A Truly Intelligent Machine

Researchers are modeling AI on human brains. Can that teach us what intelligence is?

BY GEORGE MUSSER

THE DREAM OF ARTIFICIAL INTELLIGENCE has never been just to make a grandmaster-beating chess engine or a chatbot that tries to break up a marriage. It has been to hold a mirror to our own intelligence, that we might understand ourselves better. Researchers seek not simply artificial intelligence but artificial general intelligence, or AGI—a system with humanlike adaptability and creativity.

George Musser is a contributing editor at *Scientific American* and author of *Putting Ourselves Back in the Equation* (2023) and *Spooky Action at a Distance* (2015), both published by Farrar, Straus and Giroux. Follow him on Mastodon, Bluesky and Threads.

Large language models have acquired more problem-solving ability than most researchers expected they ever would. But they still make silly mistakes and lack the capacity for open-ended learning: once they are trained on books, blogs, and other material, their store of knowledge is frozen. They fail what Ben Goertzel of the AI company SingularityNET calls the "robot college student test": you can't put them through college (or indeed even nursery school).

The one piece of AGI these systems have unequivocally solved is language. They possess what experts call formal competence: they can parse any sentence you give them, even if it's fragmented or slangy, and respond in what might be termed Wikipedia Standard English. But they fail at the rest of thinking—everything that helps us deal with daily life. "We shouldn't expect them to be able to think," says neuroscientist Nancy Kanwisher of the Georgia Institute of Technology. "They're language processors." They skillfully manipulate words but have no access to reality other than through the text they have absorbed.

In a way, large language models mimic only the brain's language abilities, without the capacity for perception, memory, navigation, social judgments, and so forth. Our gray matter performs a bewildering mashup of overlapping functions, some widely distributed across the brain, others more localized. People who have suffered a stroke in one of their language areas may be unable to speak but may still be able to add numbers, compose symphonies, play chess and communicate by gestures as well as they could before. AI developers are incorporating such modularity into their systems in the hope of making them smarter.

OpenAI, the creator of the Generative Pre-trained Transformer (GPT), lets paid users select plug-ins to handle math, Internet search, and other kinds of queries. Each plug-in calls on some external bank of knowledge pertaining to its specialty. Further, and invisibly to users, the core language system may itself be modular in some sense. OpenAI has kept the specs under wraps, but many AI researchers theorize that GPT consists of as many as 16 separate neural networks, or "experts," that pool their answers to a query—although how they divide their labor is unclear. Last December Paris-based AI company Mistral made a big splash by releasing an open-source version of this "mixture of experts" architecture. The main advantage of this simple form of modularity is its computing efficiency: it is easier to train and run 16 smaller networks than a single big one. “Let’s get the best of both worlds,” says Edorado Ponti, an AI researcher at the University of Edinburgh. “Let’s get a system that has a high number of parameters while retaining the efficiency of a much smaller model.”

But modularity comes with trade-offs. No one is sure how brain regions work together to create a coherent self, let alone how a machine could mimic that. “How does information go from the language system to logical reasoning systems or to social reasoning systems?” wonders neuroscientist Anna Ivanova of M.I.T. “That is still an open question.”

One provocative hypothesis is that consciousness is the common ground. According to this idea, known as global workspace theory (GWT), consciousness is to the brain what a staff meeting is to a company: a place where modules can share information and ask for help. GWT is far from the only theory of consciousness out there, but it is of particular interest to AI researchers because it conjectures that consciousness is integral to high-level intelligence. To do simple or rehearsed tasks, the brain can run on autopilot, but novel or complicated ones—those beyond the scope of a single module—require us to be aware of what we’re doing.

Goertzel and others have incorporated a workspace into their AI systems. “I think the core ideas of the global workspace model are going to pop up in a lot of different forms,” he says. In devising electronic representations of this model, researchers are not seeking to make conscious machines; instead they are merely reproducing the hardware of a particular theory of consciousness to try to achieve a humanlike intelligence.

