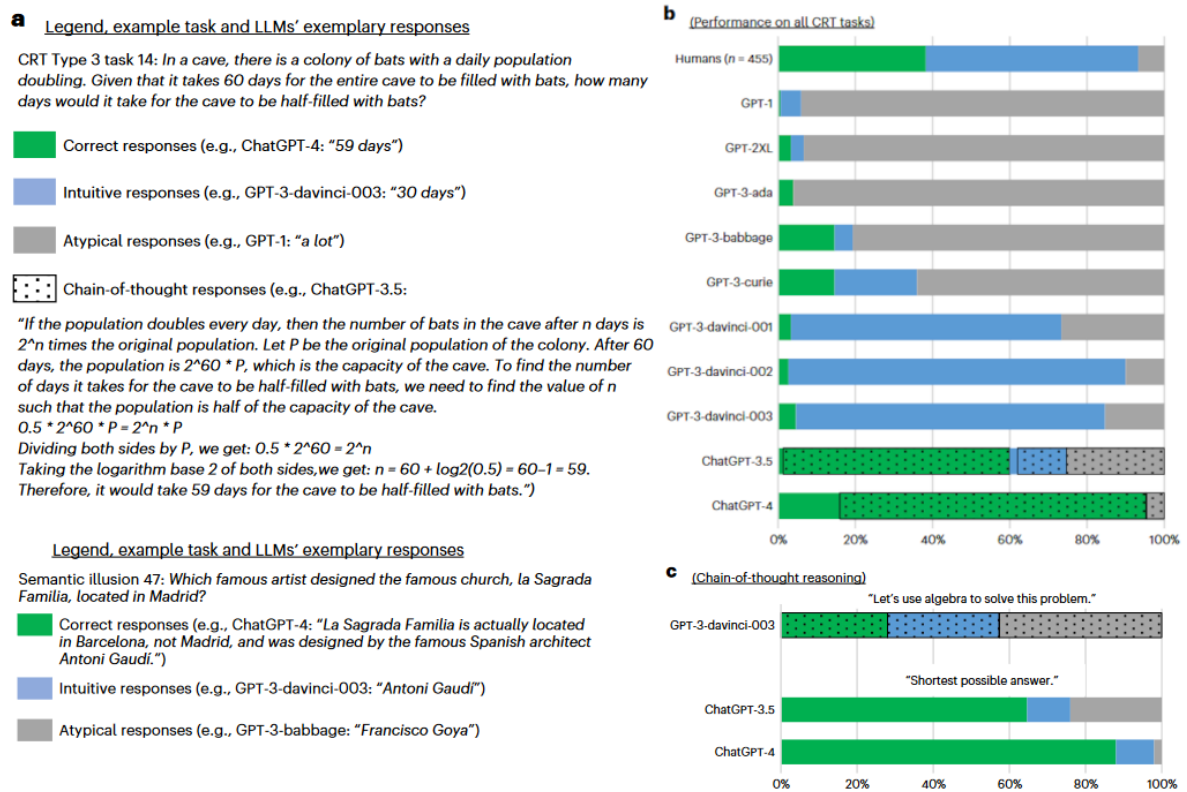


Figure 1. Performance of humans and the GPT series of LLMs on a series of cognitive reflection tasks³⁷. a) The legend provides a sample cognitive reflection task and semantic illusion used to elicit System 1 thinking alongside potential responses. b) The results of both humans and LLMs on cognitive reflection 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").

Cognitive Bias	Definition	Models Exhibiting the Bias	Differences/similarities to humans	Origins
Representativeness heuristic	Stereotyping when deciding if an object belongs to a class.	GPT-3 ¹⁶ GPT-3.5 ^{23,48,50,55,60} GPT-4 ^{48,50,55,60} Gemini/Bard ^{51,55} Llama 2-70B ⁶⁰ PaLM 2 ⁶⁰	Similar to humans, LLMs conform to social stereotyping in regards to the representativeness heuristic. Evidence is mixed when comparing accuracy between humans and LLMs in various representativeness heuristic tasks; Pantana et al. ⁵⁰ suggest that ChatGPT is less accurate and falls for the conjunction fallacy more than human responders, while Wang et al. ⁶⁰ found that GPT-3.5 and GPT-4 are more accurate than humans, with Llama 2 being the most comparable to human performance	Embedded social biases in the models' training data ⁵⁵
Availability heuristic	Responses being biased by similar events that immediately come to mind.	GPT-3.5 ^{23,50} ChatGPT ⁵⁶ GPT-4 ⁵⁰	Evidence that both GPT-3.5 and GPT-4 act more rationally and showcase less bias than humans ⁴⁸ GPT-3.5 was more often biased by anecdotal information compared to humans ^{23,50} , whereas GPT-4 exceeds human performance	Bias present in training data
Anchoring effects	How responses can be drawn or biased by initial values.	GPT-3 ⁴⁹ ChatGPT-3.5 ^{23,54} GPT-4 ⁵⁴ GPT-4o ⁵⁴ Claude 2.0 ⁶¹ Gemini Pro ⁶¹	GPT-4 and some other LLMs are less influenced by the anchor than humans ⁵⁴ The stronger the model, the less variance in responses when presented with an anchor ⁵⁴ LLMs, such as Gemini/Bard and GPTs can ignore the anchor, but still provided an inaccurate answer ⁵⁵	Bias present in the training data Algorithms used to train LLMs can exacerbate these biases Human reinforcement
Overconfidence	The tendency to overestimate one's ability to complete a task.	GPT-3.5 ^{48,95} GPT-4 ^{48,89} GPT-4o ⁹⁵ GPT-o1 ⁹⁵ Claude-Instant ⁸⁹ Llama ⁸⁹ Llama 3.2 3B ⁹⁵	LLM overconfidence is higher than that of human participants ⁹⁵ LLM overconfidence can parallel the Dunning-Kruger effect, in which people	Reinforcement learning from human feedback favouring a confident tone in LLMs ⁹⁶ Biased training data ⁹⁵

		Llama 3.1 8B ⁹⁵	overestimate their ability in an area despite their lack of competence, or training in the case of LLMs ⁸⁹	
Confirmation bias	The tendency to be favourable towards evidence that confirms a theory, dismissing any contradictory evidence.	GPT-3.5 ⁴⁸ GPT-4 ⁴⁸ Gemini ¹⁰¹ Claude 3.5 ¹⁰¹	Similar to humans, LLMs can overlook information that disconfirms their current guess to problems ¹⁰¹ GPT-3.5 resembles to human decision-making, while GPT-4 generated responses that consistently produced confirmation bias	Can fall victim to a confirmation bias when trying to convince the user its output is correct ⁵ Can generate specific information that aligns with the LLM's previous answer, residing in the same search space as the previous answer ¹⁰¹
Primacy/Recency bias	The preference for items towards the beginning or end of a list respectively	ChatGPT ^{151,152} GPT-4 ⁸¹ Llama ¹⁵¹ FLOR ¹⁵³ BLOOM ¹⁵³ Mistral ^{151,153}	Similar to humans, LLMs also display a preference towards different items depending on how they are ordered in a list	Biased training data Consequence of reinforcement learning and human feedback ¹⁵²
Framing effects	The preference for one choice over another based on how a decision is phrased.	GPT-3 ^{16,49} GPT-3.5 ^{23,50,55} GPT-4 ^{50,55} ChatGPT ⁵⁶ Bard/Gemini ⁵⁵	Pantana et al. ⁵⁰ found that the GPT-3.5 and GPT-4 did not give any correct answers, while humans responded correctly 66% of the time GPT-3.5 and GPT-4 acted rationally in Chen et al.'s ⁴⁸ study compared to humans	Potential origins of the framing effect in linguistic patterns ²³ Limited information being available causes the LLM to show framing effects more ⁵⁰
Certainty effects	A preference for choices that have a 100% probability of occurring rather than choices with a risk attached to them.	GPT-3 ¹⁶ GPT-4 ⁴⁸	GPT-4 differs from humans as it consistently chooses certainty regardless of the scenario's framing or expected payout However, GPT-4 is similar to humans in its preference for known results over uncertainty ⁴⁸	Could arise from the risk-averse nature seen in some LLMs ⁴⁸

Table 1. Table of common cognitive biases and heuristics present in LLMs, outlining their similarities or differences to that of human biases and their potential LLM origins.

Another hallmark of System 1 thinking is the use of heuristics. In humans, heuristics are mental shortcuts whereby people overlook certain information to make quicker decisions, such as deciding what phone to purchase based on anecdotal information rather than

researching different models⁴². Heuristics are usually understood to be a feature of System 1 thinking – they facilitate fast decision-making by reducing cognitive load, although sometimes at the expense of accuracy²⁶. LLMs such as Gemini, Llama and the GPT series of models recreate behaviour that is often observed in humans and attributed to heuristics and related biases, including the representativeness heuristic^{16,48,50,51,55,60}, the availability heuristic^{23,50,56} and anchoring effects^{23,40,49,54,61} (Table 1). 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 context²⁹. A well known manifestation of the representativeness heuristic is the conjunction fallacy⁶² – people erroneously believe that the probability of two specific events occurring together is higher than the probability of a single, more general event, which has been observed in the responses of LLMs such as GPT-3 (refs. ^{16,23,48}), GPT-4 (refs. ^{48,55}) and Gemini^{51,55}. The availability heuristic involves relying on similar events that easily come to mind while making decisions, which leads people to believe these events are more likely than they are⁶³. Although LLMs have broader access to information than humans do, through their training data, they remain susceptible to biases from anecdotal information²³. Anchoring effects refer to how responses can be affected by previous experience, such as thinking an item of clothing is more expensive than it is because you first saw a similar item for a cheaper price^{43,64}. Similar to this effect in humans, LLMs’ estimates of figures such as distances or prices can be biased by prior prompts. For instance, prepending a question about the length of the Mississippi with “the length of the Mississippi is greater than 1000 miles” leads the model’s estimates of the river’s length to be shifted towards this anchor value and away from the true value (~2,350 miles)^{23,40,49}.

