aps
ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2024, Vol. 19(5) 874–883

DOI: 10.1177/17456916231201401
www.psychologicalscience.org/PPS
Sage

Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet)

Eunice Yiu, Eliza Kosoy, and Alison Gopnik
Department of Psychology, University of California, Berkeley

Abstract

Much discussion about large language models and language-and-vision models has focused on whether these models are intelligent agents. We present an alternative perspective. First, we argue that these artificial intelligence (AI) models are cultural technologies that enhance cultural transmission and are efficient and powerful imitation engines. Second, we explore what AI models can tell us about imitation and innovation by testing whether they can be used to discover new tools and novel causal structures and contrasting their responses with those of human children. Our work serves as a first step in determining which particular representations and competences, as well as which kinds of knowledge or skill, can be derived from particular learning techniques and data. In particular, we explore which kinds of cognitive capacities can be enabled by statistical analysis of large-scale linguistic data. Critically, our findings suggest that machines may need more than large-scale language and image data to allow the kinds of innovation that a small child can produce.

Keywords

innovation, imitation, tool use, causal learning, children, large language models

Recently, large language and language-and-vision models, such as OpenAI's ChatGPT and DALL-E, have sparked much interest and discussion. These systems are trained on an unprecedentedly large amount of data and generate novel text or images in response to prompts. Typically, they are pretrained with a relatively simple objective such as predicting the next item in a string of text correctly. In some more recent systems, they are also fine-tuned both by their designers and through reinforcement learning by human feedback—humans judge the texts and images the systems generate and so further shape what the systems produce.

A common way of thinking about these systems is to treat them as individual agents and then debate how intelligent those agents are. The phrase "an AI" rather than "AI" or "AI system," implying individual agency, is frequently used. Some have claimed that these models can tackle complex commands (e.g., Bubeck et al., 2023), perform abstract reasoning such as inferring theory of mind (e.g., Kosinski, 2023), and demonstrate creativity (e.g., Summers-Stay et al., 2023) in a way that parallels individual human agents.

We argue that this framing is wrong. Instead, we argue that the best way to think of these systems is as powerful new cultural technologies, analogous to earlier technologies such as writing, print, libraries, the Internet, and even language itself (Gopnik, 2022a, 2022b). Large language and vision models provide a new method for easy and effective access to the vast amount of text that others have written and images that others have shaped. These AI systems offer a new means for cultural production and evolution, allowing

Correction (February 2024): This article has been updated with an additional funder, an acknowledgment, and minor grammatical or style corrections since its original publication. See https://doi.org/10.1177/17456916231222009 for additional details.

Corresponding Author:
Eunice Yiu, Department of Psychology, University of California, Berkeley
Email: ey242@berkeley.edu

information to be passed efficiently from one group of people to another (Bolin, 2012; Boyd & Richerson, 1988; Henrich, 2018). They aggregate large amounts of information previously generated by human agents and extract patterns from that information.

This contrasts with perception and action systems that intervene on the external world and generate new information about it. This contrast extends beyond perception and action systems themselves. The kinds of causal representations that are embodied in theories, either scientific or intuitive, are also the result of truth-seeking epistemic processes (e.g., Gopnik & Wellman, 2012); they are evaluated with respect to an external world and make predictions about and shape actions in that world. New evidence from that world can radically revise them. Causal representations, like perceptual representations, are designed to solve "the inverse problem" (Palmer, 1999): the problem of reconstructing the structure of a novel, changing, external world from the data that we receive from that world. Although such representations may be very abstract, as in scientific theories, they ultimately depend on perception and action—on being able to perceive the world and act on it in new ways.

These truth-seeking processes also underlie some AI systems. For example, reinforcement learning systems, particularly model-based systems, can be understood as systems that act on the world to solve something similar to an inverse problem. They accumulate data to construct models of the world that allow for broad and novel generalization. In robotics, in particular, systems such as these make contact with an external world, alter their models as a result, and allow for novel actions and generalizations, although these actions and generalizations are still very limited. Similarly, a number of AI approaches have integrated causal inference and theory formation into learning mechanisms in an attempt to design more human-like systems (Goyal & Bengio, 2022; Lake et al., 2015; Pearl, 2000). These systems are, however, very different from the typical large language and vision models that instead rely on relatively simple statistical inference applied to enormous amounts of existing data.

Truth-seeking epistemic processes contrast with the processes that allow faithful transmission of representations from one agent to another, regardless of the relation between those representations and the external world. Such transmission is crucial for abilities such as language learning and social coordination. There is considerable evidence that mechanisms for this kind of faithful transmission are in place early in development and play a particularly important role in human cognition and culture (Meltzoff & Moore, 1977; Meltzoff & Prinz, 2002).

However, such mechanisms may also be actively in tension, for good and ill, with the truth-seeking mechanisms of causal inference and theory formation. For example, in the phenomenon of "overimitation" human children (and adults) reproduce all the details of a complex action sequence even when they are not causally relevant to the outcome of that action (Lyons et al., 2011; Whiten et al., 2009).