Could they inadvertently create a sentient being with feelings and motivations? It is conceivable, although even the inventor of GWT, Bernard Baars of the Neurosciences Institute in La Jolla, Calif., thinks it’s improbable. “Conscious computing is a hypothesis without a shred of evidence,” he says. But if developers do succeed in building an AGI, they could provide significant insight into the structure and process of intelligence itself.

GWT HAS LONG BEEN a case study of how neuroscience and AI research play off each other. The idea goes back to “Pandemonium,” an image-recognition system that computer scientist Oliver Selfridge proposed in the 1950s. He pictured the system’s modules as demons shrieking for attention in a Miltonian vision of hell. His contemporary Allen Newell preferred the more sedate metaphor of mathematicians solving problems together by gathering around a blackboard. These ideas were taken up by cognitive psychologists. In the 1980s Baars put forward GWT as a theory of human consciousness. “I learned a great deal from AI my whole career, basically because it was the only viable theoretical platform that we had,” he says.

No one is sure how brain regions work together to create a coherent self, let alone how a machine could mimic that. One hypothesis is that consciousness is the common ground.

Baars inspired computer scientist Stanley Franklin of the University of Memphis to try to build a conscious computer. Whether or not Franklin’s machine was truly conscious—Baars and Franklin themselves were dubious—it at least reproduced various quirks of human psychology. For instance, when its attention was drawn from one thing to another, it missed information, so it was just as bad at multitasking as people are. Starting in the 1990s, neuroscientists Stanislas Dehaene and Jean-Pierre Changeux of the Collège de France in Paris worked out what type of neuronal wiring might implement the workspace.

In this scheme, brain modules operate mostly independently, but every tenth of a second or so they have one of their staff meetings. It is a structured shouting contest. Each module has some information to offer, and the more confident it is in that information—the more closely a stimulus matches expectations, for example—the louder it shouts. Once a module prevails, the others quiet down for a moment, and the winner places its information into a set of common variables: the workspace. Other modules may or may not find the information useful; each must judge for itself. “You get this interesting process of cooperation and competition between subagents that each have a little piece of the solution,” Baars says.

Not only does the workspace let modules communicate with one another, but it provides a forum where they can collectively mull over information even when it is no longer being presented to the senses. “You can have some elements of reality—maybe a fleeting sensation and it’s gone, but in your workspace it continues to reverberate,” Dehaene says. This deliberative capacity is essential to solving problems that involve multiple steps or that stretch out over time. Dehaene has conducted psychology experiments in which he gave such problems to people in his laboratory, and he found they had to think them through consciously.

If the system sounds anarchist, that’s the point. It does away with a boss who delegates tasks among the modules because delegating is tough to get right. In mathematics, delegation—or allocating responsibilities among different actors to achieve optimal performance—falls into the category of so-called NP-hard problems, which can be prohibitively time-consuming to solve. In many approaches, such as the mixture-of-experts architecture thought to be used by OpenAI, a "gating" network doles out tasks, but it has to be trained along with the individual modules, and the training procedure can break down. For one thing, it suffers from what Ponti describes as a "chicken-and-egg problem": because the modules depend on the routing and the routing depends on the modules, training may go around in circles. Even when training succeeds, the routing mechanism is a black box whose workings are opaque.

In 2021 Manuel Blum and Lenore Blum, mathematicians and emeritus professors at Carnegie Mellon University, worked out the details of the battle for attention in the global workspace. They included a mechanism for ensuring that modules do not overstate their confidence in the information they are bringing in, preventing a few blowhards from taking over. The Blums, who are married, also suggested that modules can develop direct interconnections to bypass the workspace altogether. These side links would explain, for example, what happens when we learn to ride a bike or play an instrument. Once the modules collectively figure out which of them need to do what, they take the task offline. "It turns processing that goes through short-term memory into processing that's unconscious," Lenore Blum says.

