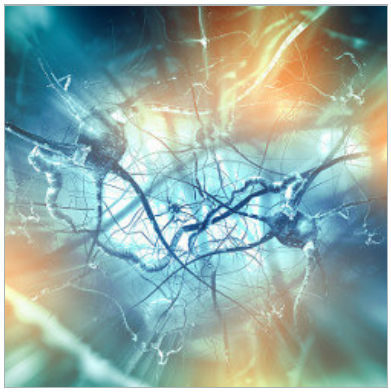
VIEWPOINT

Insights for AI from the Human Mind

By Gary Marcus, Ernest Davis

Communications of the ACM, January 2021, Vol. 64 No. 1, Pages 38-41
10.1145/3392663



Credit: Andrii Vodolazhskyi

What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single, perfect principle.
—Marvin Minsky, *The Society of Mind*

Artificial intelligence has recently beaten world champions in Go and poker and made extraordinary progress in domains such as machine translation, object classification, and speech recognition. However, most AI systems are extremely narrowly focused. AlphaGo, the champion Go player, does not know that the game is played by putting stones onto a board; it has no idea what a "stone" or a "board" is, and would need to be retrained from scratch if you presented it with a rectangular board rather than a square grid.

To build AIs able to comprehend open text or power general-purpose domestic robots, we need to go further. A good place to start is by looking at the human mind, which still far outstrips machines in comprehension and flexible thinking.

Here, we offer 11 clues drawn from the cognitive sciences—psychology, linguistics, and philosophy.

Back to Top

No Silver Bullets

All too often, people have propounded simple theories that allegedly explained all of human intelligence, from behaviorism to Bayesian inference to deep learning. But, quoting Firestone and Scholl,[4] "there is no one way the mind works, because the mind is not one thing. Instead, the mind has parts, and the different parts of the mind operate in different ways: Seeing a color works differently than planning a vacation, which works differently than understanding a sentence, moving a limb, remembering a fact, or feeling an emotion."

The human brain is enormously complex and diverse, with more than 150 distinctly identifiable brain areas, approximately 86 billion neurons, hundreds if not thousands of different types; trillions of synapses; and hundreds of distinct proteins within each individual synapse.

Truly intelligent and flexible systems are likely to be full of complexity, much like brains. Any theory that proposes to reduce intelligence down to a single principle—or a single "master algorithm"—is bound to fail.

Back to Top

Rich Internal Representations

Cognitive psychology often focuses on *internal representations*, such as beliefs, desires, and goals. Classical AI did likewise; for instance, to represent President Kennedy's famous 1963 visit to Berlin, one would add a set of facts such as part-of (Berlin, Germany), and visited (Kennedy, Berlin, June 1963). Knowledge consists in an accumulation of such representations, and inference is built on that bedrock; it is trivial on that foundation

to infer that Kennedy visited Germany.

Currently, deep learning tries to fudge this, with a bunch of vectors that capture a little bit of what's going on, in a rough sort of way, but that never directly represent propositions at all. There is no specific way to represent visited (Kennedy, Berlin, 1963) or part-of (Berlin, Germany); everything is just rough approximation. Deep learning currently struggles with inference and abstract reasoning because it is not geared toward representing precise factual knowledge in the first place. Once facts are fuzzy, it is difficult to get reasoning right. The much-hyped GPT-3 system[1] is a good example of this.[11] The related system BERT[3] is unable to reliably answer questions like "if you put two trophies on a table and add another, how many do you have?"[9]

Back to Top

Abstraction and Generalization

Much of what we know is fairly abstract. For instance, the relation "X is a sister of Y" holds between many different pairs of people: Malia is a sister of Sasha, Princess Anne is a sister of Prince Charles, and so on. We do not just know that particular pairs of people are sisters, we know what sisters are in general, and can apply that knowledge to individuals. If two people have the same parents, we can infer they are siblings. If we know that Laura was a daughter of Charles and Caroline and discover Mary was also their daughter, then we can infer Mary and Laura are sisters.

The representations that underlie cognitive models and common sense are built out of abstract relations, combined in complex structures. We can abstract just about anything: pieces of time ("10:35 PM"), pieces of space ("The North Pole"), particular events ("the assassination of Abraham Lincoln"), sociopolitical organizations ("the U.S. State Department"), and theoretical constructs ("syntax"), and use them in, an explanation, or a story, stripping complex situations down to their essentials, yielding enormous leverage in reasoning about the world.