Overimitation may increase the fidelity and efficiency of cultural transmission for complex actions. However, it also means that that transmission is not rooted in a causal understanding that could be altered by further evidence in a changing environment. Similarly, there is evidence that children begin by uncritically accepting testimony from others about the world and revise that testimony only when it is directly contradicted by other evidence (Harris & Koenig, 2006).

We argue that large language models (LLMs) enable and facilitate this kind of transmission in powerful and significant ways by summarizing and generalizing from existing text. However, nothing in their training or objective functions is designed to fulfill the epistemic functions of truth-seeking systems such as perception, causal inference, or theory formation. Even though state-of-the-art LLMs have been trained to estimate uncertainty over the validity of their claims (Kadavath et al., 2022), their output prediction probabilities do not distinguish between epistemic uncertainty, which relates to the lack of knowledge and can be resolved with more training data, and aleatoric uncertainty (relating to chance or stochasticity and so irreducible; Huang et al., 2023; Lin et al., 2023). The fact that such systems "hallucinate" (Azamfirei et al., 2023) is a well-known problem but badly posed—"hallucination" implies that the agent discriminates between veridical and nonveridical representations in the first place, and LLMs do not.

This contrast between transmission and truth is in turn closely related to the imitation/innovation contrast in discussions of cultural evolution in humans (Boyd & Richerson, 1988; Henrich, 2018; Legare & Nielsen, 2015; Tomasello et al., 1993). Cultural evolution depends on the balance between these two different kinds of cognitive mechanisms. Imitation allows the transmission of knowledge or skill from one person to another (Boyd et al., 2011; Henrich, 2016). Innovation produces novel knowledge or skill through contact with a changing world (Derex, 2022). Imitation means that each individual agent does not have to innovate—they can take advantage of the cognitive discoveries of others. But imitation by itself would be useless if some agents did not also have the capacity to innovate. It is the combination of the two that allows cultural and technological progress.

Of course, imitation and transmission may involve some kinds of generalization and novelty. A Wikipedia entry, for example, or even an old-fashioned newspaper article, is the result of multiple human editors collectively

shaping new text that none of them could have generated alone. The result involves a kind of generalization and novelty. Large language models produce similar generalizations. Similarly, it may be possible at times to produce a kind of innovation simply by generalizing from actions that are already known. If I know that a 2-ft ladder will reach a shelf, I may be able to immediately infer that a taller ladder will allow me to reach an even higher shelf, even if I have not seen the ladder used that way before.

However, striking innovations that allow novel adaptations to novel problems and environments require inferences that go beyond the information that has already been acquired. These inferences may take off from existing causal models to generate new causal possibilities that are very different from those that have been observed or transmitted earlier, or they may inspire new explorations of the external world. From the AI perspective a useful way of thinking about it is that imitation involves a kind of interpolative generalization: Within what is already known, skills and knowledge are utilized, emulated, and shared across a variety of contexts. On the other hand, innovation reflects a more extrapolative or "out-of-distribution" generalization.

In any given case, it may be difficult to determine which kinds of cognitive mechanisms produced a particular kind of representation, behavior, knowledge, or skill. For example, my answer to an exam question in school might simply reflect the fact that I have remembered what I was taught, and I can make small generalizations from that teaching. Or it might indicate that I have knowledge that would allow me to make novel predictions about or perform novel actions on the external world. Probing the responses of large language models may give us a tool to help answer that question—at least, in principle. If large models that are trained only on language internal statistics can reproduce particular competencies, for example, producing grammatical text in response to a prompt, that suggests that those abilities can be developed through imitation—extracting existing knowledge encoded in the minds of others. If not, that suggests that these capacities may require innovation—extracting knowledge from the external world.

Thus, large language and vision models provide us with an opportunity to discover which representations and cognitive capacities, in general, human or artificial, can be acquired purely through cultural transmission itself and which require independent contact with the external world—a long-standing question in cognitive science (Barsalou, 2008; Gibson, 1979; Grand et al., 2022; Landauer & Dumais, 1997; Piantadosi, 2023).

In this article, we explore what state-of-the-art large language and language-and-vision models can contribute to our understanding of imitation and innovation. We contrast the performance of models trained on a large corpus of text data, or text and image data, with that of children.

Large Language and Language-and-Vision Models as Imitation Engines

Imitation refers to the behavior of copying or reproducing features or strategies underlying a model's behavior (Heyes, 2001; Tomasello, 1990). It implies interpolative generalization. The way the question or context is presented may vary, but the underlying behavior or idea stems from a repertoire of knowledge and skills that already exist. By observing and imitating others, individuals acquire the skills, knowledge, and conventions that are essential to effectively participate in their cultural groups, promoting cultural continuity over time. An assortment of technological innovations such as writing, print, the Internet, and—we would argue—LLMs, have made this imitation much more effective over time.

Moreover, cultural technologies not only allow access to information; they also codify, summarize, and organize that information in ways that enable and facilitate transmission. Language itself works by compressing information into a digital code. Writing and print similarly abstract and simplify from the richer information stream of spoken language while allowing at the same time wider temporal and spatial access to that information. Print, in addition, allows many people to receive the same information at the same time, and this is, of course, highly amplified by the Internet. At the same time, indexes, catalogs, libraries, and, more recently, Wikis and algorithmic search engines allow humans to quickly find relevant text and images and use those texts and images as a springboard to generate additional text and images.

