Artificial Intelligence, Modularity and the Global Workspace

Chad DeChant
Psychology 4270
2017

The field of artificial intelligence (AI) has seen a great deal of progress in recent years. Computers can now perform as well as people on a range of tasks, including classifying images[1], transcribing speech[2], analyzing medical imagery[3], and playing games such as chess and Go[4]. Autonomous vehicles have driven millions of miles on public roads.[5] Despite these successes, however, the field is facing a number of challenges, including how to make AI safe, more interpretable, and able to deal with more complicated situations and reason in more sophisticated ways. It is not clear that the solutions to these challenges can come from merely improving current AI techniques. After briefly reviewing the current state of AI and the obstacles it faces, we will look at some theories and findings from Psychology that may be able to inspire new approaches within AI.

It is particularly appropriate to look to our understanding of human cognition for such inspiration because AI's recent successes are due to the use of neural networks, so-called because they are (very) loosely modeled after the functioning of collections of neurons. Neural networks are a form of machine learning, in which machine algorithms learn how to do tasks by being shown examples. Before the adoption of neural networks, traditional artificial intelligence required experts to write down rules that a computer program could follow to carry out a task. For example, linguists put together very complex systems to parse sentences and perform translation based on grammatical rules, including checks to catch the numerous exceptions to such rules. Machine learning techniques such as neural networks improved on that approach by feeding a program a huge number of examples of sample inputs and desired outputs in order to train it to derive its own rules for how to produce the desired output.

When computers carried out explicit instructions carefully crafted by experts, it was relatively easy to interpret their output and understand why they made errors and at what point in

their operation the errors cropped up. It is much less easy in the era of machine learning when an algorithm makes a decision based on thousands of learned parameters of a neural network, the meaning of which is opaque even to its designer. A program running an autonomous car might receive as input a series of images from the car's cameras and output a change in the steering wheel's direction but give no clue as to how it worked out the long chain of reasoning from input to output. This ability of machine learning algorithms to go from raw input all the way to output, called 'end-to-end learning', is regarded as one of the great successes of recent work in AI and more and more problems are being tackled in this way. Researchers are only lately coming to realize that there is a downside to this approach: by making the intermediate steps of reasoning inscrutable, it is difficult or impossible to interpret the results or understand why mistakes are made.

Interpretability is just one of the challenges posed by current end-to-end trained neural networks. Closely related are concerns about safety, for if we cannot understand the workings of a machine learning system, it is difficult to ensure its safe operation either before it is deployed or in real-time monitoring as it is working. If it does something dangerous, it would have to be taken offline and completely retrained. This introduces the third challenge, the inability to modify or upgrade just one aspect of an end-to-end trained neural network. Discrete changes to an already trained and deployed system are generally not possible, making it difficult to perform updates or corrections. The development of new systems is also impaired by this limitation, which significantly slows down the testing of potential improvements. In part because of this, the tasks that neural networks and other machine learning techniques have been applied to have remained relatively simple.

Perhaps the most difficult challenge facing AI is discovering how to create more

complex systems that can perform more complicated tasks and engage in more sophisticated reasoning than is currently possible. Neural networks tend to be trained to do one thing, in one domain — for example, speech recognition or translating text from one language to another. A small number of more sophisticated systems have been made that combine two domains, as is the case in a neural network that answers questions about images. Even AI agents or robots that have to take action in a real or simulated environment typically plan their actions based on only one kind of input, usually visual, and output relatively simple commands such as direction and speed of movement. It is not clear how more complex behavior can be learned using methods currently popular within the AI research community. Fresh perspectives and methods will be required.

These challenges could begin to be addressed if AI research underwent a paradigm shift from all or nothing end-to-end training to a more modular approach. There are many theories of modularity in cognitive science, some of which are more widely accepted than others. The simplest version would simply say that the mind or brain is at least partly made up of distinct functional components which in some respects can operate independently of other components. So far this ought to be relatively uncontroversial since it is widely accepted that the brain has some specialized regions devoted to particular functions, such as those which process sensory input. Indeed one of the early key writers on modularity, Jerry Fodor, believed that only what he called the "input systems" which process visual, auditory, and other perceptual inputs in particular domains could be modular.[6] He argued that what he called the "central systems," involving higher level reasoning capabilities which can reason across several domains, cannot be modular. In large part this was due to his very stringent and somewhat limiting definition of modularity, in place of which we will prefer that put forward by Peter Carruthers. Biological systems are messy and evolved over time so modules may be quite diverse in their attributes. Carruthers's characterization of modules is

therefore intentionally loser and more flexible.