Back to Top

Highly Structured Cognitive Systems

Marvin Minsky argued that we should view human cognition as a "society of mind," with dozens or hundreds of distinct "agents" each specialized for different kinds of tasks. For instance, drinking a cup of tea requires the interaction of a GRASPING agent, a BALANCING agent, a THIRST agent, and some number of MOVING agents. Much work in evolutionary and developmental psychology points in the same direction; the mind is not one thing, but many.

Much work in evolutionary and developmental psychology points in the same direction; the mind is not one thing, but many.

Ironically, that is almost the opposite of the current trend in machine learning, which favors end-to-end models that use a single homogeneous mechanism with little internal structure. An example is Nvidia's 2016 model of driving, which forsook classical modules like perception, prediction, and decision-making. Instead, it used a single, relatively uniform neural network that learned direct correlations between inputs (pixels) and one set of outputs (instructions for steering and acceleration).

Fans of this sort of thing point to the virtues of "jointly" training the entire system, rather than having to train modules separately. Why bother constructing separate modules when it is so much easier just to have one big network?

One issue is that such systems are difficult to debug and rarely have the flexibility that is needed. Nvidia's system typically worked well only for a few hours before intervention from human drivers, not thousands of hours (like Way-mo's more modular system). And whereas Waymo's system could navigate from point A to point B and deal with lane changes, all Nvidia's could do was to stick to a lane.

When the best AI researchers want to solve complex problems, they often use hybrid systems. Achieving victory in Go required the combination of deep learning, reinforcement learning, game tree search, and Monte Carlo search. Watson's victory in *Jeopardy!*, question-answering bots like Siri and Alexa, and Web search engines use "kitchen sink" approaches, integrating many different kinds of processes. Mao et al.[12] have shown how a system that integrates deep learning and symbolic techniques can yield good results for visual question answering and image-text retrieval. Marcus[10] discusses numerous different hybrid systems of this kind.

Back to Top

Multiple Tools for Simple Tasks

Even at a fine-grained scale, cognitive machinery often consists of many mechanisms. Take verbs and their past tense forms. In English and many other languages, some verbs form their past tense regularly, by means of a simple rule (walk-walked, talk-talked, perambulate-perambulated), while others form their past tense irregularly (*sing-sang, ring-rang, bring-brought, go-went*). Based on data from the errors that children make, one of us (Gary Marcus) and Steven Pinker argued for a hybrid model, a tiny bit of structure even at the micro level, in which regular verbs were generalized by rules, whereas irregular verbs were produced through an associative network.

Back to Top

Compositionality

The essence of language is, in Humboldt's phrase, "infinite use of finite means." With a finite brain and finite amount of linguistic data, we manage to create a grammar that allows us to say and understand an infinite range of sentences, in many cases by constructing larger sentences (like this one) out of smaller components, such as individual words and phrases. If we can say, *the sailor loved the girl*, we can use that as a constituent in a larger sentence (*Maria imagined that the sailor loved the girl*), which can serve as a constituent in a still larger sentence (*Chris wrote an essay about how Maria imagined that the sailor loved the girl*), and so on, each of which we can readily interpret.

At the opposite pole is the pioneering neural network researcher Geoff Hinton, who has been arguing that the meaning of sentences should be encoded in what he calls "thought vectors." However, the ideas expressed in sentences and the nuanced relationships between them are just way too complex to capture by simply grouping together sentences that ostensibly seem similar,[9,10] Systems built on that foundation can produce text that is grammatical, but show little understanding of what unfolds over time in the text they produce.

Back to Top

Top-Down and Bottom-Up Information, Integrated

Consider the image shown in Figure 1:[6] Is it a letter or a number? It could be either, depending on the context (see Figure 2). Cognitive psychologists often distinguish between *bottom-up information*, that comes directly from our senses, and *top-down knowledge*, which is our prior knowledge about the world (letters and numbers form distinct categories, words and numbers are composed from elements drawn from those categories, and so forth). An ambiguous symbol such as shown in the figures here looks one way in one context and different in another, as we integrate the light falling on our retina with a coherent picture of the world.

