Trends in
Cognitive Sciences

CellPress

Review

Learning by thinking in natural and artificial minds

Tania Lombrozo [1,*]

Canonical cases of learning involve novel observations external to the mind, but learning can also occur through mental processes such as explaining to oneself, mental simulation, analogical comparison, and reasoning. Recent advances in artificial intelligence (AI) reveal that such learning is not restricted to human minds: artificial minds can also self-correct and arrive at new conclusions by engaging in processes of 'learning by thinking' (LbT). How can elements already in the mind generate new knowledge? This article aims to resolve this paradox, and in so doing highlights an important feature of natural and artificial minds – to navigate uncertain environments with variable goals, minds with limited resources must construct knowledge representations 'on demand'. LbT supports this construction.

Highlights

Cognitive scientists have developed many accounts of learning from observation, in which the learner gains new knowledge from observation of the external world.

A relatively neglected form of learning is 'learning by thinking' (LbT) in which learning occurs without such external input, as in thought experiments or explaining to oneself.

Such instances of learning present a paradox: in what sense can 'old' elements in the mind lead to 'new' knowledge?

Large language models and other forms of artificial intelligence (AI) exhibit many of the same capacities for LbT in human minds.

Across both natural and artificial minds, LbT reveals important lessons about how intelligent systems construct representations to support novel contexts and goals.

The paradox of knowledge from nowhere

Learning is central to cognitive science. A great deal of research has accordingly focused on 'learning from observation' – cases in which the learner obtains new information from their senses (e.g., a visual observation or a speaker's testimony) and updates their mental representations accordingly. What this focus excludes is less canonical instances of learning that occur within the mind – a phenomenon called 'learning by thinking' (LbT) [1]. Consider the scientist who achieves new insight through a thought experiment, the driver who discovers how to navigate an obstacle through mental simulation, or the writer who learns something new by trying to explain her ideas. In each case, learning occurs in the absence of novel input from the external world.

Recent evidence from psychology demonstrates the ubiquity of LbT: both children and adults learn through explanation, simulation, comparison, and more (e.g., [2–6]). However, the latest advances in computer science bring the phenomenon of LbT into starkest relief. Consider the following dialogue with GPT-4, a large language model (LLM):

USER: Is 420 60% of 700?

GPT-4: No, 420 is not 60% of 700.
To find 60% of 700:
0.60 × 700 = 420

Actually, I made a mistake in my initial response.
420 is indeed 60% of 700.

In the course of explaining, GPT-4 corrects a misconception to arrive at a correct conclusion – without any external feedback. Less dramatic illustrations of the same phenomenon include the

[1] Department of Psychology, Princeton University, Princeton, NJ 08540, USA

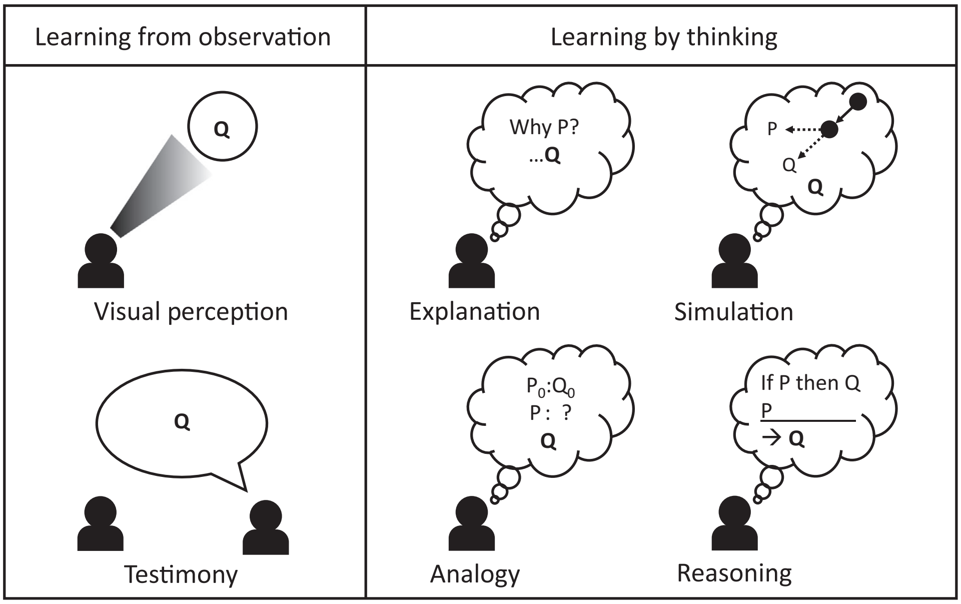
*Correspondence:
lombrozo@princeton.edu (T. Lombrozo).

effects of asking LLMs to 'think step by step' [7] or modeling a chain of thought [8], which can lead to more accurate responses without external correction.

LbT is a paradoxical phenomenon. In one sense, learners gain no new information: they have only the elements already in their minds to work with. In another sense, learning has occurred: the agent has acquired new knowledge (such as the answer to a mathematical problem) or new abilities (such as the capacity to answer a new question or draw a new inference). One aim of this article is to suggest a resolution to this paradox. Another is to highlight the parallels between LbT across natural and artificial minds, focusing on learning through explanation, simulation, comparison, and reasoning (Figure 1). Doing so reveals parallel computational problems and solutions across humans and AI: both systems make use of processes that re-represent existing information to support more reliable inferences. In so doing, LbT processes help resource-limited systems like us reach relevant conclusions 'on demand' rather than relying exclusively on the learning that occurs when observations are first made. These ideas are further developed after first reviewing evidence for LbT.

Four varieties of LbT

This section considers four examples of LbT: learning through explanation, simulation, analogy, and reasoning. This is not an exhaustive list of LbT processes – there are additional forms of LbT (such as learning through imagination [9,10]), and these forms of learning could be individuated more finely (e.g., learning through teleological explanation could differ from learning through mechanistic explanation [11]). The aim here is not to offer a fixed taxonomy of LbT processes, not least because these processes are likely to share many-to-many mappings to learning goals and to the underlying structures of attention, memory, and other cognitive mechanisms that instantiate them. Instead, these four processes help illustrate the ubiquity and power of LbT, including parallels across human minds and AI.