Conscious attention is a scarce resource. The workspace doesn't have much room in it for information, so the winning module must be very selective in what it conveys to its fellow modules. That sounds like a design flaw. "Why would the brain have such a limit on how many things you can think about at the same time?" asks Yoshua Bengio, an AI researcher at the University of Montreal. But he thinks this constraint is a good thing: it enforces cognitive discipline. Unable to track the world in all its complexity, our brains have to identify the simple rules that underlie it. "This bottleneck forces us to come up with an understanding of how the world works," he says.

For Bengio, that is the crucial lesson of GWT for AI: today's artificial neural networks are too powerful for their own good. They have billions or trillions of parameters, enough to absorb vast swaths of the Internet, but tend to get caught up in the weeds and fail to extract the larger lessons from what they are exposed to. They might do better if their vast stores of knowledge had to pass through a narrow funnel somewhat like how our conscious minds operate.

BENGIO'S EFFORTS to incorporate a consciousnesslike bottleneck into AI systems began before he started thinking about GWT as such. In the early 2010s, impressed by how our brains can selectively concentrate on one piece of information and temporarily block out everything else, Bengio and his co-workers built an analogous filter into neural networks. For example, when a language model such as GPT encounters a pronoun, it needs to find the antecedent. It does so by highlighting the nearby nouns and graying out the other parts of speech. In effect, it "pays attention" to the key words needed to make sense of the text. The pronoun might also be associated with adjectives, verbs, and so on. Different parts of a network can pay attention to different word relations at the same time.

But Bengio found that this attention mechanism posed a subtle problem. Suppose the network neglected some words completely, which it would do by assigning zero value to the computational variables corresponding to those words. Such an abrupt change would throw a wrench into the standard procedure for training networks. Known as backpropagation, the procedure involves tracing the network's output back to the computations that produced it, so that if the output is wrong, you can figure out why. But you can't trace back through an abrupt change.

So Bengio and others devised a "soft-attention mechanism" whereby the network is selective but not overly so. It assigns numerical weights to the various options, such as which words the pronoun might be related to. Although some words are weighted more highly than others, all remain in play; the network never makes a hard choice. "You get 80 percent of this, 20 percent of that, and because these attention weights are continuous, you can actually do [calculus] and apply backprop," Bengio says. This soft-attention mechanism was the key innovation of the "transformer" architecture—the "T" in GPT.

In recent years Bengio has revisited this approach to create a more stringent bottleneck, which he thinks is important if networks are to achieve something approaching genuine understanding. A true global workspace must make a hard choice—it doesn't have room to keep track of all the options. In 2021 Bengio and his colleagues designed a "generative flow" network, which periodically selects one of the available options with a probability determined by the attention weights. Instead of relying on backpropagation alone, he trains the network to work in either the forward or the reverse direction. That way it can go back to fix any errors even if there is an abrupt change. In various experiments, Bengio has shown that this system develops higher-level representations of input data that parallel those our own brains acquire.

ANOTHER CHALLENGE of implementing a global workspace is hyperspecialization. Like professors in different university departments, the brain's various modules create mutually unintelligible jargons. The vision area comes up with abstractions that let it process input from the eyes. The auditory module develops representations that are suited to vibrations in the inner ear. So how do they communicate? They must find some kind of lingua franca or what Aristotle called common sense—the original meaning of that term. This need is especially pressing in the "multimodal" networks that tech companies have been introducing, which combine text with images and other forms of data.

In Dehaene and Changeux's version of GWT, the modules are linked by neurons that adjust their synapses to translate incoming data into the local vernacular. "They transform [the inputs] into their own code," Dehaene says. But the details are hazy. In fact, he hopes AI researchers who are trying to solve the analogous problem for artificial neural networks can provide some clues. "The workspace is more an idea; it's barely a theory. We're trying to make it a theory, but it's still vague—and the engineers have this remarkable talent to turn it into a working system," he says.