Deep learning models trained on large data sets today excel at imitation in a way that far outstrips earlier technologies and so represent a new phase in the history of cultural technologies. Large language models such as Anthropic's Claude and OpenAI's ChatGPT can use the statistical patterns in the text in their training sets to generate a variety of new text, from emails and essays to computer programs and songs. GPT-3 can imitate both natural human language patterns and particular styles of writing close to perfectly. It arguably does this better than many people (M. Zhang & Li, 2021). Strikingly and surprisingly, the syntactic structure of the language produced by these systems is accurate. There is some evidence that large language models can even grasp language in more abstract ways than humans and imitate human figurative language understanding (e.g., Jeretic et al., 2020; Stowe et al., 2022). This

suggests that finding patterns in large amounts of human text may be enough to pick up many features of language, independent of any knowledge about the external world.

In turn, this raises the possibility that children learn features of language or images in a similar way. In particular, this discovery has interesting connections to the large body of empirical literature showing that infants are sensitive to the statistical structure of linguistic strings and visual images from a very young age (e.g., Kirkham et al., 2002; Saffran et al., 1996). The LLMs suggest that this may enable much more powerful kinds of learning than we might have thought, such as the ability to learn complex syntax.

On the other hand, although these systems allow skilled imitation, the imitation that they facilitate may differ from that of children in important ways. There are debates in the developmental literature about how much childhood imitation simply reflects faithful cultural transmission (as in the phenomenon of overimitation) and how much it is shaped by and in the service of broader truth-seeking processes such as understanding the goals and intentions of others. Children can meaningfully decompose observed visual and motor patterns in relation to the agent, target object, movement path, and other salient features of events (Bekkering et al., 2000; Gergely et al., 2002). Moreover, children distinctively copy intentional actions (Meltzoff, 1995), discarding apparently failed attempts, mistakes, and causally inefficient actions (Buchsbaum et al., 2011; Schulz et al., 2008) when they seek to learn skills from observing other people (Over & Carpenter, 2013). Although the imitative behavior of large language and vision models can be viewed as the abstract mapping of one pattern to another, human imitation appears to be mediated by goal representation and the understanding of causal structure from a young age. It would be interesting to see whether large models also replicate these features of human imitation.

Can Large Language and Language-and-Vision Models Innovate?

Can LLMs discover new tools?

Where might we find empirical evidence for this contrast between transmission and truth, imitation and innovation? One important and relevant set of capacities involves tool use and innovation. The most ancient representative of the human genus is called *Homo habilis* ("handy man") because of their ability to discover and use novel stone tools. Tool use is one of the best examples of the advantages of cultural transmission and of the balance between imitation and innovation. Imitation allows a novice to observe a model and reproduce their actions to bring about a particular outcome, even without understanding entirely the fine physical mechanisms and causal properties of the tool. Techniques such as "behavior cloning" in AI and robotics use a similar approach.

Again, however, the ability to imitate and use existing tools in an interpolative way depends on the parallel ability to discover new tools in an extrapolative way. Tool innovation is an indispensable part of human lives, and it has also been observed in a variety of nonhuman animals such as crows (Von Bayern et al., 2009) and chimpanzees (Whiten et al., 2005). Tool innovation has often been taken to be a distinctive mark of intelligence in biological systems (Emery & Clayton, 2004; Reader & Laland, 2002).

Tool use can then be a particularly interesting point of comparison for understanding imitation and innovation in both models and children. Both computational models and humans can encode information about objects (e.g., Allen et al., 2020), but their capabilities for tool imitation versus tool innovation might differ. In particular, our hypothesis would predict that the models might capture familiar tool uses well (e.g., predicting appropriately that a hammer should be used to bang in a nail). However, these systems might have more difficulty producing the right responses for tool innovation involving unusual or novel tools, which depends on discovering and using new causal properties, functional analogies, and affordances.

We might, however, also wonder whether young children can themselves perform this kind of innovation, or whether it depends on explicit instruction and experience. Physically building a new tool from scratch and then executing a series of actions that lead to a desired goal is a difficult task for young children (Beck et al., 2011). But children might find it easier to recognize new functions in everyday objects and to select appropriate object substitutes in the absence of typical tools to solve various physical tasks. In an ongoing study of tool innovation (Yiu & Gopnik, 2023), we have investigated whether human children and adults can insightfully use familiar objects in new ways to accomplish particular outcomes and compared the results to the output of large deep learning models such as GPT-3 and GPT-4.

Tool innovation can involve designing new tools from scratch, but it can also refer to discovering and using old tools in new ways to solve novel problems (Rawlings & Legare, 2021). We might think of this as the ability to make an out-of-distribution generalization about a functional goal. Our experiment examines the latter type of tool innovation.