In contrast to the use of the term in psychology, 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^{65,66}. 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 needed to select an appropriate answer or behaviour (even at the cost of some inaccuracy or bias). Interestingly, the heuristics observed in the decision-making processes of LLMs appear to align more closely with the intuitive System 1 heuristics of humans, rather than with the rule-based heuristics typically associated with traditional AI systems^{35,65}. This pattern is notable because the heuristics of traditional AI are explicit, engineered rules designed to reduce computation time, whereas the shortcuts observed in LLMs are learned from human language and manifest as human-like biases.

The presence of LLM outputs that mimic human-like heuristics and biases was further corroborated by a series of tests to elicit biased responses in both humans and GPT-3.5 (ref. ²³). To ensure a fair comparison between humans and GPT-3.5, testing parameters were carefully matched, a fresh session was initiated for each trial and the tests were specifically created for this study, which helped to equalise task exposure between humans and LLMs and prevented contextual carry-over between trials. The authors found that the anchoring, representativeness and availability heuristics all appeared in the outputs of GPT-3.5, which provides further evidence that LLMs exhibit decision-making tendencies that align with the intuitive, System 1 heuristics characteristic of human cognition. However, the study's findings might be somewhat constrained by the relatively small sample size and the fact that the research focused primarily on GPT-3.5, which limits the generalisability of the results given the rapid advances in LLM design with different architectures and capabilities. By contrast, other studies have suggested a trend whereby more recent iterations of GPT exhibit a reduced tendency to produce outputs that seem heuristic-driven^{16,48,67}. For example, more advanced LLMs, such as GPT-4 and GPT-4o, have an improved ability to mitigate anchoring

effects compared to earlier models like GPT-3 and GPT-3.5 (refs. ^{49,54,67}). Similarly, although the representativeness heuristic is observed in Llama 2-70B and PaLM 2, more complex models such as GPT-4 are more likely to identify when prompts are structured to trigger this heuristic and to adjust their responses accordingly⁶⁰. Although the complex GPT models with higher parameters outperformed both Llama 2 and PaLM, GPT-3.5 outperformed GPT-4 because GPT-4 requested more information before answering, which was marked as an incorrect answer. However, model complexity was not a variable of interest in this study and differences in model complexity should be considered further when investigating cognitive biases in LLMs⁶⁰. The improved performance in later models might be due in part to their training incorporating reinforcement learning from human feedback⁶⁸, which has been used in commercial model releases since early 2022 (ref. ⁶⁹). This training process involves human evaluators who assess model outputs and guide the model towards more desirable responses through reward-based fine-tuning. Unlike traditional reinforcement learning, which relies on clear, environment-defined feedback, reinforcement learning from human feedback operates in open-ended, linguistic domains where optimal responses might be subjective or ambiguous. However, this observation fails to account for discrepancies between studies that found²³ and did not find⁴⁸ evidence for the availability heuristic in GPT-3.5; these discrepancies might arise from variations in testing methods, prompts and the use of tasks with distinct cognitive demands and structure.

Although the emerging literature on biases and heuristics in LLMs provides intriguing insights, the field is nascent and rapidly evolving, so these results should be interpreted with caution. Many existing studies differ substantially in their experimental protocols – in prompt phrasing, model versioning and whether sessions were restarted between trials – making direct comparisons difficult and occasionally yielding inconsistent results⁸. Importantly, frequent updates to LLMs mean that even studies that use the same model version (such as

GPT-3.5) might be testing slightly different iterations. Additionally, some classic decision-making vignettes could have appeared in the model's training data, which further muddies the interpretation. To mitigate confounds in heuristic testing, researchers should employ diverse tasks that are not present in the models' training data and clearly specify both the model version and testing timeframe, and conduct replication studies for confirmation that these LLMs truly display these cognitive biases and heuristics^{18,23}. As the field matures, the development of clearer reporting standards and replication efforts will be crucial to generating more robust generalisable insights.

Although human heuristics are often discussed in the context of cognitive biases and inaccurate responses, they are a hallmark of intelligent adaptive behaviour. In everyday life, fast, intuitive decision strategies enable humans to make effective decisions with limited time, effort or computational resources. A substantial body of research has shown that humans' use of simple heuristics can be highly effective in real-world scenarios, especially in environments in which available cues are both informative and well matched to the heuristic's selection criteria^{30,42}. These strategies are computationally efficient and less metabolically costly than deliberative processes and therefore confer evolutionary advantages in routine decision-making contexts²⁷. This insight has also begun to inform LLM design. For instance, some approaches have incorporated architectures that are inspired by dual-process theory and separate fast, automatic responses from slower, more deliberative reasoning. These systems typically use a default mechanism to generate quick, low-cost outputs and invoke more reflective processing only when prompted by uncertainty or task complexity^{70,71}. Thus, in both biological and artificial systems, the efficiency of heuristic-driven reasoning represents a feature rather than a flaw, providing a mechanism for achieving satisfactory performance in the face of resource constraints.

System 2-like decision-making in LLMs

Although the infrastructure of LLMs inherently favours System 1 decision-making, they are still capable of engaging in more deliberative, analytical processes akin to System 2 thinking when appropriately prompted or when faced with tasks that require structured reasoning⁴¹. This System-2-like behaviour is particularly evident in problem-solving scenarios that demand step-by-step approaches or algebraic reasoning⁷². For instance, when prompted to “think step by step” or to break down complex problems, LLMs can produce outputs that resemble methodical, step-by-step reasoning processes⁷³. For example, by generating intermediate calculations in mathematical problems or outlining plot steps in creative writing, LLMs show improved performance when using step-by-step approaches that simulate slower human System 2 thinking^{66,74,75}. System-2-like outputs in LLMs typically require explicit prompting, but from GPT-3.5 onwards, the intuitive unprompted responses of these models have become more accurate and less biased, producing outcomes that resemble those typically associated with human-like System 2 problem-solving^{39,48}. This improvement in response quality – namely, fewer intuitive but biased answers – is thought to be the result of newer models automatically producing more structured and well-developed answers without explicit prompting (which might be a result of the reinforcement learning from human feedback used for these models)⁶⁹. This shift towards more structured responding resembles the automatization principle, in which humans’ complex reasoning patterns become faster and more intuitive with practice, eventually resembling System-1-type responses^{27,76–78}. Although not directly comparable to human expertise, this improvement in default outputs might parallel the way that experts in various domains rely on rapid, reliable intuitions that emerge after extensive experience. Alternatively, this improvement could be considered a departure from human-like reasoning and an indication that the model’s default

reasoning abilities might diverge from typical human cognitive processes in ways that are not yet fully understood⁷⁹. At the same time, the capacity for System-2-like reasoning in LLMs – whose infrastructure favours System 1 thinking – highlights a convergence in outcomes with human decision-making, even if the underlying processes differ.

The phrasing and information provided in a prompt can enable an LLM to generate more sophisticated and contextually appropriate outputs, thereby unlocking new potential for improved performance across a wider range of tasks or conversely leading to inaccurate and misleading responses. Slight variations in prompts can lead to vastly different interpretations and responses^{16,22,49,72}, analogous to the way in which humans are influenced by the way instructions are framed⁵². This susceptibility to framing is evident in performance on, for example, the previously mentioned cognitive reflection tests^{29,44}. However, LLM responses can be more sensitive to prompt phrasing than humans are to instructions; for example, certain wordings can cause a model to be unresponsive to the prompt⁸⁰, whereas slight variations can enable them to elicit a coherent response^{23,39,72,81,82}. Specific prompting strategies for eliciting reasoning can also be used to promote human-like System 2 thinking in LLMs, enabling them to systematically break down complex problems to achieve higher accuracy³⁵. Chain-of-thought reasoning refers to AI models' ability to generate explicit intermediate steps during problem solving, which can be elicited in two main ways^{41,72}. The first is through prompting, in which the LLM is explicitly instructed to “think step by step” or is shown sample reasoning. The second is through the system's own chain-of-thought reasoning outputs, which are the tokens (the basic units of text the model processes, such as words, parts of words or punctuation) it produces that show its reasoning process while solving the problem. These outputs seem to show a record of ‘thinking aloud’, thereby adding transparency to the model's process; however, these reasoning traces do not always faithfully represent the actual process by which the model arrives at its answer⁸³. By guiding LLMs

through a problem step by step with chain-of-thought prompting or by explicitly requesting intermediate output during the stages of problem-solving, researchers can encourage more structured outputs on tasks such as cognitive reflection tests and syllogisms, reducing reliance on shallow heuristics and increasing the likelihood of accurate responses³⁹. Chain-of-thought reasoning becomes increasingly effective as the number of parameters in an LLM grows, suggesting that larger models are better able to simulate the structured, deliberative problem-solving associated with System 2 processing, for example by eliciting specific forms of intermediate structure, such as algebraic expressions⁴¹. This reasoning process might parallel how humans use notebooks when solving problems^{39,84}, as LLMs can use their output window to iteratively process information, supporting more refined reasoning by decomposing problems into steps with the aid of external memory.

