https://doi.org/10.1038/s44159-025-00456-8

Bringing comparative cognition approaches to AI systems

Konstantinos Voudouris, Lucy Cheke & Eric Schulz

Researchers are increasingly considering the cognitive capacities of artificial intelligence systems. Comparative cognition offers a helpful framework to avoid both overstating and understating these capacities.

Most contemporary scholars agree that cognition involves mechanisms, such as learning, memory and decision-making, that drive flexible and adaptive behaviour[1].The archetypal example of a system that exhibits these sorts of cognitive capacities is humans. Indeed, human cognition is often used as a reference against which to compare other systems in both scientific and nonscientific contexts. Throughout the history of psychology and the cognitive sciences, researchers have sought evidence of human-like cognitive capacities in other systems. In this process, the definition of cognition has been stretched and adapted, often in an effort to emancipate it from its inherent anthropocentrism and to instead understand the place of human cognition in the tapestry of complex behavioural systems. We interpret this practice of extending the tools of cognitive science to study nonhuman systems as an endorsement of the 'cognition thesis' – the idea that cognition can emerge from many structurally distinct systems, if they are arranged appropriately. It is the task of cognitive science to clarify the nature of that arrangement.

For over a century, work aligned with the cognition thesis has investigated evidence for cognition in nonhuman animals, particularly primates, birds and cetaceans (such as dolphins and whales)[2]. With the rise of cognitive science in the 1960s and 1970s, cognition and related cognitive terminology have begun to be applied to invertebrates, including bees, ants and cephalopods (such as octopus and cuttlefish)[3]. In the past decade there has been an explosion in the study of cognition in non-neural systems, including bacteria[4], plants[5] and protists (such as the slime mould)[6].

Artificial intelligence (AI) systems, including large language models (LLMs) and reinforcement learning agents, now exhibit behaviours that were once assumed to be exclusive to humans and other animals. It has been claimed that LLMs have human-like behaviour, and the same vocabulary used to describe human behaviour has also been used to describe theirs[7,8]. For subscribers to the cognition thesis, an intriguing prospect is emerging as they observe the rise of sophisticated AI: perhaps cognition could emerge not only from living biological systems but also from computers[9].

However, claims of AI cognition remain controversial, and there are debates over the validity of the experiments and the strength of the available data. Human-centric psychological tests often fail to accommodate the unique architectures and experiences of nonhuman subjects. This mismatch means that researchers might fail to attribute cognitive capacities to systems that have them, as well as prematurely attribute cognitive capacities where systems lack them. These debates are remarkably similar to the debates that surrounded the field of comparative cognition – the study of nonhuman animal behaviour – in its formative years in the early twentieth century. In response to these debates, animal psychologists developed numerous tools to evidence and justify their claims about animal minds[10]; we propose that researchers interested in AI cognition should learn lessons from comparative cognition to make progress. By adopting these methods, AI research could avoid known pitfalls, join the cognitive sciences and help to clarify the very nature of cognition itself.

Avoiding underattribution

Comparative cognition abounds with examples of experimental designs that initially obscured true cognitive capacities. Take the case of domestic dogs and object permanence – the ability to track hidden objects, a key feature of human visual cognition. Early studies concluded that dogs lacked this capacity because they performed poorly on the invisible displacement task, in which a reward is hidden in a movable container that is then moved to a new location that the dog cannot see. The reward is hidden in the new location and the movable container is shown to be empty. Dogs fail to robustly locate the reward in its new location in these scenarios, which suggests that they cannot track its movement while it is out of sight. However, later research revealed that the task itself was the problem: dogs get distracted by the container and track that instead of the reward[11]. Subsequent studies using several independent measures from other tasks now suggest that dogs do have object permanence[12].

Cognitive scientists investigating AI systems can avoid the pitfall of mistakenly underattributing cognitive capacities to these systems by paying attention to the validity of the tests they use. For instance, LLMs appear to struggle with arithmetic that involves large numbers, which invites the conclusion that they lack a fundamental component of human cognition – the ability to reason about abstract numerical quantities and combine them using operations such as addition or division. However, their apparent arithmetical inability often stems from tokenization – how LLMs process text. LLMs do not encode text the way that humans do (as individual characters collected into words on a page). Instead, they combine characters together into tokens that are represented as atomic symbols. The exact characters that go into these tokens are learned from data, with a preference for grouping together characters that commonly co-occur; this means that commonly occurring numbers such as '100' and '99' are seen as single tokens. This tokenization leads to problems when the model has to compute sums such as '100,100,100 + 999,999'. Using an alternative tokenization strategy, such as forcing the model to tokenize numbers from right to left, can greatly improve LLM performance[13], which suggests that LLMs can do arithmetic but are limited by the standard way that arithmetic questions are tokenized.

Just as comparative cognition researchers refine their experimental designs to avoid misinterpreting limitations in task performance as

cognitive deficits, AI researchers must carefully consider how factors such as tokenization, input format and task structure might obscure the abilities of an AI system. By systematically varying test conditions and identifying sources of failure, they can minimize the risk of prematurely dismissing capabilities that might be obscured by artefacts of the task design.

Avoiding overattribution

The case of Clever Hans, a horse in the 1900s who appeared to solve arithmetic problems, illustrates the danger of overattributed cognitive abilities[14]. Hans's owner, Wilhelm van Osten, would say or write out a sum or a numerical puzzle, and Hans would answer correctly with taps of his hoof. Although many observers at the time, both in and outside of science, were astonished by Clever Hans's apparent numerical acumen, it later transpired that his owner was giving him subtle and unconscious cues to the right answer, and therefore creating the illusion of mathematical reasoning. This parable reminds comparative cognition researchers to rigorously rule out alternative explanations for seemingly sophisticated behaviour by generating and testing alternative hypotheses.