Carruthers affirms that modules should be thought of as different processing systems with their own functions or set of functions and their own place or places, perhaps widely dispersed, within the brain. Some modules might be completely self-sufficient or encapsulated, consulting no other modules during their operation, while others might need to tap into the knowledge contained in other modules. Modules can be composed of other modules and so the intermediate outputs between these submodules might be available to yet other modules.[7] This potential for hierarchical assemblies of modules is cited by advocates for modularity as one reason why modularity of mind is biologically plausible because such hierarchies are common throughout biology.

Can we draw any further conclusions about modularity from principles of biology or evolution? Modularity theorists believe that we can, pointing out that it is widely accepted that biological organisms are organized along modular lines. Biological modularity manifests itself in a variety of ways. During embryonic development, some parts can develop independent of context. For example, small errors in gene expression have led to fully formed, somewhat functional eyes sprouting on the wings of flies.[8] After development, parts of organisms with distinct functional roles have varying degrees of independence, including the ability to regulate their own physiological processes.[9] From organelles to cells to organs to parts of the metabolic network,[10] it is apparent that modular structure and substructure is common in biology. This is not by itself evidence that the mind is also organized in a modular fashion, but it does suggest that, as the mind/brain evolved along with the rest of an organism, it is likely that the same considerations which led to modularity in the organism as a whole might have done so as well for the mind.

Many have argued that in biological systems that evolve over time through natural

selection, modularity is a required organizing principle or, at the very least, is very advantageous and so is commonly selected for. In order for natural selection to exert positive or negative selection pressure on a functional capability, that capability has to be expressed in a way that is independent of other traits in an organism.[11] And modularity fits in some ways with how we understand evolution to operate: typically unable to completely and elegantly redesign a system in response to a new environmental condition, it usually has to "bolt on" a new capability or trait to the existing set of capabilities.[12] An error in copying may lead the genes for a particular functional trait or module to be copied an extra number of times and those copies might be mutated in some way, leading to an additional functional module.[13] Further, existing modules might be recombined in different ways that lead to additional capabilities. One of the debates surrounding modularity is the question of how many modules there might be, whether a small or 'massive' number. It seems likely that once a system has a modular architecture at all, it will tend to continue to accumulate a large number of additional modules through these duplication and modification procedures. Related to and just as important as the question of how many modules there might be that make up the mind are the questions of what kind of modules there are and how these modules interact with each other.

We saw earlier that Jerry Fodor thought that only peripheral "input" processes such as those deriving from the senses could be modular. The more interesting, complex, and characteristically human "central" reasoning processes make use of the peripheral modules but are themselves not modular, according to Fodor. He restricts which mental processes could be modular by insisting that modules have to be "encapsulated," unable to access information outside of them other than what they receive as input. Further, modules are restricted to particular stimulus domains and are not able to accept input outside their individual domain. Thinking done only with

such encapsulated modules would be quite inflexible. Even if some modules could communicate by receiving as input the output from other modules, according to Fodor such communication would be routed through rigid, predetermined "pipelines." The creativity evident in human thought, including the use of analogies to very different domains when reasoning, would apparently not be possible in such a system. But since the encapsulation of modules seems to be at the root of the inability of Fodor's "central" processes to be modular, why not just make them unencapsulated?