Trends in Cognitive Sciences

Figure 1. The varieties of learning. Schematic illustration of different forms of learning. Represented are two canonical forms of learning from observation, and four examples of learning by thinking (LbT). In each case the learner 'learns' the new proposition, Q.

Learning through explanation

In a now classic paper, researchers observed that 'good' students differed from 'poor' students in how they studied material: the good students spent more time explaining text or examples to themselves [12]. Subsequent work experimentally manipulated the extent to which students explained by offering prompts or instructions to explain, and found that engaging in explanation improved learning, and that the strongest benefits (relative to control groups) were for problems that required going beyond the material studied [13–15]. This phenomenon is known as the self-explanation effect, and it is likely to resonate with teachers, parents, and explainers of all kinds: most people have had the experience of coming to learn something better – or realizing they do not understand it as well as they thought they did – after trying to explain it [16].

When explaining occurs without external feedback, it offers an opportunity for pure LbT. Such learning can be broadly classified as corrective or generative: in corrective cases, learners recognize (and therefore represent) flaws in existing representations; in generative cases, learners construct entirely new representations. As an example of corrective learning, consider a phenomenon known as the illusion of explanatory depth (IOED) – people tend to overestimate how well they understand the workings of devices such as microwaves or zippers, but better appreciate the limits of their understanding after attempting to explain [17]. As an example of generative learning, consider studies in which participants are prompted to explain in the course of learning new categories. Relative to those in control groups, those who explain are more likely to generate abstract representations [6,18,19] and discover broad patterns in study examples [18,20–23].

Research in AI has employed 'self-explanations' in somewhat analogous ways. Most relevant to LbT are cases where the AI system benefits from generating an explanation itself. One example from machine learning is known as 'explanation-based learning', which involves generating explanations for training examples to construct generalizations from limited data [24,25]. Another example comes from more contemporary work in which deep reinforcement learning agents learn to predict either the solutions to various tasks, or those same solutions along with natural language explanations that accompany them [26]. For tasks requiring relational and causal reasoning, agents that learned to predict explanations outperformed those that did not, and even those that received explanations as inputs (versus targets for prediction). In particular, the agents that predicted explanations were less likely to over-rely on 'easy' features and were more likely to draw appropriate generalizations from confounded data. In both humans and AI, explaining seems to support the construction of more generalizable representations.

Learning through simulation

Imagine three interlocking gears arranged horizontally. The gear on the left is rotating clockwise. What direction is the right-most gear rotating? Most people solve this problem by engaging in mental simulation [27]: they create a picture in their mind of the three gears, put the left-most gear in a clockwise motion, and 'observe' the movement of the remaining two gears. More dramatic illustrations of the power of mental simulations come from thought experiments in the history of science: Einstein drew lessons about relativity by mentally simulating travel on photons and trains [28]; Galileo drew conclusions about gravity by mentally simulating the behavior of falling objects [29]. As with other forms of LbT, mental simulations and thought experiments can offer new insights in the absence of novel data external to the mind.

Like explaining, mental simulations can be corrective (used to recognize flaws) or generative (used to construct novel representations [30]). Illustrating corrective mental simulation, participants in one study were prompted with thought experiments that were expected to elicit Newtonian intuitions about force and motion, whereas other thought experiments were expected to elicit misconceptions associated with an impetus theory [31]. For many participants, these thought experiments successfully elicited inconsistent intuitions. On a subsequent assessment, participants were less likely to endorse impetus judgments, presumably because the thought experiments led them to recognize and attempt to correct what had previously been a latent inconsistency. Illustrating a generative role for mental simulation, research on causal reasoning suggests that, in making judgments about whether one event caused another, people engage in a counterfactual simulation of what would have happened had the first event not occurred [5,32]. For example, given a visual display in which one ball bumps into another, changing its trajectory such that it reaches a target that it otherwise would have missed, participant eye movements suggest that they counterfactually simulate the trajectory that the second ball would have followed had it not been bumped, and judge the first ball to have caused the second to reach the target on this basis.

Many contemporary AI systems learn through some form of simulation [33], although simulation has been a feature of AI models for some time [34]. For example, model-based methods in deep reinforcement learning can use representations of the environment to generate data that are used to train a model-free action policy [35–37]. This is analogous to drawing on an intuitive theory for mental simulation, and subsequently learning from the simulated data (potentially using a different process such as associative learning or inductive inference). Another example comes from algorithms for selecting the best course of action from a set of nested decisions (as in a decision tree): one approach is to engage in 'deep imagination', or simulation of a few long sequences of decisions, as a basis for approximating optimal solutions given limited resources [38]. Across both humans and AI, simulation can offer 'data' that serve as the input to various mechanisms for learning and reasoning.

Learning through analogical reasoning and comparison

In developing his theory of natural selection, Charles Darwin noticed a potential analogy between selective breeding and biological evolution. Based on this analogy, he compared the mechanisms of change in each case and drew new inferences about natural selection: in the same way as domestication could result in accidental variation, so too accidental variation might arise in natural selection [39]. Many such instances of analogical reasoning and comparison have been documented in the history of science (e.g., [40]) and in scientific problem solving [41]. When a reasoner already has knowledge concerning the two elements that enter into some comparison or analogy, comparison or analogical inference can be the basis for novel conclusions or insights, thus supporting LbT.

Most experimental investigations of comparison and analogical reasoning do not involve pure LbT. Participants are not solely prompted to engage in analogical thinking but are typically provided with a particular analogy or exemplars. In such cases, subsequent learning reflects both the information provided by the researchers (about which analogies or exemplars to consider) and the analogical thinking itself. However, some studies come close to isolating the effects of comparison or analogical thinking: in these studies, all participants receive the analogy or exemplars, but only some participants are cued to consider the analogy or exemplars in solving a new problem (e.g., [42]) or are prompted to explicitly engage in comparison across cases (e.g., [43,44]). These studies find both corrective and generative effects of engaging in these forms of thinking. For example, a study of mathematical learning found that participants who received more cues to compare sample solutions were less susceptible to misleading surface similarity (i.e., the topic of a word problem) in deciding how to solve new problems [45]. Illustrating generative effects, one study asked participants to identify the similarities and differences between two groups of robots [4]. Those who engaged in this comparison were significantly more likely than those in a control condition to discover a subtle rule that differentiated all members of the categories, potentially because they engaged in alignment and abstraction that helped them re-represent the features of the robots [46].