Our study has two components: an "imitation" component (making an interpolative judgment from existing knowledge about objects) and an "innovation" component (making an extrapolative judgment about the new

ways that objects could be used). In the innovation part of the study, we present a series of problems in which a goal has to be executed in the absence of the typical tool (e.g., drawing a circle in the absence of a compass). We then provide alternative objects for participants to select: (a) an object that is more superficially similar to the typical tool and is associated with it but is not functionally relevant to the context (e.g., a ruler), (b) an object that is superficially dissimilar but has the same affordances and causal properties as the typical tool (e.g., a teapot that possesses a round bottom), and (c) a totally irrelevant object (e.g., a stove). In the imitation part of the study, we present the same sets of objects but ask participants to select which of the object options would "go best" with the typical tool (e.g., a compass and a ruler are more closely associated than a compass and a teapot).

So far, we have found that both children aged 3 to 7 years old presented with animations of the scenario (n = 42, M_{age} = 5.71 years, SD = 1.24) and adults (n = 30, M_{age} = 27.80 years, SD = 5.54) can recognize common superficial relationships between objects when they are asked which objects should go together ($M_{children}$ = 88.4%, $SE_{children}$ = 2.82%; M_{adults} = 84.9%, SE_{adults} = 3.07%). But they can also discover new functions in everyday objects to solve novel physical problems and so select the superficially unrelated but functionally relevant object ($M_{children}$ = 85.2%, $SE_{children}$ = 3.17%; M_{adults} = 95.7%, SE_{adults} = 1.04%). In ongoing work, we have found that children demonstrate these capacities even when they receive only a text description of the objects, with no images.

Using exactly the same text input that we used to test our human participants, we queried OpenAI's GPT-4, gpt-3.5-turbo, and text-davinci-003 models; Anthropic's Claude; and Google's FLAN-T5 (XXL). Because we noticed that the models could alter their responses depending on how the order of options was presented, we queried the models six times for every scenario to account for the six different orders that could be generated by the three options. We set model outputs as deterministic with a temperature of 0 and kept the default values for all other parameters (Binz & Schulz, 2023; Hu et al., 2022). We averaged the scores (1 for selecting the relevant object and 0 for any other response) across the six repeated trials. As we predicted we found that these large language models are almost as capable of identifying superficial commonalities between objects as humans are. They are sensitive to the superficial associations between the objects, and they excel at our imitation tasks (M_{GPT4} = 83.3%, SE_{GPT4} = 4.42%; $M_{gpt\text{-}3.5\text{-}turbo}$ = 73.1%, $SE_{gpt\text{-}3.5\text{-}turbo}$ = 5.26%; $M_{davinci}$ = 59.9%, $SE_{davinci}$ = 5.75%; M_{Claude} = 69.9%, SE_{Claude} = 5.75%; M_{Flan} = 74.8%, SE_{Flan} = 5.17%)—they generally respond that the ruler goes with the compass. However, they are less capable than humans when they are asked to select a novel functional tool to solve a problem (M_{GPT4} = 75.9%, SE_{GPT4} = 4.27%; $M_{gpt\text{-}3.5\text{-}turbo}$ = 58.9%, $SE_{gpt\text{-}3.5\text{-}turbo}$ = 5.64%; $M_{davinci}$ = 8.87%, $SE_{davinci}$ = 2.26%; M_{Claude} = 58.16%, SE_{Claude} = 6.06%; M_{Flan} = 45.7%, SE_{Flan} = 5.42%)—they again choose the ruler rather than the teapot to draw a circle. This suggests that simply learning from large amounts of existing language may not be sufficient to achieve tool innovation. Discovering novel functions in everyday tools is not about finding the statistically nearest neighbor from lexical co-occurrence patterns. Rather, it is about appreciating the more abstract functional analogies and causal relationships between objects that do not necessarily belong to the same category or are associated in text. In these examples, people must use broader causal knowledge, such as understanding that tracing an object will produce a pattern that matches the object's shape, to produce a novel action that has not been observed or described before, in much the same way a scientific theory, for example, allows novel interventions on the world (Pearl, 2000). Compared with humans, large language models are not as successful at this type of innovation task. On the other hand, they excel at generating responses that simply demand some abstraction from existing knowledge.

One might ask whether success on our task also requires visual and spatial information rather than merely text. Indeed, GPT-4, a large multimodal model that is trained on larger amounts of images and text, demonstrates better performance than the other large language models on both the innovative and imitative tasks. Nevertheless, despite the massive amounts of vision and language training data, it is still not as innovative as human adults are when they discover new functions in existing objects. It is also unclear whether GPT-4's improved performance stems from its multimodal character or from reinforcement learning from human feedback—a point we return to later.

Can LLMs discover novel causal relationships and use them to design interventions?