Fodor believes that if the central reasoning processes were unencapsulated modules they would be called upon to do unrealistic — even impossible — computations, including searching through all of an agent's beliefs before acting. This so-called Frame Problem was most memorably described by Daniel Dennett, who imagined a bomb disposal robot's attempt to remove a bomb from a room before it explodes.[14] Before the robot can act it must work out all of the consequences of its actions, including the countless number of irrelevant implications of those actions. It might work out that pulling the bomb out of the room would not change the color of the paint on the room's walls, for example. It would not know whether its action would affect its beliefs or some part of the state of the world until it had systematically gone through all of its beliefs and possible effects on the world. Of course in Dennett's story the robot and the room are blown up before long; the computation required could go on indefinitely, making the task either impossible or merely practically infeasible. Fodor calls this "Hamlet's problem: when to stop thinking"[15] and concludes from the fact that the necessary thinking would be computationally impossible or impractical that the mind's central processes cannot be computations of any kind.[16]

The process the bomb disposal robot went through to check all the possible outcomes of its actions was considered computational infeasible because it checked all possible implications

one after another, in serial fashion. If this checking could be done in a massively parallel fashion, however, it might no longer be so difficult or time-consuming. Instead of one module or process painstakingly testing one possible effect after another, an action planning module might share a description of the intended course of action with a large number of other modules which could check the consequences within their particular, narrow purview. These specialized modules could operate quietly in the background, only calling attention to themselves should they find anything that might be relevant to the original module's calculations. Findings about unexpected deadly explosions would be returned to the querying module; non-effects on paint color would not. This simplified example, due to Shanahan and Baars, illustrates how a process operating in a serial fashion could draw on the resources of many other modules operating in parallel. Its generalization, the global workspace, could help resolve Hamlet's problem, and Fodor's. [17]

Global workspace theory, as originally proposed by Bernard Baars, is a theory of consciousness.[18] Baars wanted to explain what he called "the central puzzle," how to reconcile the limited capacity of conscious thought with the "vast" capacity of the brain to engage in tasks which may be mostly or entirely unconscious.[19] Most of the activity of the brain's modules or processes occurs below the level of consciousness, each module carrying out its own task unbeknownst to the other modules. But sometimes the output of a module is globally broadcast, making it available to the other modules. Baars and other advocates of global workspace theory contend that it is this global availability that makes something conscious. Baars often employs the metaphor of a theater to illustrate his theory: a large audience plays the role of the many unconscious processes observing what enters the global workspace while one or a small handful of audience members called on to the stage to perform for the others represents the broadcast of one or a small number of modules' outputs to the rest of the brain's modules. Similarly, Dennett has called consciousness "fame in

the brain."*

Global broadcast is not the only way that modules can interact. On the contrary, they must do so quite often because much of the work they do unconsciously requires the involvement of diverse processes, at least some of which will be found in other modules. Simply standing up from a chair would require the coordination of modules involving vision, balance, muscle movement, etc. Stanislas Dehaene and Lionel Naccache suggest that at least four categories of modules must participate in the global workspace, namely those involved in perception, movement, long-term memory, and evaluation of something as positive or negative, as well as attention mechanisms.[20] Attention is regarded by many, including Carruthers, as the key to selecting otherwise unconscious modular outputs and making them conscious by globally broadcasting them.[21] There may be (at least) two attention systems at work. A bottom-up attention mechanism is thought to monitor sense perceptions, bringing attention to important or relevant sensory data. A top-down attention mechanism maintains current goals and works to focus attention on perceptions and other modular outputs which are relevant to these. In a few cases, involving very startling stimuli like loud noises, attention may not be under someone's control. But Carruthers, at least, proposes that in almost all cases, the direction of top-down and bottom-up attention should be thought of as action. In part this is because the same region of the brain (the frontal eye-fields) plays an important role in both these kinds of attention mechanisms as well as in physically moving the eyes.[22] That attention is a form of action suggests that, like other forms of action, its performance can be either good or poor and could be improved.

Attention also plays a key role in working memory. The same brain regions involved in

* Without going into the details of Ned Block's theories on consciousness, we will briefly note that some thinkers, such as Carruthers, have claimed that states that are access-conscious are those that are globally broadcast and that phenomenal consciousness is also "co-extensive with" global broadcasting.