In 2021 Ryota Kanai, a neuroscientist and founder of the Tokyo-based AI company Araya, and another neuroscientist who has crossed over into AI, Rufin VanRullen of the University of Toulouse in France, suggested a way for artificial neural networks to perform the translation. They took their inspiration from language-translation systems such as Google Translate. These systems are one of the most impressive achievements of AI so far. They can do their job without being told, for example, that "love" in English means the same thing as "*amour*" in French. Rather they learn each language in isolation and then, through their mastery, deduce which word plays the same role in French that "love" does in English.

Suppose you train two neural networks on English and French. Each gleans the structure of its respective language, developing an internal representation known as a latent space. Essentially, it is a word cloud: a map of all the associations that words have in that language, built by placing similar words near one another and unrelated words farther apart. The cloud has a distinctive shape. In fact, it is the same shape for both languages because, for their all their differences, they ultimately refer to the same world. All you need to do is rotate the English and French word clouds until they align. You will find that "love" lines up with "*amour.*" "Without having a dictionary, by looking at the constellation of all the words embedded in the latent spaces for each language, you only have to find the right rotation to align all the dots," Kanai says.

Because the procedure can be applied to whole passages as well as single words, it can handle subtle shades of meaning and words that have no direct counterpart in the other language. A version of this method can translate between unrelated languages such as English and Chinese. It might even work on animal communication.

VanRullen and Kanai have argued that this procedure can translate not just among languages but also among different senses and modes of description. "You could create such a system by training an image-processing system and language-processing system independently, and then actually you can combine them together by aligning their latent spaces," Kanai says. As with language, translation is possible because the systems are basically referring to the same world. This insight is just what Dehaene was hoping for: an example of how AI research may provide insight into the workings of the brain. "Neuroscientists never have thought about this possibility of aligning latent spaces," Kanai says.

To see how these principles are being put into practice, Kanai—working with Arthur Juliani, now at Microsoft, and Shuntaro Sasai of Araya—studied the Perceiver model that Google DeepMind released in 2021. It was designed to fuse text, images, audio, and other data into a single common latent space; in 2022 Google incorporated it into a system that automatically writes descriptions for YouTube Shorts. The Araya team ran a series of experiments to probe Perceiver's workings and found that, though not deliberately designed to be a global workspace, it had the hallmarks of one: independent modules, a process for selecting among them and working memory—the workspace itself.

One particularly interesting implementation of workspacelike ideas is *AI People,* a forthcoming *Sims*-like game created by Prague-based AI company GoodAI. The version I saw last summer was set in a prison yard filled with convicts, corrupt guards and earnest psychiatrists, but the company also plans more peaceful scenarios. The game uses GPT as the characters' brains. It controls not just their dialogue but also their behavior and emotions so that they have some psychological depth; the system tracks whether a character is angry, sad or anxious and selects its actions accordingly. The developers added other modules—including a global workspace in the form of short-term memory—to give the characters a consistent psychology and let them take actions within the game environment. "The goal here is to use the large language model as an engine, because it's quite good, but then build long-term memory and some kind of cognitive architecture around it," says GoodAI founder Marek Rosa.

A POTENTIALLY GROUNDBREAKING advance in AI comes from researcher Yann LeCun of Meta. Although he does not directly cite the global workspace as inspiration, he has come by his own path to many of the same ideas while challenging the present hegemony of generative models—the "G" in GPT. "I'm advocating against a number of things that unfortunately are extremely popular at the moment in the AI/machine-learning community," LeCun says. "I'm telling people: abandon generative models."

Generative neural networks are so named because they generate new text and images based on what they have been exposed to. To do that, they have to be fastidious about detail: they must know how to spell each word in a sentence and place each pixel in an image. But intelligence is, if anything, the selective neglect of detail. So LeCun advocates that researchers go back to the now unfashionable technology of "discriminative" neural networks, such as those used in image recognition, so called because they can perceive differences among inputs—pictures of a dog versus a cat, for example. Such a network does not construct its

The quest for artificial general intelligence teaches us that tasks we find easy are computationally demanding, and the things we find hard, such as chess, are really the easy ones.

own image but merely processes an existing image to assign a label.