Unlike the linear progression facilitated by chain-of-thought prompting, the tree-of-thought approach simultaneously considers numerous potential thought paths originating from a given input. Doing so can be achieved purely through prompting, by instructing the model to propose several different answers or approaches to a problem, or guided algorithmically by sampling proposed answers from the model and maintaining a data structure in the form of a tree where each node is an answer or next step proposed by the model. When visually represented, these paths resemble the branching structure of a tree (for instance, if A, then B1, B2 or B3, followed by C1, C2 or C3 until a suitable output is reached). Tree-of-thought prompting can simulate human System 2 thinking by prompting a language model to consider a range of possibilities⁵⁹. This tree-of-thought approach mirrors early cognitive theories of human problem-solving⁶⁵ and aligns with dual-process models of cognition, in which intuitive responses (System 1) are subject to metacognitive evaluation (System 2), which might take the form of verifiable checks (such as program execution or calculation) or rely on human-provided feedback. Research has begun to explore how both

chain-of-thought and tree-of-thought methods can be augmented by reinforcement mechanisms, such as reward signals based on task accuracy or human preferences. In LLMs released since 2022 (InstructGPT, GPT3.5 and later models from other developers), this reinforcement has primarily taken place through reinforcement learning from human feedback⁶⁹.

Combining prompting strategies like the tree-of-thought strategy with reward-based evaluation is a promising direction for encouraging more reliable and interpretable reasoning outputs. These approaches might help to bridge rapid output generation with deliberative, self-monitoring behaviours that are characteristic of System 2 thinking. There is then a clear although imperfect parallel with dual-process theory and heuristic search models of cognition⁶⁵. The LLM produces a tree of possible next steps in the problem-solving process using fast, unidirectional passes through the fixed parameters learned from the training data. Each node in this tree can be evaluated using direct human feedback, a learned function, an algorithm to verify calculations or a heuristic for generating more analytic nodes at the next stage. In practice, this evaluation-and-reward stage can be a fast process, but it fulfils the deliberative, monitoring role characteristic of System 2.

Origins of decisions and errors

Although LLMs and humans often arrive at similar decisions, the processes underlying these outputs are fundamentally different. Human reasoning is shaped by cognition, memory and lived experience, whereas LLMs generate responses based on statistical associations learned during training. As a result, LLMs are prone to particular shortcomings, such as hallucinations and overconfidence, which can lead to types of error that differ from those typically made by humans. These contrasts raise important questions about

whether LLMs can be said to ‘understand’ prompts in any meaningful sense and about how their decision-making should be interpreted. The following section examines the origins and consequences of these divergences between human and LLM decision-making processes.

Bias and confidence

Although LLMs can replicate cognitive biases often observed in human thinking, they also display distinctive failure modes, including ‘hallucinations’ and overconfidence.

Hallucinations refer to cases in which LLMs generate outputs that include incorrect or fabricated information, producing responses that seem plausible but are factually inaccurate^{8,85–87}. There is some debate surrounding the use of the term ‘hallucination’ to describe this facet of LLM outputs, with some arguing that this term erroneously anthropomorphises LLMs by suggesting human-like subjective experiences⁸⁸. As such, the term ‘confabulation’ has been suggested to be a more fitting description⁸⁸, denoting how information can be fabricated without it being recognised as being false. Although ‘confabulation’ might be a more accurate description, we continue to use the term ‘hallucination’ here because it remains the most widely recognised term in the literature. A related issue is overconfidence, in which models produce outputs with high confidence when answering questions incorrectly, owing to their inability to extend beyond the provided prompt or express uncertainty^{48,86,89,90}. Overconfidence often compounds the impact of hallucinations because fabricated or inaccurate outputs are presented with unwarranted certainty, increasing the likelihood that users will accept them as accurate^{86,90}.

What distinguishes AI hallucinations and overconfidence from similar phenomena in humans – such as false memories, confabulations in clinical contexts and overconfidence bias in decision-making – are their origins. In humans, overconfidence is associated with judgement and decision-making biases that can lead people to overestimate the accuracy of

their beliefs⁹¹, whereas confabulations are linked to neuropsychological processes such as memory distortions⁹². In LLMs, these properties arise from architectural features and training methods, such as autoregression and reinforcement learning from human feedback. LLMs are trained to predict the next word in a sequence on the basis of large-scale text corpora. Although this training is effective at producing linguistically valid responses, in the absence of validation from external sources the LLM often produces factually inaccurate or fictional but plausible-sounding answers based on potentially limited or flawed training data (that is, hallucinations^{87,93,94}). By predicting the next word based on patterns observed in training data, LLMs can produce fluent but unsupported continuations when no reliable fact is available. Although reinforcement learning from human feedback helps to align outputs with user expectations, it does not provide real-time fact checking and therefore cannot fully prevent hallucinations. The precise interactions between flawed training data, lack of external validation and autoregressive generation are not yet fully understood, but together they contribute to the persistence of hallucinations in LLM outputs. Overconfidence is similarly shaped by reinforcement learning from human feedback, given that human raters tend to favour fluent and confident-sounding responses even when they cannot verify the accuracy of the response, thereby inadvertently rewarding a confident tone over correctness^{69,95,96}. As a consequence, models can learn to present incorrect outputs with unwarranted certainty, reflecting the reinforcement of a confident style over accuracy^{97,98}. Thus, whereas human overconfidence often serves adaptive functions such as bolstering credibility or morale^{92,99}, LLM overconfidence arises only as a by-product of how the models are trained and functions mainly to make incorrect responses seem more persuasive to users. Furthermore, current LLMs lack the capacity to critically evaluate their decisions, rendering them unable to assess the accuracy of their outputs or the appropriateness of a given prompt in relation to their training data. As a consequence of both the lack of internal mechanisms for self-evaluation

and the reinforcement of confident responses, LLMs can generate content that is expressed with unwarranted certainty but lacks the factual grounding needed to convince users of their accuracy of their output, particularly when prompted with inputs that fall outside their training data⁹⁵. Although fine-tuning approaches such as introspection-based methods⁸⁵ and self-reflection strategies¹⁰⁰ can encourage more cautious responses by reducing overconfidence and requiring model confidence to more accurately reflect the likelihood that a response is correct, further work is needed to understand and mitigate these biases, particularly in applied settings in which confidently stated but incorrect outputs can distort reasoning or decision-making¹⁰¹.

Performance and understanding

The mechanisms underlying the System 2-like decision-making of LLMs are likely to be fundamentally different from human cognition. Although some research groups argue that advanced models might approximate aspects of conceptual reasoning¹⁰², most evidence suggests that their outputs are better explained by statistical pattern recognition from learned examples^{86,103,104}. They excel when queries are similar to patterns present in their training data, but struggle with novel problems. For instance, although LLMs often perform poorly when confronted with entirely unfamiliar problems (‘zero-shot learning’), they can exhibit ‘few-shot learning’ whereby they produce more accurate responses after being trained in a given scenario or when provided with an example^{105,106}. When confronted with unfamiliar questions – such as questions about information that was discovered following model training or nonsense reasoning problems – LLMs falter, producing hallucinations or flawed reasoning^{85,105–107}. For instance, prompting ChatGPT with a question based on a medical discovery that did not appear in its training data set (which can be thought of as a rough analogue to long-term memory) led it to generate responses that were factually inaccurate or illogical¹⁰⁷, leading to instances of misinformation and unwarranted confidence in its outputs.

In addition to these gaps in the training data set, humans also differ from LLMs in their use of working memory, which enables them to actively manipulate information to explore problems more fully, whereas LLMs are restricted to processing only the information contained within their immediate context window^{22,85,107–109}. Furthermore, a language model could never truly grasp the sensation of being tickled given that they lack physical bodies; this absence of embodied experience highlights a fundamental difference from human cognition and might result in a narrower and less accurate representation of reality¹⁰⁸ (Box 3). Hence, even when LLMs match human performance benchmarks, they might reach correct conclusions along very different pathways. Accordingly, their skills might 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 repeating what they were trained on^{22,103,110}. This lack of transparent reasoning is problematic in applied contexts because without knowing why a response is correct, it is difficult to assess its reliability or how the model will perform when faced with new or high-stakes situations.

This pattern-matching approach to decision-making raises questions about whether AI systems can achieve genuine human-like understanding and the ability to respond to questions. Explanations in favour of LLMs’ ability to understand a prompt surrounds their ability to often ‘zero-shot’ a problem or how they relate a prompt to their training data. In contrast to ‘few-shot’ problem-solving, in which the model is provided with several in-context examples of solutions or reasoning steps before generating an answer, zero-shot approaches prompt LLMs to solve problems 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)^{72,111}. This ability to respond to problems without a sample solution or to recognise a problem and correctly apply it to the appropriate part of their training data⁷² suggests that LLMs do generalise beyond superficial reproduction of training examples.

Another observed phenomenon of this type is referred to as ‘grokking’. Originally coined by a science-fiction author to refer to deep understanding¹¹², in the context of LLM problem-solving the term refers to generalisation beyond overfitting by recognising crucial patterns, for instance, making use of circular geometry to perform modular addition¹¹³. When trained far beyond the point of overfitting on certain simple tasks, transformer-based neural networks can progress from merely interpolating the training data in high-dimensional space to learning the algorithm or structure underlying the data generating process¹¹³. This study focused on modular arithmetic problems and the network eventually developed a geometric representation of the data that enabled full generalisation, undergoing a sudden shift after an extended period of seeming performance stagnation. Although these demonstrations of zero-shot capabilities and emergent ‘grokking’ phenomena suggest more sophisticated processes than simple pattern matching, the question remains whether these constitute genuine understanding or merely more advanced forms of statistical association.