top-down attention are also at work in working memory tasks.[23] According to Baars, everything in the global workspace is in working memory (along with other things), while Carruthers goes further in identifying the two. According to him, working memory is the global workspace where the stream of consciousness is experienced and conscious reflection happens.[24] If the global workspace depends on and in some sense happens within working memory, its serial nature and limited capacity become more understandable, for the capacity of working memory is known to be quite limited. The limits of working memory, in turn, might play a significant role in the limits of human intelligence. Measuring intelligence, particularly along one scale, is clearly controversial and not obviously fruitful. However, some — including Carruthers — claim that the measure known as fluid *g* is meaningful and that it is highly correlated with working memory capacity, with various studies finding correlations from 0.6 to nearly 1.0. Carruthers suggests that almost all the variance in fluid *g* among people can be explained by working memory capacity along with speed of processing.[25] What does this mean for speculation about the future intelligence of machines? Computers will certainly be able to compute faster than humans do, as neurons operate quite slowly even in comparison with today's digital technology, so they will certainly beat us on speed of processing. Would machines have similar hard upper limits on working memory? It is hard to see why, suggesting that it is at least possible that AI could be, in some meaningful way, more intelligent than humans. If consciousness happens within working memory and machines had greater working memory abilities, what, if anything, would that mean about their being conscious?

So far we have laid out the basics of a modular theory of mind and a global workspace within which modules can interact and will shortly describe how what we have covered could be applied to AI. But we should pause to note that the global workspace theory is conceived by its proponents as explicitly a theory about consciousness. Are we proposing conscious AI? Not

necessarily. Though it seems quite likely that consciousness, as a natural phenomenon found in evolved, biological creatures such as humans, contains nothing magical which would prevent it from one day also manifesting in artificial entities, it is almost equally likely that the mechanisms underlying it are not yet understood and that it requires more than organizing an AI program along modular lines with an analogue to the global workspace (even if we threw in metacognition, despite recent conjectures in *Science* that having metacognition and a global workspace might be sufficient for consciousness in machines[26]). However, it is believed that consciousness aids in — or is even required for — certain mental operations, so a proposal to structure AI on lines analogous to the mechanisms that underlie consciousness is best understood more conservatively as a proposal to allow for deeper, more flexible, and more human-like cognitive processes. It has been suggested, for example, that consciousness is required for integrating several streams of evidence to come to a decision and then maintaining that decision in mind through subsequent action steps.[27] Others have suggested that the slow, deliberate reasoning typical of so-called System Two thinking, perhaps following explicitly learned rules of inference, can only be done consciously.[28] Spontaneously generating intentional behavior and combining mental operations in a novel way to address a new or unusual task have also been given as examples of actions requiring consciousness.[29] Once a novel action or sequence of actions has been carried out and learned, it is possible for it to become automated in a way that does not require it to occupy conscious attention. In that case it would become one of the many unconscious processes or modules that we have suggested make up the mind.

Introducing modularity would by itself represent an improvement over current standard practices within the AI research community, where the seamless, integrated, end-to-end architectures described earlier have become standard. A modular system would be much more

easily updated than a unified system without parts. Just as modularity of the mind was thought to be required or at least advantageous for its evolvability, so would a modular AI architecture allow for much more rapid testing of new ideas and techniques. It would greatly speed up research by allowing a much larger community of researchers to effectively cooperate in their efforts.

By its nature modularity would introduce break points where modular outputs could be checked before being passed on to other modules, which would be beneficial in several ways. First, these would introduce the possibility of greater interpretability because there would at least be standardized places to examine intermediate outputs. Actually achieving greater interpretability would require additional work, of course, but this would be facilitated by the modular architecture as it would allow modules devoted to interpretation to be trained alongside the other modules. It is worth noting that some thinkers in cognitive science, including Carruthers, believe that much of our own mind is essentially opaque to us, including our intentions, goals, and values, and we have to learn to interpret ourselves as we learn to interpret others, so some of his and others' work in this direction would be useful. It has been suggested by some, including Carruthers at an earlier stage of his writing, that the language module is the primary interface between unrelated modules and that natural language is the primary way of combining the output from such modules. Natural language would certainly be the most easily interpretable description for what goes on in a modular system, if it could be made to work.

Modularity would help address some of the safety concerns around AI mentioned earlier in a few ways. First, it would allow for testing of isolated components, potentially making them less likely to behave unexpectedly during use. Second, the intermediate outputs discussed above would allow for better real-time monitoring of the internal behavior of AI systems, which are now essentially black boxes. Third, if unsafe behavior were detected in any module, a replacement

module could be more quickly trained than an entirely new system, and swapping the modules should be very rapid and easy.