Discovering novel tools depends on being able to infer a novel causal relationship, such as drawing a circle by tracing the bottom of a teapot. A substantial amount of research shows that even very young children excel at discovering such relationships. Information about causal structure can be conveyed through imitation and cultural transmission. In fact, from a very young age, even infants will reproduce an action they have observed to bring about an effect (Waismeyer et al., 2015). However, very young children can also infer novel causal structure by observing complex statistical

relations among events, and most significantly, by acting on the world themselves to bring about effects like a scientist performing experiments (Cook et al., 2011; Gopnik et al., 2004, 2017; Gopnik & Tenenbaum, 2007; Schulz et al., 2007). Causal discovery is a particularly good example of a cognitive process that is directed at solving an inverse problem and discovering new truths through perception and action. Moreover, these processes of causal discovery do not depend on particular assumptions about "intuitive physics." Very young children can make such inferences about psychological and social relationships as well as physical ones, and they can discover new causal relations that actually contradict the assumptions of intuitive physics (Gopnik & Wellman, 2012).

In another line of research (Kosoy et al., 2022, 2023), we have explored whether LLMs and other AI models can discover and use novel causal structure. In these studies we use a virtual "blicket detector"—a machine that lights up and plays music when you put some objects on it but not others. The blicket detector can work on different abstract principles or "overhypotheses": individual blocks may activate it, or you may need a combination of blocks to do so. An overhypothesis refers to an abstract principle that reduces a hypothesis space at a less abstract level (Kemp et al., 2007), and a causal overhypothesis refers to transferable abstract hypotheses about sets of causal relationships (Kosoy et al., 2022). If you know that it takes two blocks to make the machine go, you will generate different specific hypotheses about which blocks are blickets.

The blicket detector tasks intentionally involve a new artifact, described with new words, so that the participants cannot easily use past culturally transmitted information, such as the fact that flicking a light switch makes a bulb go on. Assumptions about intuitive physics will also not enable a solution. In these experiments, we simply ask children to figure out how the machines work and allow them to freely explore and act to solve the task and determine which blocks are blickets. Even 4-year-old children spontaneously acted on the systems and discovered their structure—they figured out which ones were blickets and used them to make the machine go.

We then gave a variety of LLMs, including OpenAI's ChatGPT, Google's PaLM, and most recently LaMDA, the same data that the children produced, described in language (e.g., "I put the blue one and the red one on the machine and the machine lit up") and prompted the systems to answer questions about the causal structure of the machine (e.g., "Is the red one a blicket?").

LLMs did not produce the correct causal overhypotheses from the data. Young children, in contrast, learned novel causal overhypotheses from only a handful of observations, including the outcome of their own experimental interventions, and applied the learned structure to novel situations. In contrast, large language models and vision-and-language models, as well as both deep reinforcement learning algorithms and behavior cloning, struggled to produce the relevant causal structures, even after massive amounts of training compared with children. This is consistent with other recent studies: LLMs produce the correct text in cases such as causal vignettes, in which the patterns are available in the training data, but often fail when they are asked to make inferences that involve novel events or relations in human thought (e.g., Binz & Schulz, 2023; Mahowald et al., 2023), sometimes even when these involve superficially slight changes to the training data (e.g., Ullman, 2023).

Challenges of Studying Large Language and Language-and-Vision Models: The Questions Left Unanswered

It is difficult to escape the language of individual agency, for example, to ask whether AI can or cannot innovate (e.g., González-Díaz & Palacios-Huerta, 2022; Stevenson et al., 2022), solve a causal problem (e.g., Kıcıman et al., 2023), or even can or cannot be sentient or intelligent (e.g., Mitchell, 2023). A great deal of the discussion about AI has this character. But we emphasize again that the point of this work is neither to decide whether or not LLMs are intelligent agents nor to present some crucial comparative "gotcha" test that would determine the answer to such questions in AI systems. Instead, the research projects we have briefly described here are a first step in determining which representations and competences, as well as which kinds of knowledge or skill, can be derived from which learning techniques and data. Which kinds of knowledge can be extracted from large bodies of text and images, and which depend on actively seeking the truth about an external world?

We want to emphasize again that other AI systems, such as model-based reinforcement learning or causal inference systems, may indeed more closely approximate truth-seeking cognitive systems. In fact, we also evaluated the performance of other AI systems, including two popular deep reinforcement learning algorithms, Advantage Actor Critic (A2C) and Proximal Policy Optimization Version 2 (PPO2), which are trained on all possible overhypotheses prior to the test trials. Although these systems are conceptually closer to the truth-seeking systems children use, they are still extremely limited in comparison.

There is a great deal of scope for research that uses developmental-psychology techniques to investigate AI

systems and vice versa (Frank, 2023). Of course, there are anecdotal instances in which large language models seem to exhibit intelligent-like behaviors (Bubeck et al., 2023), solving arguably novel and complex tasks from physics and mathematics to story writing and image generation. Nonetheless, developmentalists have long realized that superficially similar behaviors, including creative- and innovative-like behaviors in humans and AI models, can have very different psychological origins and can be the result of very different learning techniques and data. As a result, we have put considerable methodological energy into trying to solve this problem. A particular conversation with a child, however compelling, is just the start of a proper research program including novel, carefully controlled tests such as our tests of tool innovation and novel causal inference. The conversation may reflect knowledge that has come through imitation within the training data set, statistical pattern recognition, reinforcement from adults, or conceptual understanding; the job of the developmental psychologist is to distinguish these possibilities. This should also be true of our assessments of AI systems.