Analogical reasoning and inference have also been the subject of interest in AI (e.g., [47–49]). As with human experiments, most demonstrations of analogical reasoning do not involve pure LbT: more often, AI systems are asked to solve analogical problems that involve source analogies (e.g., [50]). However, some recent applications do involve prompting systems to engage in analogical reasoning or comparison without providing the source analogies themselves. For example, in ‘analogical prompting’, LLMs are prompted to self-generate exemplars that can then be used in solving new problems [51]. Across a range of mathematical questions, code-generation tasks, and other reasoning problems, the most effective prompts were those that requested the LLM to generate three to five relevant but diverse exemplars: each was described and the solution was explained before the LLM offered a solution to the new problem. Using this approach (which perhaps combines the benefits of both analogical reasoning and explanation), analogical prompting out-performed a variety of state-of-the-art benchmarks for LLM performance (although the effects were not always large). At a coarse grain, this mimics the effects of comparison prompts observed with both adults and children, although it remains unclear whether the benefits for LLMs resulted from processes that correspond to those in humans.

Learning through reasoning

I might know that today is Wednesday, and that on Wednesdays I should avoid parking in a particular campus lot. However, I might not realize that I should avoid parking in that campus lot today. Mundane examples such as this illustrate that even trivial inferences can require attending to the right information and processing it in exactly the right way. That is, despite representing 'P' and 'if P then Q', drawing the conclusion 'Q' (a simple instance of *modus ponens*) can require some form of reasoning – the construction or evaluation of an argument, including the reasons that support a conclusion ([52]; also [53]). When the conclusion is successfully drawn ('Q'), it can feel like a new insight. If this example of *modus ponens* is not compelling, consider more complex instances of deductive or inductive reasoning. For example, from the premises that 'everyone loves anyone who loves someone' and that 'my neighbor Sarah loves Taylor Swift', it follows (deductively) that 'Donald Trump loves Kamala Harris' – but this is unlikely to be immediately apparent [54].

In some cases, simply reasoning, reasoning more, or reasoning better can serve as a corrective (*cf* [55]). For example, one study asked participants to evaluate arguments about the solutions to reasoning problems [56]. Unknown to the participants, the arguments they were evaluating were in some cases their own that were generated in earlier trials of the same study. About two-thirds of the time, participants were able to correctly reject their own previous invalid reasoning (and did so more often than they rejected their own previous valid reasoning). Of course, reasoning can also be generative: for instance, participants who spend more time reasoning about the answer to tricky problems from the 'cognitive reflection test' (CRT), such as the 'bat and ball problem', are more likely to generate the correct response ([57]; also [58]).

Within AI, various forms of reasoning have been realized in traditional symbolic architectures that implement explicit rules [59] and/or implement probabilistic computation (e.g., [60]). A newer development is the emergence or elicitation of reasoning (or behavior that ‘looks’ like reasoning) in deep learning systems such as LLMs [61] (Box 1). For instance, prompting LLMs to explicitly engage in step-by-step reasoning can lead them to accurate solutions that they fail to reach

Box 1. Learning by thinking in large language models

LLMs such as OpenAI's ChatGPT are deep neural networks trained on vast quantities of text to generate next-word predictions. Despite many impressive capabilities, LLMs currently fall short of human-level performance on basic reasoning tasks such as logic and mathematics problems [91]. These shortcomings have spurred a new wave of efforts to improve LLM reasoning capabilities through 'prompt engineering' – designing inputs to LLMs that increase the probability of eliciting desired outputs. Chain-of-thought prompting, an approach that holds particular promise for improving reasoning, involves LbT.

In chain-of-thought prompting [8], LLMs are prompted with a sample input that includes a series of intermediate reasoning steps. Performance is compared with 'standard' prompts in which the LLM receives a sample input but without those intermediate steps. For instance, consider the following example of input in a standard prompt:

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

With chain-of-thought prompting, the modeled answer instead includes a chain of thought (indicated in bold):

A: **Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11.**
The answer is 11.

When so prompted, LLMs perform considerably better on mathematical word problems, commonsense reasoning, and other tasks.

Why is prompting LLMs to produce intermediate reasoning steps effective? One source of insight comes from comparing the effects of 'direct prompting' to 'thinking step-by-step' (i.e., with intermediate steps) in models trained on data with different statistical structures ([7]; *cf* [92]). This approach finds – both analytically and experimentally – that thinking step-by-step is effective when the training data involve local clusters of related variables, such that variables that were not directly observed in training can be chained together.

This suggests that chain-of-thought reasoning exploits the accuracy of 'local' inferences to generate better guesses about distant connections. More precisely, when outputs are conditioned on intermediate inputs that raise the conditional probability of a correct response, the result is more accurate reasoning. This provides a concrete model for how LbT can operate in a statistical next-word prediction system – a distinctly 'un'human kind of mind [93] that may only mimic human-like reasoning [94,95]. Nonetheless, LLM chain-of-thought prompting and human learning through reasoning converge in their role for LbT: in both cases, the learner benefits from generating intermediate inputs to thinking that offer a more reliable path to the desired output.

with direct prompts to simply report the solution [8]. The benefits of such step-by-step reasoning (vs. direct prompts) are larger for more complex problems [8], and become greater still when LLMs are prompted to ask themselves questions corresponding to intermediate steps in reaching problem solutions [62]. Although the effects of step-by-step reasoning likely arise from multiple sources (not all of which are likely to have analogs in human minds), one is 'locality' [7] – when systems perform long-range inferences that require connecting pieces of information that have not been observed together in training, a direct inference will introduce greater bias. When systems instead chain together more local inferences, the result is more accurate conclusions. Systems can also improve themselves by learning from their own chain-of-thought rationales [63], a type of self-instruction that has long been recognized in humans [64].

Dissolving the paradox of LbT

We have now seen several instances of LbT across natural and artificial minds. While the operative mechanisms across cases and types of minds surely differ, the puzzle of LbT is shared. How can 'thinking' be sufficient for 'learning'?