Learning the dynamics of the global workspace will be a challenge for any AI system. Carruthers's conception of the attention mechanism which brings something into the global workspace as an action is potentially helpful, because AI researchers have many resources for modeling and training actions, such as reinforcement learning. Other writers have different ideas on how modules' output reaches the global workspace, including notions of competition for attention. The fact that there is so much disagreement within Psychology about the nature (or even existence) of modules is a boon to AI researchers, who will be able to draw on all the different theories. No doubt nature settled on just one (or none) of them, and it would of course be best to be guided by the correct theory. But nature may not have found the only means to creating intelligent agents, and it should be beneficial to try diverse approaches.

We suggested earlier that introducing modularity, as well as analogues of the global workspace, would have a beneficial effect on AI. But perhaps it would be better to say *reintroducing*, for several of the works mentioned in this paper referenced what they took to be "lessons from AI," which were often about the value of modularity and even about an approach similar to the global workspace but which predated Baars by a decade.[30] Sadly, the lessons of older approaches to AI which they were drawing upon have been all but lost in today's mainstream AI community, focused as it is on machine learning techniques such as neural networks. In order to continue to make progress, old lessons will have to be relearned, a prospect made much easier and more likely by the work done in the intervening years within cognitive science.

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034).

[2] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... & Zweig, G. (2016). The Microsoft 2016 conversational speech recognition system. *arXiv preprint arXiv:1609.03528*.

[3] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Lungren, M. P. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225*.

[4] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354.

[5]Lumb, D. Waymo's autonomous cars have driven 4 million miles. *Engadet.com* https://www.engadget.com/2017/11/27/waymo-autonomous-cars-drove-4-million-miles/ accessed December 12, 2017.

[6] Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT press.

[7] Carruthers, P. (2006). *The architecture of the mind*. Oxford University Press.

[8] Wagner, G. P., & Altenberg, L. (1996). Perspective: complex adaptations and the evolution of evolvability. *Evolution*, *50*(3), 967-976.

[9] Wagner, G. P., Mezey, J., Calabretta, R., Wagner, G. P., Mezey, J., & Calabretta, R. (2001). Natural selection and the origin of modules. In *Modularity: Understanding the development and evolution of complex natural systems*. MIT Press.

[10] Rohwer, J. M., Schuster, S., & Westerhoff, H. V. (1996). How to recognize monofunctional units in a metabolic system. *Journal of theoretical biology*, *179*(3), 213-228.

[11] Clune, J., Mouret, J. B., & Lipson, H. (2013, March). The evolutionary origins of modularity. In *Proc. R. Soc. B* (Vol. 280, No. 1755, p. 20122863). The Royal Society.

[12] Carruthers, P. (2005). Distinctively human thinking. *The innate mind: Structure and contents*, 69.

[13] Marcus, G. F. (2005). What developmental biology can tell us about innateness. *The innate mind: structure and contents*, 23.

[14] Dennett, D. (1998). Cognitive Wheels'(1984). *Brainchildren, Essays on Designing Minds*.

[15] Fodor, J. A. (1983), p. 140.

[16] Fodor, J. A. (2001). *The mind doesn't work that way: The scope and limits of computational psychology*. MIT press.

[17] Shanahan, M., & Baars, B. (2005). Applying global workspace theory to the frame problem. *Cognition*, *98*(2), 157-176.

[18] Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

[19] Baars, B. J. (2007). The global workspace theory of consciousness. *The Blackwell companion to consciousness*, 227-242.

[20] Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1), 1-37.

[21] Carruthers, P. (2015). *The centered mind: what the science of working memory shows us about the nature of human thought*. OUP Oxford, p. 58.

[22] Carruthers, P. (2015), p. 149.

[23] Carruthers, P. (2015), p. 89.

[24] Carruthers, P. (2015), p. 75.

[25] Carruthers, P. (2015), p. 121.

[26] Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it?. *Science*, *358*(6362), 486-492.

[27] Dehaene, S. et al (2017), p. 489.

[28] Carruthers, P. (2015), p. 174.

[29] Dehaene, S. and Naccache, L. (2001), p.10 - 11.

[30] Nii, H. Penny. "The blackboard model of problem solving and the evolution of blackboard architectures." *AI magazine* 7.2 (1986): 38.