Figure 1. Possible number or letter.

Figure 2. Context-dependent interpretation.

Whatever we see and read, we integrate into a cognitive model of the situation and with our understanding of the world as a whole.

Back to Top

Concepts Embedded in Theories

In a classic experiment, the developmental psychologist Frank Keil[5] asked children whether a raccoon that underwent cosmetic surgery to look like a skunk, complete with "super smelly" stuff embedded, could become a skunk. The children were convinced the raccoon would remain a raccoon nonetheless, presumably as a consequence of their theory of biology, and the notion that it's what is inside a creature that really matters. (The children didn't extend the same theory to human-made artifacts, such as a coffeepot that was modified to become a bird feeder.)

Concepts embedded in theories are vital to effective learning. Suppose that a preschooler sees a photograph of an iguana for the first time. Almost immediately, the child will be able to recognize not only other photographs of iguanas, but also iguanas in videos and iguanas in real life, easily distinguishing them from kangaroos. Likewise, the child will be able to infer from general knowledge about animals that iguanas eat and breathe and that they are born small, grow, breed, and die.

No fact is an island. To succeed, a general intelligence will need to embed the facts that it acquires into richer overarching theories that help organize those facts.[13]

Back to Top

Causal Relations

As Judea Pearl[14] has emphasized, a rich understanding of causality is a ubiquitous and indispensable aspect of human cognition. If the world was simple, and we had full knowledge of everything, perhaps the only causality we would need would be physics. We could determine what affects what by running simulations; if I apply a force of so many micronewtons, what will happen next?

But that sort of detailed simulation is unrealistic; there are too many particles to track, and too little time, and our information is too imprecise.

Instead, we often use approximations; we know things are causally related, even if we don't know exactly why. We take aspirin, because we know it makes us feel better; we don't need to understand the biochemistry. We know that having sex can lead to babies and can act on that knowledge, even if we don't understand the exact mechanics of embryogenesis. Causal knowledge is everywhere, and it underlies much of what we do.

Back to Top

Tracking Individuals

As you go through daily life, you keep track of all kinds of individual objects, their properties and their histories. Your spouse used to work as a journalist. Your car has a dent on the trunk, and you replaced the transmission last year. Our experience is made up of entities that persist and change over time, and a lot of what we know is organized around those things, and their individual histories and idiosyncrasies.

Strangely, that is not a point of view that comes at all naturally to deep learning systems. For the most part, current deep learning systems focus on learning general, category-level associations, rather than facts about specific individuals. Without a notion something like a database record and an expressive representation of time and change, it is difficult to keep track of individual entities distinct from their categories.

Back to Top

Innate Knowledge

How much of the structure of the mind is built in, and how much of it is learned? The usual "nature versus nurture" contrast is a false dichotomy. The evidence from biology—from developmental psychology and developmental neuroscience—is overwhelming: nature and nurture work together.

Learning from an absolutely blank slate, as most machine-learning researchers aim to do, makes the game much more difficult than it should be. It is nurture without nature, when the most effective solution is obviously to combine the two. Humans are likely born understanding that the world consists of enduring objects that travel on connected paths in space and time, with a sense of geometry and quantity, and the basis of an intuitive psychology.

AI systems similarly should not try to learn everything from correlations between pixels and actions, but rather start with a core understanding of the world as a basis for developing richer models.[7]

Back to Top

Conclusion

The discoveries of the cognitive sciences can tell us a great deal in our quest to build artificial intelligence with the flexibility and generality of the human mind. Machines need not replicate the human mind, but a thorough understanding of the human mind may lead to major advances in AI.

In our view, the path forward should start with focused research on how to implement the core frameworks[15] of human knowledge: time, space, causality, and basic knowledge of physical objects and humans and their interactions. These should be embedded into an architecture that can be freely extended to every kind of knowledge, keeping always in mind the central tenets of abstraction, compositionality, and tracking of individuals.[10] We also need to develop powerful reasoning techniques that can deal with knowledge that is complex, uncertain, and incomplete and that can freely work both top-down and bottom-up,[16] and to connect these to perception, manipulation, and language, in order to build rich cognitive models of the world. The keystone will be to construct a kind of human-inspired learning system that leverages all the knowledge and cognitive abilities that the AI has; that incorporates what it learns into its prior knowledge; and that, like a child, voraciously learns from every possible source of information: interacting with the world, interacting with people, reading, watching videos, even being explicitly taught.