At the same time AI systems have their own properties that need to be considered when we compare their output to that of humans. These can sometimes be problematic; for example, once a particular cognitive test is explicitly described in Internet text, it then becomes part of a large language model's training sample—we found that in some cases the systems referred to our earlier published blicket-detector articles as the source for their answers. There are many more cases in the literature in which large language models, including GPT-4, seem to respond to novel examples sampled from the training distribution almost perfectly (again reinforcing that they are perfect imitators) but then fail miserably to generalize to out-of-distribution examples that require the discovery of more abstract causal hypotheses and innovation (e.g., Chowdhery et al., 2022; Talmor et al., 2020; H. Zhang et al., 2022).

In addition, the more recent versions of GPT, GPT-4, and GPT-3.5, have also been fine-tuned through reinforcement learning from human feedback. This also raises problems. Reinforcement learning from human feedback may itself be considered a method for enabling cultural transmission. However, in practice, we know very little about exactly what kinds of feedback these systems receive or how they are shaped by that feedback. Reinforcement learning from human feedback is both opaque and variable and may simply edit out the most obvious mistakes and errors.

On the other hand, these systems, particularly the "classic" LLMs, also have the advantage that we know more about their data and learning techniques than we do about those of human children. For example, we know that the data for GPT systems is Internet text and that the training function involves predicting new text from earlier text. We know that large language models and language-and-vision models are built on deep neural networks and trained on immense amounts of unlabeled text or text-image pairings.

These kinds of techniques may indeed contribute to some kinds of human learning as well. Children do learn through cultural transmission and statistical generalizations from data. But human children also learn in very different ways. Although we do not know the details of children's learning algorithms or data, we do know that, unlike large language and language-and-vision models, children are curious, active, self-supervised, and intrinsically motivated. They are capable of extracting novel and abstract structures from the environment beyond statistical patterns, spontaneously making overhypotheses and generalizations, and applying these insights to new situations.

Because performance in large deep learning models has been steadily improving with increasing model size on various tasks, some have advocated that simply scaling up language models could allow task-agnostic, few-shot performance (e.g., Brown et al., 2020). But a child does not interact with the world better by increasing their brain capacity. Is building the tallest tower the ultimate way to reach the moon? Putting scale aside, what are the mechanisms that allow humans to be effective and creative learners? What in a child's "training data" and learning capacities is critically effective and different from that of LLMs? Can we design new AI systems that use active, self-motivated exploration of the real external world as children do? And what we might expect the capacities of such systems to be? Comparing these systems in a detailed and rigorous way can provide important new insights about both natural intelligence and AI.

Conclusion

Large language models such as ChatGPT are valuable cultural technologies. They can imitate millions of human writers, summarize long texts, translate between languages, answer questions, and code programs. Imitative learning is critical for promoting and preserving knowledge, artifacts, and practices faithfully within social groups. Moreover, changes in cultural technologies can have transformative effects on human societies and cultures—for good or ill. There is a good argument that the initial development of printing technology contributed to the Protestant Reformation. Later improvements in printing technology in the 18th century were responsible for both the best parts of the American Revolution and the worst parts of the French Revolution

(Darnton, 1982). Large language and language-and-vision models may well have equally transformative effects in the 21st century.

However, cultural evolution depends on a fine balance between imitation and innovation. There would be no progress without innovation—the ability to expand, create, change, abandon, evaluate, and improve on existing knowledge and skills. Whether this means recasting existing knowledge in new ways or creating something entirely original, innovation challenges the status quo and questions the conventional wisdom that is the training corpus for AI systems. Large language models can help us acquire information that is already known more efficiently, even though they are not innovators themselves. Moreover, accessing existing knowledge much more effectively can stimulate more innovation among humans and perhaps the development of more advanced AI. But ultimately, machines may need more than large-scale language and images to match the achievements of every human child.

Transparency

Action Editor: Mirta Galesic
Editor: Interim Editorial Panel
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
We also gratefully acknowledge the support of the following agencies: DARPA 047498-002 Machine Common Sense, John Templeton Foundation 61475 The Development of Curiosity, Templeton World Charity Foundation TWCF0434 Play: A Computational Account, DOD ONR MURI Self-Learning Perception Through Real World Interaction.

ORCID iD

Eunice Yiu https://orcid.org/0000-0002-3505-5525

Acknowledgments

We are grateful to Shiry Ginosar, Jitendra Malik, and the participants of the DARPA MCS group, the SFI Institute and the Simons Theoretical Computer Science Group for discussion, and to the participants and their parents, the museums and preschools who participated, and the undergraduates who assisted, Jalaya Allen, Yuki Bian, Fei Dai, Megan Lui, Sophia Liu, Iran Torres Aleman, Luc LaMontagne,Bryanna Kauffman, Zane Levin, Elijah Phipps, Janie Dent, Jenna Levin, Nikita Kumar and Athena Leong.