AI research has its own Clever Hans moments. For instance, deep neural networks excel at identifying common objects from visual images and discriminate between millions of objects with an accuracy that sometimes surpasses that of humans, which leads to the claim that these models have a human-like object recognition capacity. However, their accuracy can unexpectedly collapse for images that are imperceptibly different to a human observer, which leaves claims of human-like behaviour in tatters. Instead, these models mostly rely only on superficial image features (such as textures or patterns), which leads to brittle, non-human-like forms of object recognition[15]. In this example, apparently sophisticated performance is driven by an unexpected and unintended mechanism that researchers initially overlooked, partly because of the complexity and human-like nature of the observed behaviour. Just as Clever Hans was not actually capable of arithmetic but instead responded to human body language, deep neural networks seem to perform sophisticated object recognition but actually rely on brittle, non-human-like heuristics.

To prevent overattribution, researchers must systematically consider and rule out alternative explanations. By designing tests that explicitly control for these alternative explanations, researchers can more accurately assess AI capabilities and avoid misattributing human-like cognition to systems that rely on fundamentally different mechanisms.

Towards a comparative science of AI

The rise of AI demands a new comparative science that draws from the science of animal cognition to investigate intelligence across both biological and artificial systems. This approach enables researchers to empirically test the cognition thesis: the idea that AI systems might share fundamental cognitive capacities with humans and other animals.

The comparative cognition approach to AI emphasizes the importance of considering alternative explanations before making claims about cognitive capacities. Researchers must rigorously design experiments to reveal the true cognitive capacities of these AI systems, without overestimating or underestimating them. Sometimes, failures on a task arise owing to the way it is presented, such as tokenization artefacts in language models, and successes might be driven by different mechanisms than those thought to drive human behaviour. These alternative explanations can be counterintuitive because they would not typically apply to human cognition. The key lesson is to step outside anthropomorphic assumptions and rigorously test all possible explanations for observed behaviour before attributing sophisticated cognitive abilities, or the lack thereof, to AI systems.

Individual researchers as well as the wider scientific community must scrutinize their experiments to determine whether they adequately control for alternative explanations and account for the constraints of the system under study. Only then can researchers truly compare cognition — explicitly evaluating the similarities and differences between AI systems, humans and other animals in terms of their cognitive capacities. By embracing this challenge, researchers can make progress on the questions of what cognition is and where it comes from.

Konstantinos Voudouris [1,2] ✉, **Lucy Cheke**[2,3] **& Eric Schulz**[1]
[1]Helmholtz Institute for Human-Centered AI, Helmholtz Munich, Munich, Germany. [2]Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK. [3]Department of Psychology, University of Cambridge, Cambridge, UK.
✉e-mail: k.voudouris14@googlemail.com

Published online: 13 May 2025

References

1. Bayne, T. et al. What is cognition? *Curr. Biol.* **29**, R608–R615 (2019).
2. Shettleworth, S. J. *Cognition, Evolution, and Behavior* (Oxford Univ. Press, 2009).
3. Perry, C. J., Barron, A. B. & Cheng, K. Invertebrate learning and cognition: relating phenomena to neural substrate. *Wiley Interdiscip. Rev. Cogn. Sci.* **4**, 561–582 (2013).
4. Lyon, P. The cognitive cell: bacterial behavior reconsidered. *Front. Microbiol.* **6**, 264 (2015).
5. Segundo-Ortin, M. & Calvo, P. Consciousness and cognition in plants. *Wiley Interdiscip. Rev. Cogn. Sci.* **13**, e1578 (2022).
6. Reid, C. R. Thoughts from the forest floor: a review of cognition in the slime mould *Physarum polycephalum*. *Anim. Cogn.* **26**, 1783–1797 (2023).
7. Hagendorff, T. et al. Machine psychology. Preprint at https://doi.org/10.48550/arXiv.2303.13988 (2024).
8. Shevlin, H. & Halina, M. Apply rich psychological terms in AI with care. *Nat. Mach. Intell.* **1**, 165–167 (2019).
9. Simon, H. A. Cognitive science: the newest science of the artificial. *Cogn. Sci.* **4**, 33–46 (1980).
10. Boakes, R. *From Darwin to Behaviourism: Psychology and the Minds of Animals* (CUP Archive, 1984).
11. Müller, C. A., Riemer, S., Range, F. & Huber, L. The use of a displacement device negatively affects the performance of dogs (*Canis familiaris*) in visible object displacement tasks. *J. Comp. Psychol.* **128**, 240–250 (2014).
12. Zentall, T. R. & Pattison, K. F. Now you see it, now you don't: object permanence in dogs. *Curr. Dir. Psychol. Sci.* **25**, 357–362 (2016).
13. Singh, A. K. & Strouse, D. J. Tokenization counts: the impact of tokenization on arithmetic in frontier LLMs. Preprint at https://doi.org/10.48550/arXiv.2402.14903 (2024).
14. Beran, M. J. Did you ever hear the one about the horse that could count? *Front. Psychol.* **3**, 357 (2012).
15. Ilyas, A. et al. Adversarial examples are not bugs, they are features. In *Adv. Neural Inf. Process. Syst.* 32 (eds Wallach, H. et al.) 125–136 (Curran Associates, 2019).

Competing interests

The authors declare no competing interests.