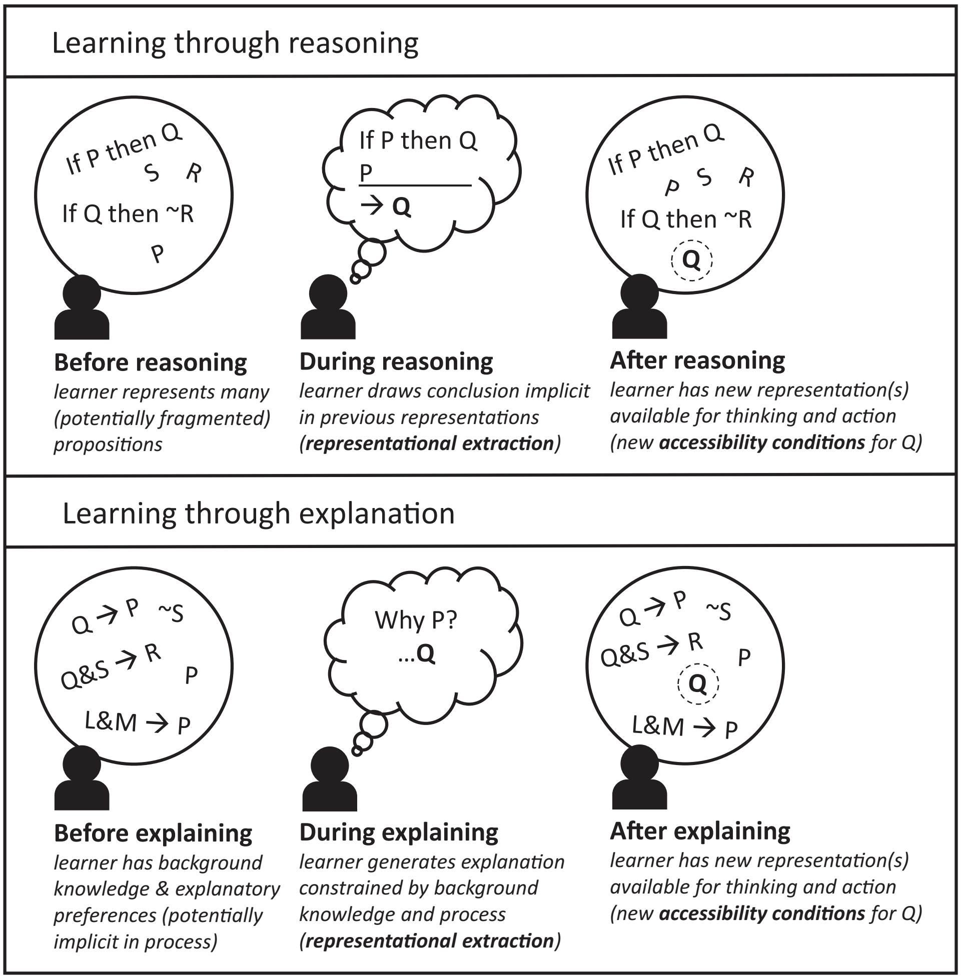
The poet and playwright Heinrich von Kleist argued for the value of 'learning by speaking' – that through the process of articulating our thinking to others, we can lubricate the fabrication of ideas in 'the workshop of reason' [65]. He also identified the contours for dissolving the paradox of LbT – 'For it is not we who know, but at first it is only a *certain state of mind* of ours that knows' (p. 45). On this view, the ‘learning’ in LbT comes from making some knowledge newly accessible to the learner (the ‘we’). The reason thinking is sufficient – without external input – is because the basis for that knowledge was already in the mind (or in a 'certain state of mind').

We can unpack these ideas in more cognitive terms using the case of learning through reasoning. In the simplest case, the learner combines two premises to support a new conclusion. Although the conclusion was already deductively implied or inductively supported by the premises, explicitly recognizing the conclusion is a cognitive accomplishment: the process of reasoning yields a representation with affordances that differ from those of the premises in isolation [1,66–68]. For example, when I recognize that ‘today’ is the day I should avoid parking in a particular campus lot, I am in a better position to direct my actions ('turn left here, not right'), draw further inferences ('I will need to leave 10 minutes to walk back to my car'), and share information with others ('park in the east lot!'). Moreover, a process such as reasoning can support learning in a factive sense (i.e., where ‘learning’ that Q implies that Q is true). This is because deductive and inductive reasoning are truth-conducive: if the premises are true, the conclusion is guaranteed to be true (for deduction) or likely to be true (for induction) ([69,70] for more nuanced discussions of deduction and induction).

Recent work has developed these ideas for learning through explanation [1] and simulation [71], respectively. In these cases, it is less clear what functions as a 'premise' in thinking, and it is less obvious why the variety of thinking involved would be conducive to learning. An important observation is that different representations have different affordances. These can be characterized in terms of the ‘accessibility conditions’ of a representation ([67,68]. The basic idea is that cognitive processes operate more or less effectively or efficiently given different types of inputs, much in the same way as algorithms for addition will operate differently (if at all) depending on whether they receive input in the form of Arabic or Roman numerals. As a result, a representational change can have dramatic consequences – for example, once a number in Roman numerals is re-represented in Arabic numerals, different features become more apparent (e.g., whether a number is even or odd) and different algorithms become available. Processes such as explanation, comparison, and simulation can change the accessibility conditions for representations and thus support ‘representational extraction’ – the creation of new representations with new accessibility conditions.

When representational extraction occurs, information that was represented in a proprietary format or embodied in a process can function as a premise that constrains the output of an LbT process. To make this more concrete, consider the case of explanation. Some research suggests that, when learners engage in explanation, they favor causal structures with fewer 'root causes' (i.e., causes that are themselves unexplained [72]; also [73–77]), such that they will treat some causal hypotheses as preferable or more probable than others. A proposition such as 'explanations with fewer root causes are more probable' thus functions as a premise would in reasoning, but not one that the explainer would necessarily recognize or exploit in explicit reasoning [1,29,72,78]. When the process of explaining yields some output (e.g., a preference for one explanation over another), the learner gains access to a conclusion that was constrained by this implicit premise. This process is illustrated in Figure 2, along with an example of learning through reasoning.