References

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences, USA*, *117*(47), 29302–29310. https://doi.org/10.1073/pnas.1912341117

Azamfirei, R., Kudchadkar, S. R., & Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care*, *27*(1), Article 120. https://doi.org/10.1186/s13054-023-04393-x

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Beck, S. R., Apperly, I. A., Chappell, J., Guthrie, C., & Cutting, N. (2011). Making tools isn't child's play. *Cognition*, *119*(2), 301–306. https://doi.org/10.1016/j.cognition.2011.01.003

Bekkering, H., Wohlschlager, A., & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(1), 153–164. https://doi.org/10.1080/713755872

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

Bolin, G. (2012). Introduction: Cultural technologies in cultures of technology. In *Cultural technologies* (pp. 1–15). Routledge.

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago Press.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences, USA*, *108*(Suppl. 2), 10918–10925. https://doi.org/10.1073/pnas.1100290108

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 1877–1901). Curran Associates, Inc.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with gpt-4*. arXiv. https://doi.org/10.48550/arXiv.2303.12712

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331–340. https://doi.org/10.1016/j.cognition.2010.12.001

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., . . . Fiedel, N. (2022). *PaLM: Scaling language modeling with pathways*. arXiv. https://doi.org/10.48550/arXiv.2204.02311

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349. https://doi.org/10.1016/j.cognition.2011.03.003

Darnton, R. (1982). What is the history of books? *Daedalus*, *111*, 65–83.

Derex, M. (2022). Human cumulative culture and the exploitation of natural phenomena. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1843), Article 20200311. https://doi.org/10.1098/rstb.2020.0311

Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science*, *306*(5703), 1903–1907. https://doi.org/10.1126/science.1098410

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *2*, 451–452. https://doi.org/10.1038/s44159-023-00211-x

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*(6873), Article 755. https://doi.org/10.1038/415755a

Gibson, J. J. (1979). *The ecological approach to visual perception*. Psychology Press.

González-Díaz, J., & Palacios-Huerta, I. (2022). *AlphaZero ideas*. SSRN. https://ssrn.com/abstract=4140916

Gopnik, A. (2022a, July 15). What AI still doesn't know how to do. *The Wall Street Journal*. https://www.wsj.com/articles/what-ai-still-doesnt-know-how-to-do-11657891316

Gopnik, A. (2022b, October 31). Children, creativity, and the real key to intelligence. *Observer*. https://www.psychologicalscience.org/observer/children-creativity-intelligence

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*(1), 3–32. https://doi.org/10.1037/0033-295X.111.1.3

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences, USA*, *114*(30), 7892–7899. https://doi.org/10.1073/pnas.1700811114

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental Science*, *10*(3), 281–287. https://doi.org/10.1111/j.1467-7687.2007.00584.x

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108. https://doi.org/10.1037/a0028044

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. Proceedings of the Royal Society A: Mathematical, *Physical and Engineering Sciences*, *478*(2266), Article 20210068. https://doi.org/10.1098/rspa.2021.0068

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, *6*(7), 975–987. https://doi.org/10.1038/s41562-022-01316-8

Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development*, *77*(3), 505–524. https://doi.org/10.1111/j.1467-8624.2006.00886.x

Henrich, J. (2016). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Henrich, J. (2018). Human cooperation: The hunter-gatherer puzzle. *Current Biology*, *28*(19), R1143–R1145. https://doi.org/10.1016/j.cub.2018.08.005

Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Sciences*, *5*(6), 253–261. https://doi.org/10.1016/S1364-6613(00)01661-2

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2022). *A fine-grained comparison of pragmatic language understanding in humans and language models*. arXiv. https://doi.org/10.48550/arXiv.2212.06801

Huang, Y., Song, J., Wang, Z., Chen, H., & Ma, L. (2023). *Look before you leap: An exploratory study of uncertainty measurement for large language models*. arXiv. https://doi.org/10.48550/arXiv.2307.10236

Jeretic, P., Warstadt, A., Bhooshan, S., & Williams, A. (2020). *Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition*. arXiv. https://doi.org/10.48550/arXiv.2004.03066

Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). *Causal reasoning and large language models: Opening a new frontier for causality*. arXiv. https://doi.org/10.48550/arXiv.2305.00050

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., & Kaplan, J. (2022). *Language models (mostly) know what they know*. arXiv. https://doi.org/10.48550/arXiv.2207.05221

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321. https://doi.org/10.1111/j.1467-7687.2007.00585.x

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. https://doi.org/10.1016/S0010-0277(02)00004-5

Kosinski, M. (2023). *Theory of mind may have spontaneously emerged in large language models*. arXiv. https://doi.org/10.48550/arXiv.2302.02083

Kosoy, E., Chan, D. M., Liu, A., Collins, J., Kaufmann, B., Huang, S. H., Hamrick, J. B., Canny, J., Ke, N. R., & Gopnik, A. (2022). *Towards understanding how machines can learn causal overhypotheses*. arXiv. https://doi.org/10.48550/arXiv.2206.08353

Kosoy, E., Reagan, E. R., Lai, L., Gopnik, A., & Cobb, D. K. (2023). *Comparing machines and children: Using developmental psychology experiments to assess the strengths and weaknesses of LaMDA responses*. ArXiv. https://doi.org/10.48550/arXiv.2305.11243

