Trends in
Cognitive Sciences

25th Anniversary Series: Looking Forward

Forum

Capturing advanced human cognitive abilities with deep neural networks

James L. McClelland [1,*]

How can artificial neural networks capture the advanced cognitive abilities of pioneering scientists? I suggest they must learn to exploit human-invented tools of thought and human-like ways of using them, and must engage in explicit goal-directed problem solving as exemplified in the activities of scientists and mathematicians and taught in advanced educational settings.

The capabilities of artificial computational systems have advanced dramatically over the past decade. Progress has been driven by a combination of huge data sets, massive computational resources, and innovations in deep artificial neural networks. New papers appear frequently, accompanied by press releases touting breakthroughs. Are we on the verge of creating truly intelligent artificial systems? Or are these systems just mindless statistical machines?

In my view, making a binary judgment or categorizing these systems into preexisting but ill-defined categories is unhelpful. It is more useful to ask whether such systems will ever capture advanced human abilities underlying the achievements of pioneering mathematicians and scientists. The 25th anniversary of *Trends in Cognitive Sciences* provides an opportunity to consider what approaches might allow them to someday capture these advanced cognitive abilities.

Capturing advanced cognitive abilities

We are awed by the achievements of pioneering thought leaders in science, mathematics, and other domains. To me, the ability to recognize an unsolved problem or untapped opportunity and then address or exploit it through a process that may take weeks or years is at the heart of their accomplishments. Newton surmised that gravity might explain both the falling of an apple from a tree and the orbits of smaller celestial bodies around larger ones. Darwin realized that the traits of organisms might have evolved, but he was unsatisfied with other people's explanations. Rumelhart recognized that although neural networks had many appealing properties, the absence of a method that would allow these networks to learn to perform arbitrary nonlinear computations was a profound limitation. Once these problems were identified, these great scientists set out to solve them. Their achievements are certainly beyond the reach of today's artificial systems – identifying the problem to be solved and organizing an effort to solve it remains the province of these systems' human designers. However, it may be useful to set these kinds of achievements as long-term goals because the same abilities are tapped on a smaller scale whenever we engage in advanced cognitive activities like explicit reasoning and problem-solving.

Will neural networks be part of the solution?

Most of the breakthroughs in artificial intelligence over the past decade have been achieved with deep neural networks. These systems recognize objects, master games, and translate language in ways that have sometimes been surprising, and I believe their ability to exploit context and experience implicitly in graded, multilayer, connection-based systems will be central to the effort to capture the intuitive aspects of advanced cognitive abilities. However, to describe a computational system as a neural network is only to describe the microstructure of the system. Its macrostructure is also crucial. The achievements of recent years depend on the emergence of a network architecture called the transformer [1]. Transformers enhance a network's ability to exploit context, and I have argued that they should be extended to capture multiple input modalities and to rely on human-like memory systems [2] to extend their capabilities still further. By itself, however, this is unlikely to be enough to capture the advanced cognitive abilities of pioneering scientists.

Experience: human-invented tools for thought and schools for learning them

Many cognitive scientists believe that advanced cognitive abilities rely on systematic thought, defined as thought that exploits formal systems like those found in logic, mathematics, and computer science. I start from the premise that these forms of thought are human cultural inventions [3], and that the ability to exploit them depends on immersion in educational systems that promote and encourage them [4]. To acquire cognitive abilities that begin to mirror those of pioneering scientists, neural networks must be immersed in similar settings.

What is the nature of our human-invented systems of thought and what are the characteristics of the experiences from which humans acquire them? These systems are not just systems for manipulating expressions according to category- and structure-sensitive rules, as proposed, for example, by Fodor and Pylyshyn [5]. Although categories and rules are central to these systems, their role is to characterize the properties of idealized objects and their relationships and to provide tools for making inferences about unknown properties and relations from properties and relations that are given. Presentations of these systems in educational materials

do not simply rely on formal argument and symbol manipulation. Instead, they often (e.g., in textbooks [6,7]) focus on reasoning about these objects and their properties, appealing to general principles and using semiformalized mixtures of text, diagrams, and formal expressions, even in discourse described as presenting a proof. I provide one illustration of this from the writings of the mathematical psychologist Roger Shepard [8] in Box 1.