Recognizing the role of accessibility conditions and representational extraction allows us to generalize the simple story of learning through reasoning to less transparent processes such as explanation, comparison, and simulation, among others (Box 2). However, why should


Learning through reasoning
If P then Q
S
R
If Q then ~R
P
If P then Q
P
→ Q
If P then Q
P
S
R
If Q then ~R
Q
Before reasoning
learner represents many (potentially fragmented) propositions
During reasoning
learner draws conclusion implicit in previous representations (**representational extraction**)
After reasoning
learner has new representation(s) available for thinking and action (new **accessibility conditions** for Q)
Learning through explanation
Q → P
~S
Q&S → R
P
L&M → P
Why P?
…Q
Q → P
~S
Q&S → R
P
Q
L&M → P
Before explaining
learner has background knowledge & explanatory preferences (potentially implicit in process)
During explaining
learner generates explanation constrained by background knowledge and process (**representational extraction**)
After explaining
learner has new representation(s) available for thinking and action (new **accessibility conditions** for Q)

Trends in Cognitive Sciences

Figure 2. Learning by thinking (LbT) in action. Two examples of LbT, schematically representing the role of representational extraction and its consequences for the accessibility conditions of representations. Note that LbT processes can change multiple representations; for example, explaining might not only result in the explicit representation of some explanation, Q, but also in the target of that explanation being represented at a different level of abstraction [6,18].

we expect the output of these processes to yield new 'knowledge' (i.e., learning in a factive sense)? Of course, there is no guarantee that the conclusions reached through explanation, simulation, or other LbT processes will be true – indeed, there is no guarantee that the conclusions of inductive reasoning will be true, nor even those of deductive reasoning when the premises are false. However, to the extent that these processes have been shaped by evolution, experience, or (in the case of artificial minds) intentional design, we might expect them to at least partially track the structure of the world, and therefore produce somewhat reliable outputs [79]. More modestly, even when these outputs are not strictly accurate, they may still prove useful for guiding thinking and action [80]. One example of this comes from the process of learning through explanation: even when the explanations generated are false, the process of explaining can sometimes improve subsequent inquiry and later judgments (e.g., by leading learners to recognize conflicts between representations, or to represent the domain in more abstract terms [6,13,81]).

Box 2. Rationalization as learning by thinking

Humans often engage in the peculiar activity of rationalization: generating explanations for their behavior that would make that behavior rational in light of their beliefs and desires. Yet, decades of research suggest that such explanations are often unfaithful in the sense that they fail to correspond to the causal mechanisms that actually generated the behavior in question (e.g., [96,97]). Why then do humans engage in rationalization?

One recent proposal is that it is rational to pursue rationalization: when we generate beliefs and desires that make sense of adaptive behaviors generated by introspectively opaque mechanisms, we engage in a form of 'representational exchange' that extracts useful information from those otherwise opaque mechanisms, rendering that information available to reasoning [66]. On this view, representational exchange is 'the process of translating information from one psychological system, or representational format, into another' (p. 9), and can thus be a basis for LbT.

To illustrate, consider a child who carefully checks inside their shoe before putting it on. The mechanism generating this behavior could be simple imitation – copying what they have seen a parent or sibling do. However, when someone asks them why they are checking inside their shoe (or they ask themselves), they generate a plausible response: they are checking for spiders. If this is in fact the reason that others in their community check their shoes, then rationalization has led the child to learn something new.

Underlying this approach is the idea that many 'non-rational' processes, such as instinct, conformity to social norms, and habits, are adaptive: because they are shaped by adaptive processes (biological evolution, cultural evolution, and reinforcement learning, respectively), they contain implicit beliefs that may be justified, but are unavailable to other cognitive systems. For this reason representational exchange can result in new knowledge or learning – not merely the acquisition of new (potentially false or unjustified) beliefs.

Given that different forms of thinking will shape representations in different ways, figuring out when to rely on such processes is an important meta-reasoning problem that faces natural and artificial minds alike (e.g., [82,83]). For example, a scientist might learn to discern whether and when to trust the outputs of mental simulations that rely on intuitive physics [84,85], and a robot might evaluate whether to continue reasoning or engage in action [86]. That said, the trade-offs faced by humans and AI are likely to differ in important ways, with implications for the conditions under which pursuing each process is likely to be reliable and rational. For example, a system with extremely reliable storage and retrieval might engage in simple recall whereas a system with nosier retrieval will engage in simulation or reasoning.

So far we have considered how thinking can be 'sufficient' for learning (the inputs to learning are already in the mind of the learner), and how such thinking might result in 'learning' (LbT processes are themselves shaped by evolution, learning, and/or design). Across both natural and artificial minds, LbT allows agents to capitalize on intermediate processes through which they can arrive at more reliable conclusions. However, we are left with the question of when and why LbT is 'necessary'. Would an ideal learner already represent all conclusions upon learning the relevant premises, rather than needing to think their way to new conclusions as the circumstance demands? A computational analogy is helpful once again. Within an artificial system, limitations on memory and processing time will dictate how much anticipatory computation can occur. Since the implications of current beliefs are boundless, the propagation of error detrimental, and future needs uncertain, a system needs to be selective in what (and how much) it concludes from either observation or inference. LbT offers a way to generate new and useful representations 'on demand' [87,88] rather than relying exclusively on learning that has already occurred. We should therefore expect LbT to be especially common for intelligent agents that (i) face limitations in terms of time, processing, or other resources [89,90]; and (ii) have uncertainty concerning their future contexts and goals. In other words, we should expect LbT processes to be especially pervasive for creatures like us.

These observations generate predictions about the conditions under which natural and artificial minds might diverge when it comes to the role of LbT. As artificial minds overcome the resource limitations faced by human minds, or when artificial minds face problems involving lower uncertainty concerning future contexts and goals (perhaps because they operate within a very narrow domain), we should expect greater departures between natural and artificial minds in the role of LbT.

Concluding remarks

LbT is ubiquitous: humans learn not only through observation but also through explanation, comparison, simulation, reasoning, and beyond. Recent developments reveal that AI systems can also learn in these ways. In both cases we can resolve the paradox of LbT by recognizing that representations can vary in their accessibility conditions; through LbT, representations with novel accessibility conditions can be extracted and put to use to yield new knowledge and abilities.

In one sense, LbT reflects cognitive limitations: a system with unlimited resources and limited uncertainty would be able to work out the consequences of observations as they occur. By contrast, natural and artificial minds face limited resources and considerable uncertainty about what will be relevant to future judgments and decisions. In such cases, LbT offers a way to support 'on demand' learning that capitalizes on the strengths of existing representations (Box 1) in the context of the agent's current situation and goals. However, many questions remain open about how LbT processes are implemented in natural and artificial minds, including how they contribute to human intelligence (Box 3) and when they might lead us astray (see Outstanding questions). Learning the answers to these questions will ultimately require more than merely thinking – it will take the full, interdisciplinary toolkit of cognitive science.