A parallel perspective that addresses whether LLMs truly understand their outputs suggests that these systems can use their training data similarly to how humans use their past experiences to inform present decision-making¹¹⁴. This perspective proposes that through exposure to diverse textual information across domains such as medicine, business or law, LLMs can acquire capabilities that mirror human domain expertise, which potentially enables them to reason beyond the limitations of their training data (such as past temporal cutoffs after which they have no training data)^{12,115}. 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 to increase their facility within a topic¹¹⁶. Additionally, even if LLMs do not understand their prompt in a human sense, they might have a nonhuman form of ‘understanding’ that emerges through statistical correlations¹⁰⁸. The pattern-matching capabilities of LLMs – which enable 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 clarify this debate, 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. There is much discussion both in the academic and public-facing literature on questions of whether LLMs have, or could possibly exhibit, ‘true’ understanding of the problems they are presented with^{108,117,118}. The ability to generalise beyond training data, to show explicit steps in the problem-solving process and to learn structural patterns that enable generalisation beyond overfitting certainly encompass some of the properties that people associate with understanding a problem¹¹⁹. However, the literature is far from settled on each of these questions: zero-shot problem-solving is often brittle and it is often difficult to identify problem-evaluation sets that are entirely absent from the training data. The reasoning traces elicited with chain-of-thought prompting have sometimes been shown not to be faithful to the model’s actual reasoning process and ‘grokking’ has so far only been demonstrated in limited contexts. Finally, ‘understanding’ is not a strictly well-defined scientific or philosophical concept and arguments about whether LLMs can understand are intertwined with arguments about which properties and mechanisms constitute the concept. Although such discussions can drive innovation, the purely semantic dispute might offer limited operational guidance, as Edsger Dijkstra suggested when he wrote that the question of whether computers can think “is just as relevant and just as meaningful as the question of whether submarines can swim”¹²⁰. Human and LLM responses might seem comparable at first, but the underlying mechanisms in how they reach the same or different decisions are distinct. Although both humans and LLMs can be incorrect in their reasoning, LLMs have inherent limitations that humans do not, which can exacerbate their flaws and lead to incorrect decisions. LLMs’ algorithms and training data can cause them to reach a correct conclusion without truly understanding their prompt in the same way as humans do. These differences in internal

mechanisms highlight the need for caution when assessing the performance of LLMs and future research needs to thoroughly contemplate the reasons for why and how LLMs output the responses they do in a given paradigm.

Assisting human decision-makers

As the decision-making abilities of LLMs improve, it is important to consider how they can assist with human decision-making. LLMs are particularly effective at supporting decision-making by synthesising large amounts of information, highlighting relevant patterns that might not be immediately obvious to humans and generating alternative perspectives. These strengths have already been successfully leveraged in several real-world decision-making applications. For instance, LLMs have the potential to enhance medical decision-making thanks to their training on openly available medical databases^{13,121}. For example, GPT-4 has demonstrated impressive performance on sample questions from the United States Medical Licensing Examination¹²² and GPT-3.5's o3 model has been shown to outperform physicians on sequential diagnostic tasks¹²³. In business, LLMs can aid decision-making by summarising information, generating reports and creating strategic plans based on real-time data to assist professionals in management, finance and corporate decision-making¹²⁴. Similarly, LLMs can be used as a research tool, aiding in tasks such as literature reviews and hypothesis generation^{17,125,126}. These applications point to the promise that LLMs hold for improving decision-making practices across a variety of fields.

However, LLMs have notable limitations that must be considered for applications in augmenting human decision-making. Although they can excel at generating plausible responses, their outputs can be prone to inaccuracies⁵ and raise concerns of potentially plagiarising from existing sources¹²⁵, complicating the reliability and authenticity of the content¹²⁷. Moreover, the capabilities of these models do not equate to an expert in the

relevant field. For example, ChatGPT's input on business decisions lacked expert-level insight and instead provided generic overviews and explanations¹²⁸. Beyond concerns about accuracy, the widespread adoption of LLMs also raises broader societal risks, including the potential erosion of job quality and increased unemployment caused by the automation of tasks traditionally carried out by humans^{129,130}. Thus, present attempts to integrate LLMs into the workplace should focus on how they can be used to complement human expertise rather than replace it, particularly in high-stakes decision-making contexts^{12,115,122}.

The increasing reliance on LLMs in real-world decision-making raises important ethical concerns, particularly regarding biases and overconfidence. LLMs are prone to errors and biases, including gender and racial prejudices, that undermine their reliability and fairness in real-life scenarios^{90,122,133,134}. LLMs can reflect and even amplify biases from their training data, which is particularly concerning in high-stakes fields such as healthcare and law, where accuracy is paramount¹¹⁰. Their optimisation for fluency rather than accuracy leads to a tendency to present incorrect information with certainty, which further complicates their reliability and producing risks in high-stakes contexts in which overreliance on AI systems could have serious consequences. Addressing these risks requires an interdisciplinary approach involving researchers in ethics, AI and cognitive psychology. Ethical analysis provides guidelines for responsible AI use that emphasise fairness, accountability and transparency¹³⁵. AI research can help to reduce biases and improve reliability through better data curation and model refinement. Meanwhile, cognitive psychology contributes valuable insights into decision-making mechanisms – both human and AI – to help shape models that more closely align with human reasoning and avoid common cognitive traps^{16,136}. Together, these disciplines can offer a framework for mitigating the ethical challenges posed by LLMs and ensure that they support rather than undermine human decision-making.

To harness LLMs as effective decision-making aids, it is imperative to better understand how to integrate them into workflows that leverage their strengths and offset their limitations. Research should determine what forms of oversight best mitigate the risk of hallucinations in workplace settings and how future LLM architectures can be optimised for transparency and accountability without compromising their practical utility. For instance, it is an open question how prompt design and context tailoring can improve decision-making performance. Humans can consider a multitude of different factors in making decisions, drawing on social understanding, embodied and situational awareness (Box 3), accumulated experience and long-term foresight^{12,22}, whereas LLMs rely solely on context windows and past training^{12,22,107,115,130}. As such, LLMs should function as support tools under expert supervision, providing alternative perspectives and suggestions to aid in decision-making while remaining open to human correction^{3,4,114,131}. This model of use aligns with current usage guidelines, such as those from OpenAI¹³², which emphasise that LLMs should be used as assistive tools rather than autonomous decision-makers^{4,12,122}. Furthermore, extended use of AI-generated solutions also pose the risk of diminishing humans' capacity for critical and innovative thinking (Box 4). As the field advances, further research is needed to clarify when and how LLMs add value to human reasoning. Addressing these questions will be critical to ensuring that the deployment of LLMs in applied settings is dependable, fair and aligned with human values.

Summary and future directions

As new LLMs are released and older models are updated, their abilities have steadily improved. They have demonstrated promising skills in a variety of cognitive domains, being able to match or outperform humans in some areas^{7,16}. Despite these advances, the internal decision-making processes of LLMs remain a 'black box' from a cognitive perspective¹²⁶. Although researchers are beginning to gain insights into how these models behave,

fundamental questions remain about whether their statistically driven pattern recognition can mirror aspects of human cognition. Concerns also persist regarding their susceptibility to biases, errors and misinformation, which can substantially affect their reliability in real-world applications. Addressing these challenges requires methodologically rigorous studies and carefully designed psychological experiments to better characterise the nature of LLM decision-making and the development of improved bias detection, misinformation filtering and transparency mechanisms to ensure the responsible deployment of LLMs in the world.

One important step towards improving our understanding of the decision-making capabilities of LLMs is for researchers to incorporate a wide range of prompt variations and psychological tests to provide a comprehensive picture of how these models operate. Previous studies have indicated that slight alterations in prompt wording can produce markedly different responses, owing to a specific prompt or test being present in the model's training data or biases caused by the wording of prompts^{16,22,39,80,82}. Using diverse prompts can help to account for this sensitivity to prompt phrasing and enable research to become more robust and resistant to these confounding variables. Furthermore, incorporating diverse testing scenarios enables researchers to identify specific contexts in which biases and heuristic-driven behaviours are exhibited or absent – an important limitation of existing research investigating biases and heuristics in LLMs. When creating prompts for research, it is important to validate them first to ensure that they accurately target the desired variable¹⁸. Additionally, the exact wording of the prompts used in studies should be provided by the researchers to make readers aware of any potential flaws in prompt design and to enable future replications (a study argues that prompts function like code and therefore must be documented transparently)¹³⁷.

A limitation of current research is that the comparisons between LLMs 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 human performance and when they do these comparisons tend to be underpowered with small sample sizes and forgo reporting statistics such as standard deviations⁷. Future studies that aim to compare LLMs to humans should address these limitations while also moving beyond merely using humans as benchmarks for accuracy. The ability of LLMs to match or surpass human performance on specific tasks does not imply equivalence in real-world decision-making^{103,122}. Simply comparing outcomes overlooks key factors that contribute to decision-making processes, such as the extent to which LLMs rely on statistical pattern recognition rather than comprehension, sensitivity to contextual cues and integration of prior knowledge in a meaningful way²². Additionally, it is critical to consider factors such as adaptability and metacognitive awareness and how uncertainty estimates shape decision-making in order to draw meaningful comparisons. Without considering these broader cognitive and computational factors, assessments of LLM decision-making risk conflating superficial performance similarities with deeper cognitive equivalence. Probing LLMs for justifications about their decisions could provide valuable insights into these decision-making processes. Analysing these justifications could help to reveal whether LLMs rely on genuine reasoning or heuristics – even when they seem to be engaging in System-2-like thinking – as well as to uncover biases, gaps in understanding and limitations in generalisation.