To allow artificial systems to capture this kind of thinking ability, we must expose them to mathematical and scientific arguments of these types. Indeed, several recent transformer-based neural network systems [9] have demonstrated a new level of success in mathematical problem solving after training them with a corpus of text- and symbolic-formula-based presentations of mathematical ideas and arguments found in web pages and scientific articles. Extending these models with visual input systems and effectors that allow them to perceive and manipulate external depictions (e.g., diagrams) of idealized mathematical objects may allow these systems to exploit the visuospatial intuitions illustrated in Box 1. However, these systems remain challenged when a structured argument must be assembled over a series of intermediate steps. This is where, I argue, explicit goal-directed thinking should come in.

Becoming goal directed

For the most part, today's neural networks are successful because human programmers have specified their learning objectives and engineered their exposure to experience. Maximizing reward and minimizing error in predicting some aspects of their input from other aspects of their input are the two main learning objectives. However, humans represent and work toward achieving specified goals under specified constraints and often engage in dialog with others about these goals and constraints and how to achieve the desired outcome within them. This can involve relying on and even formulating general principles and rules in explicit form. Whenever young children engage with others to bake cookies, play a game, or build a toy ice cream parlor with Lego

Box 1. Visuospatial and informal presentation of a mathematical proof

Shepard [8] presents Figure I as a proof of the Pythagorean Theorem: in right-angled triangles, the area of the square on the side opposite the right angle equals the sum of the areas of the squares on the sides containing the right angle.

The proof proceeds by noting that the four triangles in Figure IB are translated, rotated, and/or flipped copies of the triangle in Figure IA with sides of arbitrary lengths a and b and hypotenuse of length c. These triangles are placed within a larger square leaving unoccupied regions that are both square with areas a^2 and b^2, as shown in panel C. Shepard invites us to imagine the triangles translating within the larger square to fit into its corners as illustrated in panel D. Here, we see the empty (shaded) portion of the larger square as a quadrilateral with sides of length c. Shepard notes that the shaded quadrilateral in D is invariant under 90° rotations to prove that it is a square and hence has area c^2. He then notes that the region of the larger square not occupied by the triangles must be invariant under nonoverlapping rearrangements of the triangles. Hence, the sum of the areas of the two shaded squares in C (i.e., $a^2 + b^2$) must equal the area of the single shaded square in D (i.e., c^2). He then writes Q.E.D. Figure reprinted, with permission, from [8].