LeCun developed a special training regimen to make the discriminative network extract the essential features of text, images, and other data. It may not be able to autocomplete a sentence, but it creates abstract representations that, LeCun hopes, are analogous to those in our own heads. For instance, if you feed in a video of a car driving down the road, the representation should capture its make, model, color, position and velocity while omitting bumps in the asphalt surface, ripples on puddles, glints off blades of roadside grass—anything that our brains would neglect as unimportant unless we were specifically watching for it. "All of those irrelevant details are eliminated," he says.

Those streamlined representations are not useful on their own, but they enable a range of cognitive functions that will be essential to AGI. LeCun embeds the discriminative network in a larger system, making it one module of a brainlike architecture that includes key features of GWT, such as a short-term memory and a "configurator" to coordinate the modules and determine the workflow. For instance, the system can plan. "I was very much inspired by very basic things that are known about psychology," LeCun says. Just as the human brain can run thought experiments, imagining how someone would feel in different situations, the configurator will run the discriminative network multiple times, going down a list of hypothetical actions to find the one that will achieve the desired outcome.

LeCun says he generally prefers to avoid drawing conclusions about consciousness, but he offers what he calls a "folk theory" that consciousness is the working of the configurator, which plays roughly the role in his model that the workspace does in Baars's theory.

IF RESEARCHERS SUCCEEDED in building a true global workspace into AI systems, would that make them conscious? Dehaene thinks it would, at least if combined with a capacity for self-monitoring. But Baars is skeptical, in part because he is still not entirely convinced by his own theory. "I'm constantly doubting whether GWT is really all that good," he says. To his mind, consciousness is a biological function that is specific to our makeup as living beings. Franklin expressed a similar skepticism when I interviewed him several years ago. (He passed away last year.) He argued that the global workspace is evolution's answer to the body's needs. Through consciousness, the brain learns from experience and solves the complex problems of survival quickly. Those capacities, he suggested, aren't relevant to the kinds of problems that AI is typically applied to. "You have to have an autonomous agent with a real mind and a control structure for it," he told me. "That agent has got to have kind of a life—it doesn't mean it can't be a robot, but it's got to have had some sort of development. It's not going to come into the world full-blown."

Anil Seth, a neuroscientist at the University of Sussex in England, agrees with these sentiments. "Consciousness is not a matter of being smart," he says. "It's equally a matter of being alive. However smart they are, general-purpose AIs, if they're not alive, are unlikely to be conscious."

Rather than endorsing GWT, Seth subscribes to a theory of consciousness known as predictive processing, by which a conscious being seeks to predict what will happen to it so it can be ready. "Understanding conscious selfhood starts from understanding predictive models of the control of the body," he says. Seth has also studied integrated information theory, which associates consciousness not with the brain's function but with its complex networked structure. By this theory, consciousness is not integral to intelligence but might have arisen for reasons of biological efficiency.

AI is an ideas-rich field at the moment, and engineers have plenty of leads to follow up already without having to import more from neuroscience. "They're killing it," notes neuroscientist Nikolaus Kriegeskorte of Columbia University. But the brain is still an existence proof for generalized intelligence and, for now, the best model that AI researchers have. "The human brain has certain tricks up its sleeve that engineering hasn't conquered yet," Kriegeskorte says.

The quest for AGI over the past several decades has taught us much about our own intelligence. We now realize that tasks we find easy, such as visual recognition, are computationally demanding, and the things we find hard, such as math and chess, are really the easy ones. We also realize that brains need very little inborn knowledge; they learn by experience almost everything they need to know. And now, through the importance of modularity, we are confirming the old wisdom that there isn't any one thing called intelligence. It is a toolbox of abilities—from juggling abstractions to navigating social complexities to being attuned to sights and sounds. As Goertzel notes, by mixing and matching these diverse skills, our brains can triumph in realms we've never encountered before. We create novel genres of music and solve scientific puzzles that earlier generations couldn't even formulate. We step into the unknown—and one day our artificial cousins may take that step with us. ⬢