A major challenge for both human cognition and AI systems is determining when deliberative reasoning should be deployed. Early dual-process accounts often implied that System 2 intervenes only after System 1 produces an error but contemporary work highlights the role of conflict detection and metacognitive monitoring in triggering the escalation to System 2 (ref. ¹³⁸). Moreover, some researchers have argued that humans have an intuitive sensitivity to conflict that can flag the need for further processing before conscious deliberation begins¹³⁹. In this view, System 2 is conditionally activated when initial responses

feel uncertain, inconsistent or otherwise unsatisfactory (two studies further discuss the conditions that trigger System 2 engagement)^{77,138}. The problem of knowing when to engage System 2 has clear implications for LLM architecture and use. Although most LLMs rely on static prompting strategies, certain models employ dynamic routing mechanisms that shift from fast, low-cost inference to more resource-intensive, multistep reasoning when internal uncertainty or error likelihood is high^{70,71}. These architectures mirror human metacognition by allocating resources conditionally, engaging System-2-like reasoning only when the situation warrants it. This effort to allocate computational resources conditionally, based on internal uncertainty, is a clear example of how LLM architectures are increasingly drawing inspiration from studies of human cognition. A key future direction will be to develop more refined mechanisms for monitoring and regulating uncertainty that will enable models to adjust their processing strategies in ways that improve their reliability while also making their decision processes more transparent to users.

Building on these architectural developments, advances increasingly position LLMs as interactive agents that are capable of autonomously interacting in an environment and can perceive their environment, maintain memory, use tools and select actions based on explicit goals or reward functions. Such agent-based systems offer structured frameworks for guiding LLMs through complex, goal-oriented decision-making, which can help to mitigate limitations associated with static prompting and context-dependence. Developments in agentic frameworks such as ReAct¹⁴⁰, Tree-of-Thoughts⁶⁶ and Voyager¹⁴¹ echo dual-process distinctions between fast, reactive responses and slow, deliberative reasoning. Beyond system design, new benchmarks in single-agent environments^{142–144} and multi-agent settings^{145–148} are emerging to evaluate how LLMs make decisions over time, adapt strategies and interact with others. These studies have begun to reveal both convergence and divergence between LLM and human behaviour: although LLMs sometimes replicate heuristics and biases

observed in human decision-making (such as satisficing or anchoring), they diverge substantially in consistency, adaptability and long-term reasoning, especially under uncertainty¹⁴⁷. As such, these interactive environments provide a valuable new lens through which to examine whether LLM outputs align with predictions from dual-process theory and to systematically compare their cognitive profiles with those of human agents.

As LLMs become increasingly integrated into real-world decision-making processes, clear guidelines and regulatory frameworks must be established to ensure their responsible and effective use. We recommend that future research investigates methods and regulations aimed at avoiding misuse^{122,130,149}. Developing methods and guidelines to use LLMs in different settings (such as flowcharts for academic writing with AI)⁴ could help to counter risks such as misinformation and loss of data security as well as to facilitate application of LLMs in these areas. Similarly, 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 to reduce the variability in the results that these investigations produce and promote replicable studies^{4,18,150}.

In conclusion, LLMs offer powerful decision-support capabilities, yet their reasoning processes remain largely opaque and it remains unclear how they make decisions. Their biases, susceptibility to misinformation and lack of genuine understanding also necessitate ongoing scrutiny, particularly in high-stakes applications. As these models become more integrated into society, addressing these challenges requires rigorous research to uncover the mechanisms behind LLMs decision-making, to develop robust safeguards against biases and errors and to enhance their transparency. Future work should prioritise interdisciplinary collaboration across psychology, cognitive science, AI research and policy regulation to

ensure that LLMs are not only high-performing but also interpretable, reliable and ethically aligned with human decision-making.

Boxes

Box 1 – Large language model architectures

Following the rapid advances in performance of deep neural systems in the early 2010s, different architectures have been at the forefront for various tasks: convolutional neural networks for image recognition, recurrent architectures such as long short-term memory networks¹⁵⁴ for sequence data and, most recently, the transformer architecture for modelling dependencies in sequential input. The transformer architecture uses a mechanism called self-attention to compute representations of input sequences by relating each element to every other. The original transformer was proposed in an encoder–decoder format, in which the encoder maps an input sequence (such as a sentence) into a contextual representation and the decoder generates an output sequence (such as a translation) based on that representation. However, modern large language models (LLMs) such as GPT-3, GPT-4, Claude and Llama are based on decoder-only transformers that are optimised for autoregressive next-token prediction: during training the model learns to predict each token from the tokens that come before it and during generation it appends each new token to the context and uses this expanded sequence to predict the next. Autoregressive language models can achieve good performance at multiple tasks without further fine-tuning¹⁵⁵ and it is these generative, pre-trained decoder-only models that we primarily discuss in this Review.

By this definition, LLM architectures are almost inherently System-1-like in their computation: an output is generated with a single pass propagating in one direction through the layers of the network; there is no reflection or explicit symbolic logical inference. However, much of the progress on LLM benchmarks has come with the expansion of few-

shot-learning to include chain-of-thought tokens⁴¹, ‘hidden’ reasoning tokens¹⁵⁶ or even meaningless filler tokens used only to aid quantification¹⁵⁷. These methods, often collectively referred to as ‘test-time compute’ or ‘in-context learning’, much more closely resemble System 2 processes: slow, sequential, explicit and self-monitoring. In fact, the reinforcement learning of DeepSeek-R1 specifically learns to reward backtracking tokens (such as ‘wait’) in context from a combination of a formatting reward that encourages reflection through <think> tags and automatically verifiable accuracy rewards for programming and mathematical tasks¹⁵⁸.

In-context learning has led to improvements on common benchmarks such as question answering and language understanding¹⁰², accompanied by improved performance on several tasks designed to evaluate cognitive behaviour in humans, such as theory of mind¹⁵⁹ and analogical reasoning¹¹¹. On semantic illusions and cognitive reflection tests, early OpenAI generative pre-trained transformer (GPT) models give evasive and atypically incorrect answers, the GPT-3-davinci series of models primarily produce the intuitive but incorrect answers that the tests are designed to elicit in humans, and ChatGPT-3.5 and later models primarily arrive at the correct answer with the use of chain-of-thought prompting or ‘reasoning’ tokens³⁹ (although it is not always the case that the reasoning tokens are faithful to the models underlying the problem-solving process). Although these improvements suggest that newer LLMs display more System-2-like behaviour, their ability to sustain complex, multistep reasoning in dynamic environments remains limited. The finding that meaningless filler tokens can improve performance, along with studies that demonstrate that LLMs often generate chain-of-thought explanations that do not faithfully reflect the underlying computations used to reach an answer¹⁶⁰, highlight a lack of faithfulness between supposed chain-of-thought explanations and actual problem-solving processes. This is an ongoing and contested area, investigating the conditions under which reasoning traces might

or might not be faithful to underlying mechanisms¹⁶¹. Future research into LLM architectures will need to integrate more advanced mechanisms to enhance robustness, adaptability and self-reflective reasoning, which are core components of System 2 cognition in humans.

Box 2- The potential dangers of anthropomorphic seduction

The very qualities that make artificial intelligence (AI) systems most useful and accessible to humans simultaneously pose the greatest risks to unsuspecting users. This phenomenon, described as ‘anthropomorphic seduction’¹⁶², represents a fundamental shift in how humans must conceptualise and regulate artificial intelligence. Traditionally, anthropomorphism was framed as a cognitive illusion on the part of the user; humans anthropomorphise because they perceive minds everywhere¹⁶³. But this view becomes untenable when the systems themselves increasingly exhibit anthropomorphic characteristics in their outputs¹⁶². For example, the Turing test was once considered ‘anthropomorphism-proofed’ because it explicitly instructs judges to assume they are evaluating a machine, creating a built-in scepticism that should prevent them from projecting human qualities onto the AI system¹⁶². However, when large language models (LLMs) consistently succeed in fooling judges¹⁶⁴, the explanation lies not in user bias, but in the systems’ ability to convincingly emulate human-like traits. In this sense, anthropomorphism is no longer merely a user-side illusion – it is embedded in the interaction itself.

This situation raises a fundamental design question of whether developers should continue refining LLMs to mimic human interaction or instead work to dehumanise their outputs¹⁶². On the one hand, LLMs that feel conversational, empathetic and context-sensitive unlock a host of benefits. They make digital interfaces more accessible, enabling users to interact with complex systems through natural language. This access facilitates medical question-and-answer tools¹⁶⁵, automated tutoring¹⁶⁶ and decision-support systems in financial and legal settings, for example. At the same time, the very qualities that make

anthropomorphic agents potentially helpful also make them uniquely dangerous. When machines can write persuasively, match emotional tone and mimic human-like conversation, the potential for deception, manipulation and misplaced trust rises rapidly, with evidence to suggest that many users already attribute some possibility of consciousness to LLMs¹⁶⁷. Furthermore, LLMs present new risks in domains such as political persuasion¹⁶⁸, mental health¹⁶⁹ and digital companionship¹⁷⁰. Unlike humans, LLMs do not possess intrinsic motives or ethical intuition and therefore have the potential to spread misinformation with unprecedented speed and subtlety. In contexts such as customer service, education or therapy, anthropomorphic LLMs could build trust that might be unwarranted and, in some cases, harmful. If users are unaware that they are interacting with a nonhuman system, a system designed to obfuscate this fact, the resulting confusion has serious ethical and regulatory implications.