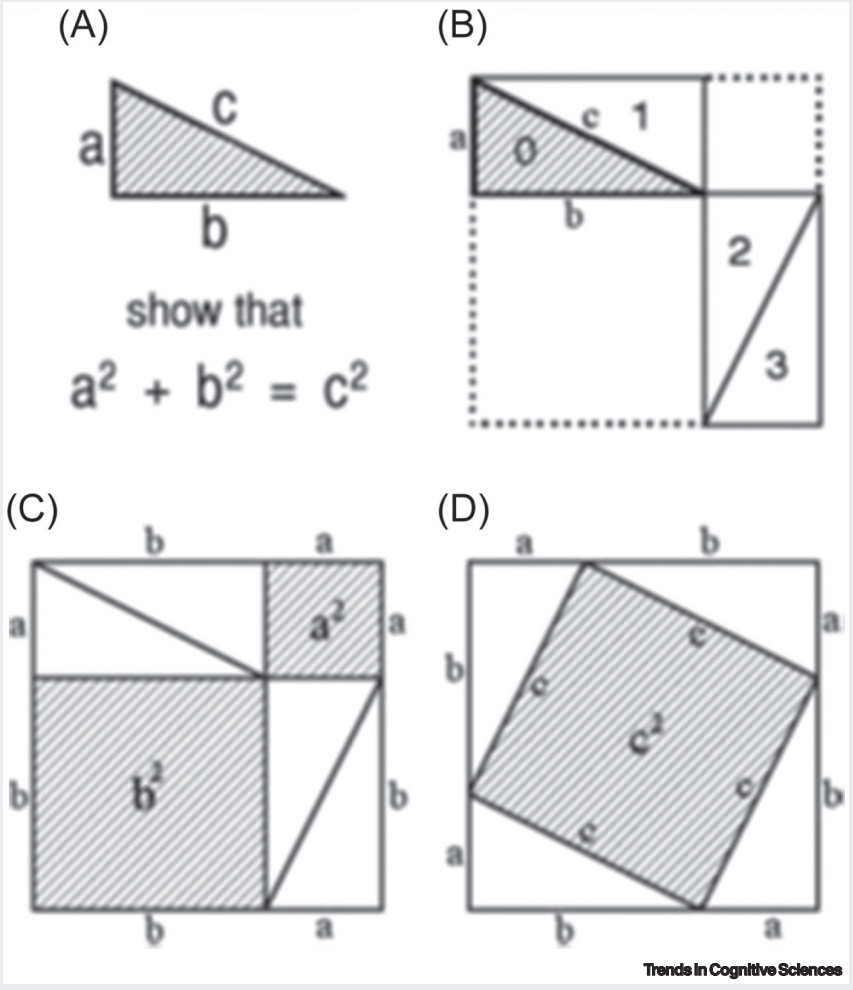

Figure I. A proof of the Pythagorean Theorem.

bricks, they are participating in such an activity. Complex mathematical problem-solving involving the need to establish intermediate results as subgoals are extensions of such goal-directed activity, using (and even inventing as in Newton's case) the formal reasoning systems that support this activity.

However, being goal-directed is more than just contextualizing behavior in terms of goals. Once goals and relevant task constraints are known, the system must be able to evaluate how well current conditions meet its goals and must organize its activities toward achieving them, allowing for repeated attempts [10], and learning how to refine and extend its skill set through a combination of engagement with others, self-practice, and self-teaching. The puzzle in Box 2 provides an example that illustrates many of these points. I argue that these activities are extensions of everyday cognitive activities, and play a role in guiding our thought processes in ways not currently captured in neural networks. By engaging them in such activities, starting with everyday situations and continuing onward into logical, mathematical, and scientific problem-solving situations, our neural networks may eventually learn to become explicitly goal-directed as well.

Are learning-based solutions enough?

Thus far, I have argued that progress toward capturing advanced cognitive abilities may come from structuring experience in extensions of contemporary transformer-based learning systems. However, we know that human goal-directed behavior depends on a network of interconnected brain regions in the frontal and parietal lobes that work together with all other brain areas to orchestrate overall behavior [11]. I agree with others who have argued for the importance of such control in artificial systems [12] and I believe that capturing these capabilities in our neural networks will be crucial. Future architectural innovations beyond the transformer may be necessary for this. My hunch is that there will be several such innovations as cognitive scientists and AI researchers seek to capture the level and extent of explicit goal-directed cognitive activity we often see in human problem-solving.

Concluding remarks

To go as far as thought leaders like Newton or others in discovering and solving novel scientific problems, I have argued that neural network-based learning systems must be immersed in goal-directed activities guided by the human-invented tools and practices of scientists and mathematicians. Building neural networks and systems for teaching them so that they can emulate these abilities provides an exciting challenge for the next 25 years.

Box 2. A puzzle illustrating explicit goal-directed thinking

The puzzle in Figure I illustrates several human abilities: to engage in explicit goal-directed reasoning subject to specified constraints, to formulate and use explicit rules and subgoals, and to communicate them to others. Grab a pencil, read the puzzle description, and go as far as you can in solving the puzzle before reading the further text below, and observe the thoughts that come to mind as you proceed!