It's a tall order, but it's what has to be done.

Back to Top

References

1. Brown, T.B. et al. Language models are few-shot learners. (2020); arXiv preprint arXiv:2005.14165

2. Darwische, A. Human-level intelligence or animal-like abilities? *Commun. ACM 61*, 10 (Oct. 2018), 56–67.

3. Devlin, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-2019. (2019), 4171–4186.

4. Firestone, C. and Scholl, B.J. Cognition does not affect perception: Evaluating the evidence for 'top-down' effects. *Behavioral and Brain Sciences 39*, e229. (2016.)

5. Keil, F.C. Concepts, *Kinds, and Cognitive Development*. MIT Press, Cambridge, MA, 1992.

6. Lupyan, G. and Clark, A. Words and the world: Predictive coding and the language=perception-cognition interface. *Current Directions in Psychological Science 24*, 4 (2015), 279–284.

7. Marcus, G. Innateness, alphazero, and artificial intelligence. (2018); arXiv preprint arXiv:1801.05667).

8. Marcus, G. Deep Understanding: The Next Challenge for AI. *NeurIPS-2019* (2019).

9. Marcus, G. GPT-2 and the nature of intelligence. *The Gradient*. (Jan. 25, 2020).

10. Marcus, G. The next decade in AI: four steps towards robust artificial intelligence. (2020); arXiv preprint arXiv:2002.06177

11. Marcus, G. and Davis, E. GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *Technology Review* (Aug. 22, 2020).

12. Mao, J. et al. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. arXiv preprint arXiv:1904.12584.

13. Murphy, G. *The Big Book of Concepts*. MIT Press, 2002.

14. Pearl, J. and MacKenzie, D. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 2018.

15. Spelke, E. Initial knowledge: Six suggestions. *Cognition 50*, 1–3 (1994), 431–445.

16. van Harmelen, F., Lifschitz, V., and Porter, B., Eds. *The Handbook of Knowledge Representation*. Elsevier, Amsterdam, 2008.

Back to Top

Authors

Gary Marcus (gary.marcus@nyu.edu) is Founder and CEO of Robust.AI, and Professor Emeritus at NYU.

Ernest Davis (davise@cs.nyu.edu) is Professor of Computer Science at NYU. This Viewpoint is adapted from their new book, *Rebooting AI: Building Artificial Intelligence We Can Trust*.

Comments

Nandakumar Ramanathan

December 24, 2020 10:34

I'd like to bring to the notice of authors about a book entitled "Thinking, fast and slow" by Daniel Kahneman, Penguin Books, 2011 (ISBN: 978-1-846-14606-0). DK explains "the two systems that drive the way we think and make choices. One system is fast, intuitive and emotional; the other is slower, more deliberative and more logical." Both systems have their plus and minuses. I wonder how AI could be made to mimic human thinking in view of this dichotomy. R. Nandakumar (r_nand) aka Nandakumar Ramanathan.

Sathyanaraya Raghavachary

December 28, 2020 03:07

Nandakumar Ramanathan, the AI community is indeed aware of his work. He was a guest at a AAAI Fireside Chat event in April - https://vimeo.com/390814190, and last week, a speaker at #aidebate2: https://www.youtube.com/watch?v=VOI3Bb3p4GM

My own ("personal") view is this - there is no dichotomy, ie. no separate System1, System2 modes of processing for the brain to switch between; instead, it is a continuum, based on how much deliberate/conscious *attention* is needed in a situation. Familiar, routine (on account of past experience, practice...) situations are processed with very little attention, whereas novel, 'surprise', out-of-the-ordinary ones [for which there is no simple "lookup" based on the past] do need more attention.

Vinay Chaudhri

December 28, 2020 05:33

These insights reinforce the fundamentals and principles of AI that have been around for a long time. What is missing, however, is an articulation of a specific, measurable, research program that could be undertaken to make advances along one or more of these insights.

Displaying **all 3** comments