Box 3. Is learning by thinking uniquely human?

Annette Karmiloff-Smith introduced the important idea of ‘representational redescription’, which she characterized as 'basically a hypothesis about the specifically human capacity to enrich itself from within, by exploiting knowledge already stored rather than by simply exploiting the human and physical environment' ([98], p. 706; also [99]). In the vocabulary of this article, representational redescription offers a basis for LbT.

Karmiloff-Smith posited multiple levels of knowledge representations, ranging from those contained implicitly within procedural knowledge (level I), to representations of that procedural knowledge that can be used as data to the system (level E1), to consciously accessible representations (level E2), and finally to representations that are available in a domain-general code that supports verbal report (level E3). The process of representational redescription involves re-representing knowledge representations at increasingly higher levels, leading to greater representational manipulability and flexibility. As children 'spontaneously seek to understand their own cognition' ([98], p. 706), they engage in representational redescription and obtain the representational sophistication that ultimately supports intuitive theories of the world.

Karmiloff-Smith also speculated that representational redescription is uniquely human (while recognizing the limitations of extant data): 'in the human, internal representations become objects of cognitive manipulation such that the mind extends well beyond its environment and is capable of creativity' ([98], p. 706).

As a hypothesis about what differentiates humans from non-human animals, representational redescription (and LbT more generally) has two attractive features. First, the capacity for representational redescription admits of degrees – for example, it could be that rats engage in mental simulations that generate representations at Karmiloff-Smith's level E1, and that some primates can go beyond E1 but without achieving the perhaps uniquely human level E3 (that supports verbal report). Second, representational redescription arguably subsumes other common claims about human uniqueness. For example, language is often touted as uniquely human and corresponds to level E3, the level at which we see domain-general representations that readily translate to linguistic encoding. Another hypothesis is that humans are unique in their capacity for relational or analogical reasoning [100], an ability that requires the type of abstraction that begins to emerge at level E1.

Although Karmiloff-Smith focused on the contrast between human and non-human animals, she also entertained the possibility that representational redescription could be captured in connectionist networks. She ultimately concluded that the connectionist networks of the 1980s and 1990s fell short in a number of ways, but her analysis raised the possibility that representational redescription might be realized in future forms of AI.

Outstanding questions

When are LbT processes most likely to support reliable conclusions, and when do they lead learners astray? To what extent are these conditions for reliability similar or different across types of LbT (explanation, simulation, etc.) and across natural and artificial minds?

How do humans and AI systems solve the meta-reasoning problems of deciding when to engage in LbT versus other forms of activity (such as exploration) and of deciding which LbT process to engage?

How do humans and AI systems learn to balance the costs and benefits of immediate computation (e.g., reasoning about the consequences of a new observation the moment it is made) versus those of engaging in later LbT?

What can we learn about how to improve prompts for LLMs from research in psychology on LbT? What can the success of LLM prompting tell us about how to engage in 'prompt engineering' for humans?

How do learning through observation and LbT work together to support intelligent behavior? What can the study of LbT teach us about the role of mental processing in shaping learning in canonical cases of learning from observation? To what extent does LbT involve the same mechanisms as learning from observation, albeit with internally generated observations such as thoughts and mental imagery?

Acknowledgments

I am grateful to Andre Beukers and Tom Griffiths for comments on earlier drafts of this manuscript, and to Agnes Callard, Noah Goodman, and Joseph Sommer for pointing me to relevant references.

Declaration of interests

The author declares no competing interests.