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Legare, C. H., & Nielsen, M. (2015). Imitation and innovation: The dual engines of cultural learning. *Trends in Cognitive Sciences*, *19*(11), 688–699. https://doi.org/10.1016/j.tics.2015.08.005

Lin, Z., Trivedi, S., & Sun, J. (2023). *Generating with confidence: Uncertainty quantification for black-box large language models*. arXiv. https://doi.org/10.48550/arXiv.2305.19187

Lyons, D. E., Damrosch, D. H., Lin, J. K., Macris, D. M., & Keil, F. C. (2011). The scope and limits of overimitation in the transmission of artefact culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1158–1167. https://doi.org/10.1098/rstb.2010.0335

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). *Dissociating language and thought in large language models: A cognitive perspective*. arXiv. https://doi.org/10.48550/arXiv.2301.06627

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*(5), 838–850. https://doi.org/10.1037/0012-1649.31.5.838

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*(4312), 75–78. https://doi.org/10.1126/science.198.4312.75

Meltzoff, A. N., & Prinz, W. (Eds.). (2002). *The imitative mind: Development, evolution and brain bases* (Vol. 6). Cambridge University Press.

Mitchell, M. (2023). How do we know how smart AI systems are? *Science*, *381*(6654), Article eadj5957. https://doi.org/10.1126/science.adj5957

Over, H., & Carpenter, M. (2013). The social side of imitation. *Child Development Perspectives*, *7*(1), 6–11. https://doi.org/10.1111/cdep.12006

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Piantadosi, S. T. (2023). *Modern language models refute Chomsky's approach to language*. LingBuzz. https://lingbuzz.net/lingbuzz/007180

Rawlings, B., & Legare, C. H. (2021). Toddlers, tools, and tech: The cognitive ontogenesis of innovation. *Trends in Cognitive Sciences*, *25*(1), 81–92. https://doi.org/10.1016/j.tics.2020.10.006

Reader, S. M., & Laland, K. N. (2002). Social intelligence, innovation, and enhanced brain size in primates. *Proceedings of the National Academy of Sciences, USA*, *99*(7), 4436–4441. https://doi.org/10.1073/pnas.062041299

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, *43*(5), 1124–1139. https://doi.org/10.1037/0012-1649.43.5.1124

Schulz, L. E., Hooppell, C., & Jenkins, A. C. (2008). Judicious imitation: Children differentially imitate deterministically and probabilistically effective actions. *Child Development*, *79*(2), 395–410.

Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022). *Putting GPT-3's creativity to the (alternative uses) test*. arXiv. https://doi.org/10.48550/arXiv.2206.08932

Stowe, K., Utama, P., & Gurevych, I. (2022). IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 5375–5388). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.369

Summers-Stay, D., Voss, C. R., & Lukin, S. M. (2023, February 13). *Brainstorm, then select: A generative language model improves its creativity score* [Paper presentation]. AAAI-23 Workshop on Creative AI Across Modalities, Washington, DC, United States.

Talmor, A., Tafjord, O., Clark, P., Goldberg, Y., & Berant, J. (2020). *Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge*. arXiv. https://doi.org/10.48550/arXiv.2006.06609

Tomasello, M. (1990). Cultural transmission in the tool use and communicatory signaling of chimpanzees? In S. T. Parker & K. R. Gibson (Eds.), *"Language" and intelligence in monkeys and apes: Comparative developmental perspectives* (pp. 274–311). Cambridge University Press. https://doi.org/10.1017/CBO9780511665486.012

Tomasello, M., Kruger, A., & Ratner, H. (1993). Cultural learning. *Behavioral and Brain Sciences*, *16*, 495–552. https://doi.org/10.1017/S0140525X0003123X

Ullman, T. (2023). *Large language models fail on trivial alterations to theory-of-mind tasks*. arXiv. https://doi.org/10.48550/arXiv.2302.08399

Von Bayern, A. M., Heathcote, R. J., Rutz, C., & Kacelnik, A. (2009). The role of experience in problem solving and innovative tool use in crows. *Current Biology*, *19*(22), 1965–1968. https://doi.org/10.1016/j.cub.2009.10.037

Waismeyer, A., Meltzoff, A. N., & Gopnik, A. (2015). Causal learning from probabilistic events in 24-month-olds: An action measure. *Developmental Science*, *18*(1), 175–182. https://doi.org/10.1111/desc.12208

Whiten, A., Horner, V., & de Waal, F. (2005). Conformity to cultural norms of tool use in chimpanzees. *Nature*, *437*(7059), 737–740. https://doi.org/10.1038/nature04047

Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2417–2428. https://doi.org/10.1098/rstb.2009.0069

Yiu, E., & Gopnik, A. (2023). Discovering new functions in everyday tools by children, adults and LLM's. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 45, No. 45). https://escholarship.org/uc/item/5247k5m

Zhang, H., Li, L. H., Meng, T., Chang, K. W., & Broeck, G. V. D. (2022). *On the paradox of learning to reason from data*. arXiv. https://doi.org/10.48550/arXiv.2205.11502

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, *1*(6), 831–833. https://doi.org/10.1016/j.fmre.2021.11.011