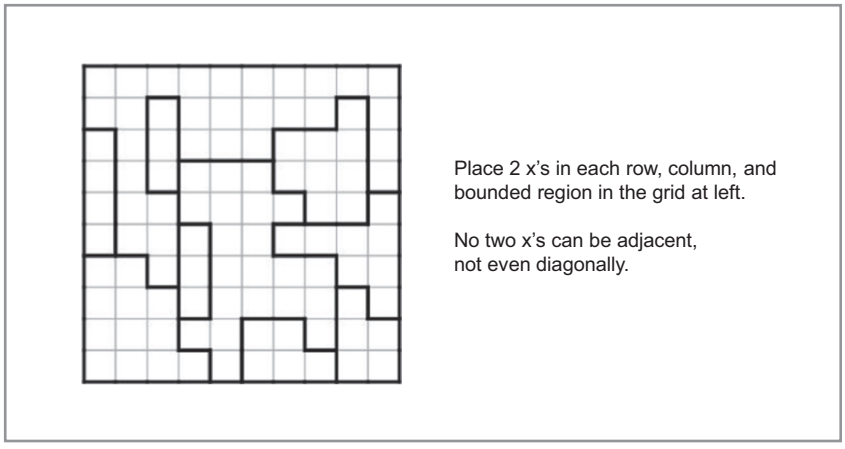

Figure I. A two-not-touch puzzle

On the first encounter with this puzzle, many people notice and can state explicitly, that xs must be placed at each end of each 3 × 1 bounded region. If asked why they know this, they note that these are the only two cells in the region that are not adjacent. One might also notice that the bounded region along the left edge of the grid must contain 2 xs and infer that no other cells in the same column could contain xs. One might formulate the abstract rule: if all the cells in a bounded region are in a single column, there can be no xs elsewhere in that column. One can also realize that it is possible to identify many cells in the grid where xs cannot go. One can then adopt the subgoal: find all of the cells that must not contain xs and mark them with a dot, and someone who has discovered this approach can communicate it to others in a statement like the one you have just read. You may have discovered these rules and subgoals for yourself, and even if not, you may be able to exploit them to find the solution to this puzzle. Puzzle reprinted from Bumgardner, J. (2020). Two Not Touch Puzzles, https://krazydad.com/twonottouch/, accessed January 3, 2021. Copyright © 2020 www.krazydad.com.

Declaration of interests

No interests are declared.

[1]Stanford University and DeepMind, Stanford, CA, 94305, USA

*Correspondence:
jlmcc@stanford.edu (J.L. McClelland).

https://doi.org/10.1016/j.tics.2022.09.018

References

1. Vaswani, A. *et al.* (2017) Attention is all you need. In *Advances in Neural Information Processing Systems 30. Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA*
2. McClelland, J.L. *et al.* (2020) Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proc. Natl. Acad. Sci. U. S. A.* 117, 25966–25974
3. Hersh, R. (1997) *What Is Mathematics, Really?* Oxford University Press
4. Burger, W.F. and Shaughnessy, J.M. (1986) Characterizing the van Hiele levels of development in geometry. *J. Res. Math. Educ.* 17, 31–48
5. Fodor, J.A. and Pylyshyn, Z.W. (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71
6. Lange, S. *Basic Mathematics*, Springer
7. Needham, T. *Visual Complex Analysis*, Oxford University Press
8. Shepard, R.N. (2008) The step to rationality: the efficacy of thought experiments in science, ethics, and free will. *Cogn. Sci.* 32, 3–35
9. Lewkowycz, A. *et al.* (2022) Solving quantitative reasoning problems with language models. *arXiv* Published online July 1, 2022. https://doi.org/10.48550/arXiv.2206.14858
10. Newell, A. *et al.* (1958) Elements of a theory of human problem solving. *Psychol. Rev.* 65, 151
11. Duncan, J. *et al.* (2020) Integrated intelligence from distributed brain activity. *Trends Cogn. Sci.* 24, 838–852
12. Russin, J. *et al.* (2020) Deep learning needs a prefrontal cortex. *Work Bridging AI Cogn. Sci.* 107, 603–616