References

1. Lombrozo, T. (2019) 'Learning by thinking' in science and in everyday life. In *The Scientific Imagination* (Levy, A. and Godfrey-Smith, P., eds), pp. 230–249, Oxford University Press
2. Brockbank, E. and Walker, C.M. (2022) Explanation impacts hypothesis generation, but not evaluation, during learning. *Cognition* 225, 105100
3. Brockbank, E. *et al.* (2023) Ask me why, don't tell me why: asking children for explanations facilitates relational thinking. *Dev. Sci.* 26, e13274
4. Edwards, B.J. *et al.* (2019) Explanation recruits comparison in a category-learning task. *Cognition* 185, 21–38
5. Gerstenberg, T. *et al.* (2021) A counterfactual simulation model of causal judgments for physical events. *Psychol. Rev.* 128, 936
6. Ruggeri, A. *et al.* (2019) Effects of explanation on children's question asking. *Cognition* 191, 103966
7. Prystawski, B. *et al.* (2024) Why think step by step? Reasoning emerges from the locality of experience. *Adv. Neural Inf. Proces. Syst.* 36, 70926–70947
8. Wei, J. *et al.* (2022) Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Proces. Syst.* 35, 24824–24837
9. Kind, A. (2018) How imagination gives rise to knowledge. In *Perceptual Imagination and Perceptual Memory* (Macpherson, F. and Dorsch, F., eds), pp. 227–246, Oxford University Press
10. Mguidich, H. *et al.* (2023) Does imagination enhance learning? A systematic review and meta-analysis. *Eur. J. Psychol. Educ.*, Published online October 25, 2023. https://doi.org/10.1007/s10212-023-00754-w
11. Lombrozo, T. and Wilkenfeld, D.A. (2019) Mechanistic versus functional understanding. In *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology* (Grimm, S.R., ed.), pp. 209–229, Oxford University Press
12. Chi, M.T. *et al.* (1989) Self-explanations: how students study and use examples in learning to solve problems. *Cogn. Sci.* 13, 145–182
13. Chi, M.T. *et al.* (1994) Eliciting self-explanations improves understanding. *Cogn. Sci.* 18, 439–477
14. Lombrozo, T. (2006) The structure and function of explanations. *Trends Cogn. Sci.* 10, 464–470
15. Lombrozo, T. (2016) Explanatory preferences shape learning and inference. *Trends Cogn. Sci.* 20, 748–759
16. Fonseca, B.A. and Chi, M.T. (2011) Instruction based on self-explanation. In *Handbook of Research on Learning and Instruction* (Mayer, R.E. and Alexander, P.A., eds), pp. 310–335
17. Rozenblit, L. and Keil, F. (2002) The misunderstood limits of folk science: an illusion of explanatory depth. *Cogn. Sci.* 26, 521–562
18. Williams, J.J. and Lombrozo, T. (2010) The role of explanation in discovery and generalization: evidence from category learning. *Cogn. Sci.* 34, 776–806
19. Walker, C.M. and Lombrozo, T. (2017) Explaining the moral of the story. *Cognition* 167, 266–281
20. Williams, J.J. and Lombrozo, T. (2013) Explanation and prior knowledge interact to guide learning. *Cogn. Psychol.* 66, 55–84
21. Williams, J.J. *et al.* (2013) The hazards of explanation: overgeneralization in the face of exceptions. *J. Exp. Psychol. Gen.* 142, 1006
22. Walker, C.M. *et al.* (2017) Explaining constrains causal learning in childhood. *Child Dev.* 88, 229–246
23. Kon, E. and Lombrozo, T. (2019) Scientific discovery and the human drive to explain. In *Advances in Experimental Philosophy of Science* (Wilkenfeld, D.A. and Samuels, R., eds), pp. 15–40, Bloomsbury Academic
24. DeJong, G. and Mooney, R. (1986) Explanation-based learning: an alternative view. *Mach. Learn.* 1, 145–176
25. Mitchell, T.M. *et al.* (1986) Explanation-based generalization: a unifying view. *Mach. Learn.* 1, 47–80
26. Lampinen, A.K. *et al.* (2022) Tell me why! Explanations support learning relational and causal structure. In *International Conference on Machine Learning*, pp. 11868–11890, ICML
27. Hegarty, M. (2004) Mechanical reasoning by mental simulation. *Trends Cogn. Sci.* 8, 280–285
28. Norton, J. (1991) Thought experiments in Einstein's work. In *Thought Experiments in Science and Philosophy* (Horowitz, T. and Massey, G.J., eds), pp. 129–144, Rowman & Littlefield
29. Gendler, T.S. (1998) Galileo and the indispensability of scientific thought experiment. *Br. J. Philos. Sci.* 49, 397–424
30. Brown, J.R. (1986) Thought experiments since the scientific revolution. *Int. Stud. Philos. Sci.* 1, 1–15
31. Bascandziev, I. (2024) Thought experiments as an error detection and correction tool. *Cogn. Sci.* 48, e13401
32. Gerstenberg, T. *et al.* (2017) Eye-tracking causality. *Psychol. Sci.* 28, 1731–1744
33. Hamrick, J.B. (2019) Analogues of mental simulation and imagination in deep learning. *Curr. Opin. Behav. Sci.* 29, 8–16
34. Yip, K. (1991) *KAM: A System for Intelligently Guiding Numerical Experimentation*, MIT Press
35. Gershman, S.J. *et al.* (2017) Imaginative reinforcement learning: Computational principles and neural mechanisms. *J. Cogn. Neurosci.* 29, 2103–2113
36. Mattar, M.G. and Daw, N.D. (2018) Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* 21, 1609–1617
37. Sutton, R.S. (1990) Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine Learning Proceedings 1990*, pp. 216–224, Morgan Kaufmann
38. Mastrogiuseppe, C. and Moreno-Bote, R. (2022) Deep imagination is a close to optimal policy for planning in large decision trees under limited resources. *Sci. Rep.* 12, 10411
39. Millman, A.B. and Smith, C.L. (1997) Darwin's use of analogical reasoning in theory construction. *Metaphor. Symb.* 12, 159–187
40. Nersessian, N.J. (1992) How do scientists think? Capturing the dynamics of conceptual change in science. In *Cognitive Models of Science (Minnesota Studies in the Philosophy of Science Vol. 15)* (Giere, R.N., ed.), pp. 3–44, University of Minnesota
41. Clement, J.J. (2009) The role of imagistic simulation in scientific thought experiments. *Top. Cogn. Sci.* 1, 686–710
42. Gick, M.L. and Holyoak, K.J. (1980) Analogical problem solving. *Cogn. Psychol.* 12, 306–355
43. Loewenstein, J. *et al.* (2003) Analogical learning in negotiation teams: comparing cases promotes learning and transfer. *Acad. Manag. Learn. Educ.* 2, 119–127
44. Rittle-Johnson, B. and Star, J.R. (2007) Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *J. Educ. Psychol.* 99, 561
45. Richland, L.E. and McDonough, I.M. (2010) Learning by analogy: discriminating between potential analogs. *Contemp. Educ. Psychol.* 35, 28–43
46. Gentner, D. and Maravilla, F. (2017) Analogical reasoning. In *International Handbook of Thinking and Reasoning* (Ball, L.J. and Thompson, V.A., eds), pp. 186–203
47. Forbus, K.D. *et al.* (2017) Extending SME to handle large-scale cognitive modeling. *Cogn. Sci.* 41, 1152–1201