Despite these concerns, efforts to discourage anthropomorphism are likely to become less effective over time as AI systems become increasingly sophisticated and can mimic humans with even greater precision. The uncomfortable truth is that the genie is already out of the bottle – human-like dialogue in AI systems is a publicly available product. Thus, the path forwards is to govern anthropomorphism, which will require thoughtful design practices, such as mandating that AI systems clearly disclose their status as an artificial intelligence at the beginning of interactions with human users¹⁷¹. It might also require stronger regulatory oversight, particularly in high-risk domains such as healthcare settings. Most importantly, these developments demand a new kind of digital literacy that equips the public to engage with seemingly emotionally intelligent machines with both clarity and informed scepticism. Scientists and researchers also have a key part to play in resisting the seductive pull of anthropomorphism. Doing so requires discipline and the deliberate practice of repeatedly stepping back to reflect on how LLMs – and the broader systems in which they are embedded – actually work¹⁷². Maintaining this critical distance will help us not to misattribute human-

like qualities to systems whose behaviour arises from pattern recognition and probabilistic modelling rather than conscious intent and understanding.

In short, anthropomorphic LLMs offer powerful opportunities to democratise access to information and support. But the same qualities, if left unchecked, could also destabilise trust, privacy and public discourse. The challenge is to balance the promise of human-like interaction with safeguards that mitigate its most seductive and dangerous effects.

Box 3- Embodied Cognition and Decision-Making in LLMs

Unlike artificial intelligence (AI) systems, human cognition is inherently grounded in sensorimotor experiences, enabling humans to adapt to changing contexts through feedback from their bodies. This embodied cognition enables flexibility and generalisation in novel situations – capabilities that remain limited in large language models (LLMs). Although LLMs excel in processing and generating language, they sometimes struggle with causal reasoning and lack the physical intuition required for decision-making in dynamic environments¹⁷². An ‘embodied Turing test’ was recently proposed to evaluate the ability of AI systems to interact and make decisions in the physical world¹⁷³. This test would assess sensorimotor abilities and adaptive decision-making, similar to the flexible behaviour observed in humans. Humans rapidly acquire new skills without large labelled data sets, instead relying on memory, cognitive maps and hierarchical knowledge representations to efficiently adapt to changing conditions^{174,175}. By contrast, AI systems often falter when confronted with slight perturbations or rule changes, because they lack these structured representations and real-time feedback from the environment¹⁷⁶. Neuroscience-inspired approaches, such as integrating episodic memory and planning mechanisms, could help to develop more flexible, context-aware AI systems¹⁷⁷. Moreover, grounding models in real-world sensorimotor interactions and adaptive learning could move AI systems closer to passing the embodied Turing test, equipping them to navigate complex, unpredictable

environments more as humans do^{173,178}. Such approaches might bridge the gap between artificial and biological intelligence, enhancing the decision-making capabilities of AI systems in real-world scenarios.

Box 4- Large Language Models and Human Cognition

Although the focus of this Review is on the decision-making capabilities of large language models (LLMs), the use of these models can also affect human decision-making. If used correctly, LLMs can enhance human decision-making processes^{179–181} and extend human decision-making capabilities by acting as a cognitive artefact or external decision-making agent^{24,182,183}. LLMs might take on part of the cognitive load, freeing humans to focus on more complex and strategic aspects of decision-making. However, there are also potential indirect negative consequences on human cognition that could arise from excessively outsourcing cognitive tasks to LLMs. For instance, in one study the presence of a biased artificial intelligence (AI) system led to participants making more mistakes on a medical diagnostic task and had a lasting effect on responses even after it was no longer available¹⁸⁴. Additionally, 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 training¹⁸⁵. 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 who believed that the advent of writing would negatively affect human memory^{186,187}. This position has persisted, with evidence to suggest that modern technologies have greatly influenced cognition. For instance, widespread use of smartphones and the internet are believed to have altered memory processes and reduced attention spans^{188,189} and GPS has eroded navigation and spatial reasoning skills^{190,191}. In this context, there are growing concerns that overreliance on LLMs for decision-making could potentially diminish human critical thinking and independent problem-solving abilities,

leading to a form of cognitive offloading in which complex reasoning is increasingly delegated to AI systems. For instance, the concept of ‘System 0’ describes a new system of thinking in which AI serves as an external thinking tool that complements human cognition¹⁸¹. Used correctly, this system has the potential to greatly assist humanity to address societal challenges that are beyond our natural cognitive capacities (two studies illustrate how LLMs can augment learning and memory and even reduce human susceptibility to conspiracy beliefs)^{192,193}. However, it is essential to continue questioning and critically evaluating AI outputs, along with providing education to aid in preventing the risks associated with overreliance on LLMs¹⁸¹. Given the uncertain long-term impact of LLMs on human decision-making and cognitive processes, there is also a need for longitudinal studies to inform policies and safeguards¹⁹⁴. 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.

Acknowledgements

The authors thank R. O’Connell and M. Wu for feedback and comments on the review.

References

1. Pierson, B. Lawsuit claims UnitedHealth AI wrongfully denies elderly extended care. *Reuters* (2023).
2. Schreiber, M. New AI tool counters health insurance denials decided by automated algorithms. *The Guardian* (2025).
3. Au Yeung, J. AI chatbots not yet ready for clinical use. *Front. Digit. Health* (2023).
4. Azaria, A., Azoulay, R. & Reches, S. ChatGPT is a remarkable tool — for experts. *Data Intell.* (2024).
5. Meyrowitsch, D. W., Jensen, A. K., Sørensen, J. B. & Varga, T. V. AI chatbots and (mis)information in public health: impact on vulnerable communities. *Front. Public Health* (2023).
6. Kim, J. H. When ChatGPT gives incorrect answers: the impact of inaccurate information by generative AI on tourism decision-making. *J. Travel Res.* (2023).
7. Sartori, G. & Orrù, G. Language models and psychological sciences. *Front. Psychol.* (2023).
8. Dillion, D., Tandon, N., Gu, Y. & Gray, K. Can AI language models replace human participants? *Trends Cogn. Sci.* (2023).
9. Kaddour, J. Challenges and applications of large language models. *arXiv* (2023).
10. Griewing, S. Evolution of publicly available large language models for complex decision-making in breast cancer care. *Arch. Gynecol. Obstet.* (2024).
11. Jimenez, C. E. SWE-bench: can language models resolve real-world Github issues? *arXiv* (2024).
12. Jusman, I. A., Ausat, A. M. A. & Sumarna, A. Application of ChatGPT in business management and strategic decision making. *J. Minfo Polgan* (2023).

13. Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* (2023).
14. Basir, A., Puspitasari, E. D., Aristarini, C. C., Sulastris, P. D. & Ausat, A. M. A. Ethical use of ChatGPT in the context of leadership and strategic decisions. *J. Minfo Polgan* (2023).
15. Gloria, B., Melsbach, J., Bienert, S. & Schoder, D. Real-GPT: efficiently tailoring LLMs for informed decision-making in the real estate industry. *J. Real Estate Portf. Manag.* (2024).
16. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl Acad. Sci. USA* (2023).
17. Demszky, D. Using large language models in psychology. *Nat. Rev. Psychol.* (2023).
18. Hagendorf, T. Machine psychology: investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv* (2023).
19. White, J. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv* (2023).
20. Abbate, F. Natural and artificial intelligence: a comparative analysis of cognitive aspects. *Minds Mach.* (2023).
21. Rich, A. S. & Gureckis, T. M. Lessons for artificial intelligence from the study of natural stupidity. *Nat. Mach. Intell.* (2019).
22. Shiffrin, R. & Mitchell, M. Probing the psychology of AI models. *Proc. Natl Acad. Sci. USA* (2023).
23. Suri, G., Slater, L. R., Ziaee, A. & Nguyen, M. Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *J. Exp. Psychol. Gen.* (2024).

24. Qu, Y. Promoting interactions between cognitive science and large language models. *Innovation* (2024).
25. Thaler, R. H. Behavioral economics: past, present, and future. *Am. Econ. Rev.* (2016).
26. Kahneman, D. Maps of bounded rationality: psychology for behavioral economics. *Am. Econ. Rev.* (2003).
27. Evans, J. S. & Stanovich, K. E. Dual-process theories of higher cognition: advancing the debate. *Persp. Psychol. Sci.* (2013).
28. Evans, J. S. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* (2008).
29. Kahneman, D. & Frederick, S. Heuristics and Biases: The Psychology of Intuitive Judgment. *Cambridge Univ. Press* (2002).
30. Gigerenzer, G. & Brighton, H. Homo heuristics: why biased minds make better inferences. *Top. Cogn. Sci.* (2009).
31. Kahneman, D. & Klein, G. Conditions for intuitive expertise: a failure to disagree. *Am. Psychol.* (2009).
32. Smolensky, P. Connectionist AI, symbolic AI, and the brain. *Artif. Intell. Rev.* (1987).
33. Goel, A. Looking back, looking ahead: symbolic versus connectionist AI. *AI Mag.* (2022).
34. Bellini-Leite, S. C. Dual process theory: embodied and predictive; symbolic and classical. *Front. Psychol.* (2022).
35. Bellini-Leite, S. C. Dual process theory for large language models: an overview of using psychology to address hallucination and reliability issues. *Adapt. Behav.* (2023).

36. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* (2013).
37. Newell, A. Physical symbol systems. *Cogn. Sci.* (1980).
38. Marcus, G. The next decade in AI: four steps towards robust artificial intelligence. *arXiv* (2020).
39. Hagendorf, T., Fabi, S. & Kosinski, M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nat. Comput. Sci.* (2023).
40. Ma, D., Zhang, T. & Saunders, M. Is ChatGPT humanly irrational? *Res. Sq.* (2023).
41. Wei, J. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* (2022).
42. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* (2011).
43. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. *Science* (1974).
44. Frederick, S. Cognitive reflection and decision making. *J. Econ. Perspect.* (2005).
45. Erickson, T. D. & Mattson, M. E. From words to meaning: a semantic illusion. *J. Verbal Learn. Verbal Behav.* (1981).
46. Toplak, M. E., West, R. F. & Stanovich, K. E. The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Mem. Cogn.* (2011).
47. Pennycook, G., Cheyne, J. A., Koehler, D. J. & Fugelsang, J. A. Is the cognitive reflection test a measure of both reflection and intuition? *Behav. Res. Meth.* (2016).

48. Chen, Y., Kirshner, S. N., Ovchinnikov, A., Andiappan, M. & Jenkin, T. A manager and an AI walk into a bar: does ChatGPT make biased decisions like we do? *Manuf. Serv. Oper. Manag.* (2025).
49. Jones, E. & Steinhardt, J. Capturing failures of large language models via human cognitive biases. *Adv. Neural Inf. Process. Syst.* (2022).
50. Pantana, G., Castello, M. & Torre, I. Examining cognitive biases in ChatGPT 3.5 and ChatGPT 4 through human evaluation and linguistic comparison. *AMTA* (2024).
51. Ryu, J., Kim, J. & Kim, J. A study on the representativeness heuristics problem in large language models. *IEEE Access* (2024).
52. Tversky, A. & Kahneman, D. The framing of decisions and the psychology of choice. *Science* (1981).
53. Nickerson, R. S. Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* (1998).
54. Lou, J. & Sun, Y. Anchoring bias in large language models: an experimental study. *arXiv* (2024).
55. Talboy, A. N. & Fuller, E. Challenging the appearance of machine intelligence: cognitive bias in LLMs and best practices for adoption. *arXiv* (2023).
56. Azaria, A. ChatGPT: more human-like than computer-like, but not necessarily in a good way. *IEEE ICTAI* (2023).
57. Acerbi, A. & Stubbersfield, J. M. Large language models show human-like content biases in transmission chain experiments. *Proc. Natl Acad. Sci. USA* (2023).
58. Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. & Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* (2022).

59. Gallegos, I. O. Bias and fairness in large language models: a survey. *Comput. Linguist.* (2024).
60. Wang, P., Xiao, Z., Chen, H. & Oswald, F. L. Will the real Linda please stand up...to large language models? Examining the representativeness heuristic in LLMs. *arXiv* (2024).
61. Nguyen, J. K. Human bias in AI models? Anchoring effects and mitigation strategies in large language models. *J. Behav. Exp. Finance* (2024).
62. Tversky, A. & Kahneman, D. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* (1983).
63. Tversky, A. & Kahneman, D. Availability: a heuristic for judging frequency and probability. *Cogn. Psychol.* (1973).
64. Ariely, D., Loewenstein, G. & Prelec, D. Coherent arbitrariness: stable demand curves without stable preferences. *Q. J. Econ.* (2003).
65. Newell, A. & Simon, H. A. Computer science as empirical inquiry: symbols and search. *Commun. ACM* (2007).
66. Yao, S. Tree of thoughts: deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* (2023).
67. Du, M. Machine vs. human, who makes a better judgment on innovation? Take GPT-4 for example. *Front. Artif. Intell.* (2023).
68. Ziegler, D. M. Fine-tuning language models from human preferences. *arXiv* (2019).
69. Ouyang, L. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* (2022).

70. Su, D., Sukhbaatar, S., Rabbat, M., Tian, Y. & Zheng, Q. Dualformer: controllable fast and slow thinking by learning with randomized reasoning traces. *arXiv* (2025).
71. Shang, Y., Li, Y., Xu, F. & Li, Y. Defint: a default-interventionist framework for efficient reasoning with hybrid large language models. *arXiv* (2024).
72. Kojima, T., Gu, S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* (2022).
73. Weston, J. & Sukhbaatar, S. System 2 Attention (is something you might need too). *arXiv* (2023).
74. Zhang, H., Huang, J., Li, Z., Naik, M. & Xing, E. Improved logical reasoning of language models via differentiable symbolic programming. *ACL* (2023).
75. Zhu, X. Solving math word problems via cooperative reasoning induced language models. *ACL* (2023).
76. Raelison, M., Thompson, V. A. & De Neys, W. The smart intuitor: cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition* (2020).
77. De Neys, W. & Pennycook, G. Logic, fast and slow: advances in dual-process theorizing. *Curr. Dir. Psychol. Sci.* (2019).
78. Reyna, V. F. & Brainerd, C. J. Numeracy, gist, literal thinking and the value of nothing in decision making. *Nat. Rev. Psychol.* (2023).
79. Bubeck, S. Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv* (2023).
80. Sclar, M., Choi, Y., Tsvetkov, Y. & Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design. *arXiv* (2023).
81. Pezeshkpour, P. & Hruschka, E. Large language models sensitivity to the order of options in multiple-choice questions. *ACL* (2024).

82. Loya, M., Sinha, D. A. & Futrell, R. Exploring the sensitivity of LLMs' decision-making capabilities: insights from prompt variation and hyperparameters. *ACL* (2023).
83. Barez, F. Chain-of-thought is not explainability. *arXiv* (2025).
84. Zhang, Z. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.* (2024).
85. Chakraborty, N., Ornik, M. & Driggs-Campbell, K. Hallucination detection in foundation models for decision-making: a flexible definition and review of the state of the art. *ACM Comput. Surv.* (2025).
86. Stella, M., Hills, T. T. & Kenett, Y. N. Using cognitive psychology to understand GPT-like models needs to extend beyond human biases. *Proc. Natl Acad. Sci. USA* (2023).
87. Huang, L. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* (2025).
88. Smith, A. L., Greaves, F. & Panch, T. Hallucination or confabulation? Neuroanatomy as metaphor in large language models. *PLoS Digital Health* (2023).
89. Singh, A. K., Lamichhane, B., Devkota, S., Dhakal, U. & Dhakal, C. Do large language models show human-like biases? Exploring confidence–competence gap in AI. *Information* (2024).
90. Borji, A. A categorical archive of ChatGPT failures. *arXiv* (2023).
91. Kopelman, M. D. Varieties of confabulation and delusion. *Cogn. Neuropsychiatry* (2010).
92. Johnson, D. D. P. & Fowler, J. H. The evolution of overconfidence. *Nature* (2011).

93. McCoy, R. T., Yao, S., Friedman, D., Hardy, M. & Griffiths, T. L. Embers of autoregression: understanding large language models through the problem they are trained to solve. *arXiv* (2023).
94. Perković, G., Drobnjak, A. & Botički, I. Hallucinations in LLMs: understanding and addressing challenges. *MIPRO* (2024).
95. Sun, F., Li, N., Wang, K. & Goette, L. Large language models are overconfident and amplify human bias. *arXiv* (2025).
96. Casper, S. Open problems and fundamental limitations of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.* (2023).
97. Zhou, K., Hwang, J. D., Ren, X. & Sap, M. Relying on the unreliable: the impact of language models' reluctance to express uncertainty. *arXiv* (2024).
98. Zhou, K., Jurafsky, D. & Hashimoto, T. Navigating the grey area: how expressions of uncertainty and overconfidence affect language models. *arXiv* (2023).
99. Dunning, D. Advances in Experimental Social Psychology Vol. 44. *Academic* (2011).
100. Ji, Z. Towards mitigating LLM hallucination via self reflection. *ACL* (2023).
101. O'Leary, D. E. Confirmation and specificity biases in large language models: an explorative study. *IEEE Intell. Syst.* (2025).
102. Wei, J. Emergent abilities of large language models. *Trans. Mach. Learn. Res.* (2022).
103. McCoy, R. T., Pavlick, E. & Linzen, T. Right for the wrong reasons: diagnosing syntactic heuristics in natural language inference. *ACL* (2020).
104. Webson, A. & Pavlick, E. Do prompt-based models really understand the meaning of their prompts? *ACL* (2022).

105. Lampinen, A. K. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus* (2024).
106. Brown, T. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* (2020).
107. Zheng, H. & Zhan, H. ChatGPT in scientific writing: a cautionary tale. *Am. J. Med.* (2023).
108. Mitchell, M. & Krakauer, D. C. The debate over understanding in AI's large language models. *Proc. Natl Acad. Sci. USA* (2023).
109. Nelson, A. B. & Shifrin, R. M. The co-evolution of knowledge and event memory. *Psychol. Rev.* (2013).
110. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? *ACM FAccT* (2021).
111. Webb, T., Holyoak, K. J. & Lu, H. Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* (2023).
112. Heinlein, R. A. Stranger in a Strange Land. *Putnam* (1969).
113. Power, A., Burda, Y., Edwards, H., Babuschkin, I. & Misra, V. Grokking: generalization beyond overfitting on small algorithmic datasets. *arXiv* (2022).
114. Seligman, M. E. P., Railton, P., Baumeister, R. F. & Sripada, C. Navigating into the future or driven by the past. *Persp. Psychol. Sci.* (2013).
115. Liu, S. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J. Am. Med. Inform. Assoc.* (2023).
116. Gu, Y. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* (2021).
117. Silva, G. A. Can AI understand what it's telling you? *Forbes* (2025).