48. Gentner, D. and Forbus, K.D. (2011) Computational models of analogy. *Wiley Interdiscip. Rev. Cogn.* 2, 266–276
49. Mitchell, M. (2021) Abstraction and analogy-making in artificial intelligence. *Ann. N. Y. Acad. Sci.* 1505, 79–101
50. Webb, T. *et al.* (2023) Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* 7, 1526–1541
51. Yasunaga, M. *et al.* (2023) Large language models as analogical reasoners. *ArXiv*, Published online October 3, 2023. http://dx.doi.org/10.48550/arXiv.2310.01714
52. Mercier, H. and Sperber, D. (2011) Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* 34, 57–74
53. Markman, A.B. and Gentner, D. (2001) Thinking. *Annu. Rev. Psychol.* 52, 223–247
54. Cherubini, P. and Johnson-Laird, P.N. (2004) Does everyone love everyone? The psychology of iterative reasoning. *Think. Reason.* 10, 31–53
55. Tesser, A. (1978) Self-generated attitude change. *Adv. Exp. Soc. Psychol.* 11, 289–338
56. Trouche, E. *et al.* (2016) The selective laziness of reasoning. *Cogn. Sci.* 40, 2122–2136
57. Meyer, A. and Frederick, S. (2023) The formation and revision of intuitions. *Cognition* 240, 105380
58. Di Stefano, G. *et al.* (2014) *Learning by Thinking: How Reflection Aids Performance (Working Paper 14-093)*, Harvard Business School
59. Russell, S.J. and Norvig, P. (2016) *Artificial Intelligence: A Modern Approach*, Pearson
60. Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann
61. Huang, J. and Chang, K.C.C. (2023) Towards reasoning in large language models: a survey. In *61st Annual Meeting of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, ACL
62. Press, O. *et al.* (2023) Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, ACL
63. Zelikman, E. *et al.* (2022) Star: bootstrapping reasoning with reasoning. *Adv. Neural Inf. Proces. Syst.* 35, 15476–15488
64. Simon, H.A. (1980) Problem solving and education. In *Problem Solving and Education: Issues in Teaching and Research* (Tuma, D.T. and Reif, F., eds), pp. 81–96, Lawrence Erlbaum Associates
65. Von Kleist, H. (1951) On the gradual construction of thoughts during speech (M. Hamburger, Trans.). *Ger. Life Lett.* 5, 42–46
66. Cushman, F. (2020) Rationalization is rational. *Behav. Brain Sci.* 43, e28
67. Elga, A. and Rayo, A. (2022) Fragmentation and logical omniscience. *Noûs* 56, 716–741
68. Powers, L.H. (1978) Knowledge by deduction. *Philos. Rev.* 87, 337–371
69. Hawthorne, J. (2024) Inductive logic. In *The Stanford Encyclopedia of Philosophy* (Zalta, E.N. and Nodelman, U., eds)
70. Read, S. (2003) Logical consequence as truth-preservation. *Log. Anal.* 183, 479–493
71. Aronowitz, S. and Lombrozo, T. (2020) Learning through simulation. *Phil. Imprint* 20, 1–18
72. Pacer, M. and Lombrozo, T. (2017) Ockham's razor cuts to the root: simplicity in causal explanation. *J. Exp. Psychol. Gen.* 146, 1761
73. Lombrozo, T. (2007) Simplicity and probability in causal explanation. *Cogn. Psychol.* 55, 232–257
74. Walker, C.M. *et al.* (2017) Effects of explaining on children's preference for simpler hypotheses. *Psychon. Bull. Rev.* 24, 1538–1547
75. Kon, E. and Lombrozo, T. (2017) Explaining guides learners towards perfect patterns, not perfect prediction. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, pp. 682–687, Cognitive Science Society
76. Vrantsidis, T.H. and Lombrozo, T. (2022) Simplicity as a cue to probability: multiple roles for simplicity in evaluating explanations. *Cogn. Sci.* 46, e13169
77. Vrantsidis, T. and Lombrozo, T. (2024) Inside Ockham's razor: a mechanism driving preferences for simpler explanations. *Mem. Cogn.*, Published online July 24, 2024. https://doi.org/10.3758/s13421-024-01604-w
78. Blanchard, T. *et al.* (2018) Bayesian Occam's razor is a razor of the people. *Cogn. Sci.* 42, 1345–1359
79. Berke, M.D. *et al.* (2022) Flexible goals require that inflexible perceptual systems produce veridical representations: implications for realism as revealed by evolutionary simulations. *Cogn. Sci.* 46, e13195
80. McKay, R.T. and Dennett, D.C. (2009) The evolution of misbelief. *Behav. Brain Sci.* 32, 493–510
81. Wilkenfeld, D.A. and Lombrozo, T. (2015) Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Sci. Educ.* 24, 1059–1077
82. Lieder, F. and Griffiths, T.L. (2017) Strategy selection as rational metareasoning. *Psychol. Rev.* 124, 762
83. Russell, S. and Wefald, E. (1991) Principles of metareasoning. *Artif. Intell.* 49, 361–395
84. Allaire-Duquette, G. *et al.* (2021) An fMRI study of scientists with a Ph.D. in physics confronted with naive ideas in science. *NPJ Sci. Learn.* 6, 11
85. Shtulman, A. and Lombrozo, T. (2016) Bundles of contradiction. In *Core Knowledge and Conceptual Change* (Barner, D. and Baron, A.S., eds), pp. 53–72, Oxford University Press
86. Sung, Y. *et al.* (2021) Learning when to quit: meta-reasoning for motion planning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4692–4699, IEEE
87. Casasanto, D. and Lupyan, G. (2015) All concepts are ad hoc concepts. In *The Conceptual Mind: New Directions in the Study of Concepts* (Margolis, E. and Laurence, S., eds), pp. 543–566, MIT Press
88. Chater, N. (2018) *The Mind Is Flat: The Remarkable Shallowness of the Improvising Brain*, Yale University Press
89. Griffiths, T.L. (2020) Understanding human intelligence through human limitations. *Trends Cogn. Sci.* 24, 873–883
90. Lieder, F. and Griffiths, T.L. (2020) Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* 43, e1
91. Yao, S. *et al.* (2024) Tree of thoughts: deliberate problem solving with large language models. *Adv. Neural Inf. Proces. Syst.* 36, 11809–11822
92. Kojima, T. *et al.* (2022) Large language models are zero-shot reasoners. *Adv. Neural Inf. Proces. Syst.* 35, 22199–22213
93. McCoy, R.T. *et al.* (2023) Embers of autoregression: understanding large language models through the problem they are trained to solve. *ArXiv*, Published online September 24, 2023. http://dx.doi.org/10.48550/arXiv.2309.13638
94. Dziri, N. *et al.* (2024) Faith and fate: limits of transformers on compositionality. *ArXiv*, Published online October 31, 2023. https://doi.org/10.48550/arXiv.2305.18654
95. Kambhampati, S. (2024) Can large language models reason and plan? *Ann. N. Y. Acad. Sci.* 1534, 15–18
96. Nisbett, R.E. and Wilson, T.D. (1977) Telling more than we can know: verbal reports on mental processes. *Psychol. Rev.* 84, 231
97. Roser, M. and Gazzaniga, M.S. (2004) Automatic brains – interpretive minds. *Curr. Dir. Psychol. Sci.* 13, 56–59
98. Karmiloff-Smith, A. (1994) Précis of beyond modularity: a developmental perspective on cognitive science. *Behav. Brain Sci.* 17, 693–707
99. Karmiloff-Smith, A. (1995) *Beyond Modularity: A Developmental Perspective on Cognitive Science*, MIT Press
100. Penn, D.C. *et al.* (2008) Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.* 31, 109–130