118. Ananthaswamy, A. New theory suggests chatbots can understand text. *Quanta Magazine* (2024).
119. Moskvichev, A. K., Odouard, V. V. & Mitchell, M. The ConceptARC benchmark: evaluating understanding and generalization in the ARC domain. *Trans. Mach. Learn. Res.* (2023).
120. Dijkstra, E. W. On IPW's. *Univ. Texas Austin* (1983).
121. Thirunavukarasu, A. J. Large language models in medicine. *Nat. Med.* (2023).
122. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems. *arXiv* (2023).
123. Nori, H. Sequential diagnosis with language models. *arXiv* (2025).
124. Nie, Y. A survey of large language models for financial applications: progress, prospects and challenges. *arXiv* (2024).
125. Aydın, Ö. & Karaarslan, E. OpenAI ChatGPT generated literature review: digital twin in healthcare. *Emerg. Comput. Technol.* (2022).
126. Ke, L., Tong, S., Cheng, P. & Peng, K. Exploring the frontiers of LLMs in psychological applications: a comprehensive review. *Artif. Intell. Rev.* (2025).
127. Hua, S., Jin, S. & Jiang, S. The limitations and ethical considerations of ChatGPT. *Data Intell.* (2024).
128. Chuma, E. L. & de Oliveira, G. G. Generative AI for business decision-making: a case of ChatGPT. *Manag. Sci. Bus. Decis.* (2023).
129. Eloundou, T., Manning, S., Mishkin, P. & Rock, D. GPTs are GPTs: labor market impact potential of LLMs. *Science* (2024).
130. Weidinger, L. Ethical and social risks of harm from language models. *arXiv* (2021).

131. Chen, J., Liu, L., Ruan, S., Li, M. & Yin, C. Are different versions of ChatGPT's ability comparable to the clinical diagnosis presented in case reports? A descriptive study. *J. Multidisc. Healthc.* (2023).
132. OpenAI Introducing ChatGPT. *OpenAI* (2022).
133. An, J., Huang, D., Lin, C. & Tai, M. Measuring gender and racial biases in large language models. *arXiv* (2024).
134. Liang, P. P., Wu, C., Morency, L.-P. & Salakhutdinov, R. Towards understanding and mitigating social biases in language models. *PMLR* (2021).
135. Peters, D., Vold, K., Robinson, D. & Calvo, R. A. Responsible AI — two frameworks for ethical design practice. *IEEE Trans. Technol. Soc.* (2020).
136. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* (2017).
137. Liang, J. T., Lin, M., Rao, N. & Myers, B. A. Prompts are programs too! Understanding how developers build software containing prompts. *Proc. ACM Softw. Eng.* (2025).
138. De Neys, W. Bias and conflict: a case for logical intuitions. *Persp. Psychol. Sci.* (2012).
139. Neys, W. D. Advancing theorizing about fast-and-slow thinking. *Behav. Brain Sci.* (2023).
140. Yao, S. ReAct: synergizing reasoning and acting in language models. *arXiv* (2023).
141. Wang, G. Voyager: an open-ended embodied agent with large language models. *Trans. Mach. Learn. Res.* (2023).
142. Nathani, D. MLGym: a new framework and benchmark for advancing AI research agents. *arXiv* (2025).

143. Paglieri, D. BALROG: benchmarking agentic LLM and VLM reasoning on games. *ICLR* (2025).
144. Liu, X. Agentbench: evaluating LLMs as agents. *ICLR* (2024).
145. Huang, J. How far are we on the decision-making of LLMs? Evaluating LLMs' gaming ability in multi-agent environments. *arXiv* (2025).
146. Leng, Y. & Yuan, Y. Do LLM agents exhibit social behavior? *arXiv* (2024).
147. Costarelli, A. GameBench: evaluating strategic reasoning abilities of LLM agents. *NeurIPS* (2024).
148. Trencsenyi, V., Mensfelt, A. & Stathis, K. Approximating human strategic reasoning with LLM-enhanced recursive reasoners leveraging multi-agent hypergames. *arXiv* (2025).
149. Dwivedi, Y. K. Opinion paper: So what if ChatGPT wrote it? Multidisciplinary perspectives. *Int. J. Inform. Manag.* (2023).
150. Vaugrante, L., Niepert, M. & Hagendorf, T. A looming replication crisis in evaluating behavior in language models? Evidence and solutions. *arXiv* (2024).
151. Koo, R. Benchmarking cognitive biases in large language models as evaluators. *ACL* (2024).
152. Wang, Y., Cai, Y., Chen, M., Liang, Y. & Hooi, B. Primacy effect of ChatGPT. *ACL* (2023).
153. Mina, M., Ruiz-Fernández, V., Falcão, J., Vasquez-Reina, L. & Gonzalez-Agirre, A. Cognitive biases, task complexity, and result interpretability in large language models. *ACL* (2025).
154. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* (1997).

155. Radford, A. Language models are unsupervised multitask learners. *OpenAI Blog* (2019).
156. OpenAI Learning to reason with LLMs. *OpenAI* (2024).
157. Pfau, J., Merrill, W. & Bowman, S. R. Let's think dot by dot: hidden computation in transformer language models. *arXiv* (2024).
158. Guo, D. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* (2025).
159. Kosinski, M. Evaluating large language models in theory of mind tasks. *Proc. Natl Acad. Sci. USA* (2024).
160. Turpin, M., Michael, J., Perez, E. & Bowman, S. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Adv. Neural Inf. Process. Syst.* (2023).
161. Ameisen, E. Circuit tracing: revealing computational graphs in language models. *Transformer Circuits* (2025).
162. Peter, S., Riemer, K. & West, J. D. The benefits and dangers of anthropomorphic conversational agents. *Proc. Natl Acad. Sci. USA* (2025).
163. Guthrie, S. E. Faces in the Clouds: A New Theory of Religion. *Oxford Univ. Press* (1995).
164. Jones, C. R. & Bergen, B. K. Large language models pass the Turing test. *arXiv* (2025).
165. Singhal, K. Toward expert-level medical question answering with large language models. *Nat. Med.* (2025).
166. Zerkouk, M., Mihoubi, M. & Chikhaoui, B. A comprehensive review of AI-based intelligent tutoring systems: applications and challenges. *arXiv* (2025).

167. Colomatto, C. & Fleming, S. M. Folk psychological attributions of consciousness to large language models. *Neurosci. Conscious.* (2024).
168. Aldahoul, N. Large language models are often politically extreme, usually ideologically inconsistent, and persuasive even in informational contexts. *arXiv* (2025).
169. Lawrence, H. R. The opportunities and risks of large language models in mental health. *JMIR Ment. Health* (2024).
170. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R. & Yang, D. The rise of AI companions: how human-chatbot relationships influence well-being. *arXiv* (2025).
171. Akbulut, C., Weidinger, L., Manzini, A., Gabriel, I. & Rieser, V. All too human? Mapping and mitigating the risk from anthropomorphic AI. *AAAI/ACM* (2024).
172. Shanahan, M. Talking about large language models. *Commun. ACM* (2024).
173. Zador, A. Catalyzing next-generation artificial intelligence through neuroAI. *Nat. Commun.* (2023).
174. Moser, E. I., Kropff, E. & Moser, M.-B. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.* (2008).
175. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* (2017).
176. Mirzadeh, S. I. GSM-symbolic: understanding the limitations of mathematical reasoning in large language models. *arXiv* (2024).
177. Kumaran, D., Hassabis, D. & McClelland, J. L. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* (2016).
178. Connell, L. & Lynott, D. What can language models tell us about human cognition? *Curr. Dir. Psychol. Sci.* (2024).

179. Martínez, L., Ruan, D. & Herrera, F. Computing with words in decision support systems: an overview on models and applications. *Int. J. Comput. Intell. Syst.* (2010).
180. Perlis, R. H., Goldberg, J. F., Ostacher, M. J. & Schneek, C. D. Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology* (2024).
181. Chiriatti, M., Ganapini, M., Panai, E., Ubiali, M. & Riva, G. The case for human–AI interaction as system 0 thinking. *Nat. Hum. Behav.* (2024).
182. Essel, H. B., Vlachopoulos, D., Essuman, A. B. & Amankwa, J. O. ChatGPT effects on cognitive skills of undergraduate students. *Comput. Educ. Artif. Intell.* (2024).
183. Rasmequan, S. & Russ, S. Cognitive artefacts for decision support. *IEEE SMC* (2000).
184. Vicente, L. & Matute, H. Humans inherit artificial intelligence biases. *Sci. Rep.* (2023).
185. Treiman, L. S., Ho, C.-J. & Kool, W. The consequences of AI training on human decision-making. *Proc. Natl Acad. Sci. USA* (2024).
186. Frentz, T. S. Memory, myth, and rhetoric in plato’s phaedrus. *Rhetor. Soc. Q.* (2006).
187. Heersmink, R. Extended mind and cognitive enhancement: moral aspects of cognitive artifacts. *Phenomenol. Cogn. Sci.* (2017).
188. Carr, N. The Shallows: What the Internet Is Doing to Our Brains. *W. W. Norton* (2020).
189. Tanil, C. T. & Yong, M. H. Mobile phones: the effect of its presence on learning and memory. *PLoS ONE* (2020).

190. Clemenson, G. D., Maselli, A., Fiannaca, A. J., Miller, A. & Gonzalez-Franco, M. Rethinking GPS navigation: creating cognitive maps through auditory clues. *Sci. Rep.* (2021).
191. Dahmani, L. & Bohbot, V. D. Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Sci. Rep.* (2020).
192. Bai, L., Liu, X. & Su, J. ChatGPT: the cognitive effects on learning and memory. *Brain-X* (2023).
193. Costello, T. H., Pennycook, G. & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science* (2024).
194. Heersmink, R. Use of large language models might affect our cognitive skills. *Nat. Hum. Behav.* (2024).