The Tale of Two Intelligence Fields: AI and Neuroscience

A Report on the 2022 FENS Conference Debate on *Building and understanding brains: How can AI research inform neuroscience?*

Zahraa Chorghay August 7, 2022

In April 2022, DALL-E 2 was launched to immense popularity: it is an artificial intelligence (AI) model that creates an image corresponding to a phrase that the user provides. Prompted with phrases like "impressionist painting of a robot neuroscientist" or "a dragon perched on a snowy mountain," DALL-E 2 creates realistic images with interesting compositionality. Notably, these features were not explicitly designed into the generative machine learning model, and these emergent properties have been hitherto uniquely attributed to human creativity. This raises the question of what differentiates human, and more broadly, biological intelligence, from AI technologies like DALL-E 2, AlphaGo, and others? What do these similarities or differences tell us about the nature of intelligence, and how can the two communities of AI and neuroscience (which studies biological intelligence) benefit from understanding how artificial and biological intelligence intersect with each other?

These questions formed the basis of the *Building and understanding brains: How can AI research inform neuroscience?* debate held on July 9, 2022, as part of the Federation of European Neuroscience Societies (FENS) conference in Paris. The debate was moderated by Christopher Summerfield, Professor at the University of Oxford, and featured prominent panellists, in order of their presentations: Jane Wang, Staff Research Scientist at DeepMind; Kanaka Rajan, Assistant Professor at Mount Sinai; Claudia Clopath, Professor at Imperial College London; Andrew Saxe, Sir Henry Dale Fellow and Associate Professor at University College London; and Stanislas Dehaene, Director of NeuroSpin and professor at College de France.

Figure 1 Panellists at the FENS conducting the debate on Building and understanding brains: How can AI research inform neuroscience? Left to right: Stanislas Dehaene, Christopher Summerfield, Kanaka Rajan and Andrew Saxe. Not pictured: Claudia Clopath, Jane Wang. Attending virtually: Jane Wang. Photograph ©Bartosch Salmansk.

AI and neuroscience have different goals

Wang began the discussion by highlighting that AI and neuroscience have traditionally held different goals, which can likely be attributed to the vastly different nature of its agents. Biological agents are fully formed, with priors imbued by their evolutionary history, and further amenable to learning through experience. In contrast, artificial agents must be trained from scratch, such that the training and task structure often determines their behaviour. Neuroscientists thus focus on understanding the biological mechanisms underlying intelligent, while AI researchers develop engineering algorithms for optimised task performance.

Identifying areas where artificial and biological agents can learn from each other

Dehaene noted what he considers to be the major AI achievements: AI has harnessed a central dogma of neuroscience that learning occurs through synaptic changes, by creating algorithms where synaptic weight adjustments facilitate learning and task performance. AI models can thus successfully recapitulate the early stages of visual processing and certain aspects of language processing. However, Dehaene maintains the position that “brains are better than any machines for now.” Unlike artificial agents and even other biological species, humans have an ability for language and composition, which also aids sharing information or knowledge transfer within the species. How can we train AI models to show such compositional representations? Humans are thought to rely on symbolic models, which use discrete properties (e.g., parallel lines, right angles, etc.) rather than continuous variables to make decisions. Such representations could underly the superior performance of humans compared to AI in the examples that Dehaene presented, including identifying outliers and classifying certain images.

One- or few-shot learning, and the importance of data versus fine scale architecture for learning

Of note, humans can learn from a small number of examples, sometimes even a single trial, using Bayesian reasoning. This optimal inference is performed even by young children, which Dehaene suggested indicates a type of intelligence that emerges intrinsically in humans rather than is formed during development. However, it is possible that future AI systems could have such capabilities. Wang highlighted recent work showing few-shot context learning can also occur in largescale AI models trained with large enough datasets that reflect the data distribution of the real world. Understanding the priors that artificial models are operating with is an emerging neuroscience-inspired AI subfield.

This leads to the question of whether the right kind of data is sufficient for task performance, or whether the fine scale architecture still matters? Neuroscientists study precise neural connectivity throughout development and the role of different cell types in computations, among other biological details of brain structure. However, Wang’s experience with recent largescale models suggests a relative lack of importance of these details compared to the properties of the training data. A key difference between brains and artificial systems is that AI does not face the strict resource constraints that biology has to optimize for. Dehaene mentioned the wiring efficiency hypothesis which posits that biological brains are wired as they are to minimize resource loss. In contrast, AI does not rely on the restricted resources of its immediate environment, so it has different architectural constraints than that of brains.

Continual learning, and the importance of data versus temporal aspects for learning

Another domain in which humans outperform AI is continual learning, where the agent sequentially learns a number of tasks. Unlike humans and other animals who do this throughout their lifetime with relative success, algorithms suffer from catastrophic forgetting, i.e., learning a new task diminishes their performance on a previously learned task. Clopath points out that incorporating biological mechanisms into AI — including synaptic consolidation resulting from synaptic plasticity over different timescales, systems consolidation where memories from one brain region are transferred to another (e.g., from the hippocampus to the cortex), and replay during sleep — could prevent forgetting and thus improve performance on continual learning paradigms. Furthermore, comparing the performance of such biologically inspired algorithms to current AI setups can serve as a bed for hypothesis testing to assess which of the potential neural mechanisms identified by neuroscientists actually enhance learning.

With regards to Clopath's inclusion of timescales, Summerfield placed the spotlight on the importance of temporal aspects of learning. Rajan pointed out that the sampling rate, which is one of the three considerations of learning algorithms alongside the architecture and the learning rule, could formalize the effect of time on performance. Additionally, temporal learning can be affected by both the duration and sequence of information, so the precise structure of the presented data might matter.

AI can also outperform humans

Conversely, there are some domains in which AI has surpassed human performance. Machine learning has been successfully used in various domains to organise and analyse large amounts of data. Furthermore, AI can supersede the priors that human brains have developed through evolution as a way to conserve resources. For example, DeepMind's Flamingo performs well on the Stroop Test because it does not need to have much more efficient word recognition compared to colour recognition like humans do in their environment. Clopath noted that while computational neuroscientists aim to incorporate biologically plausible aspects and mechanistic insights into their models, AI researchers have often found successful functional outcomes, i.e., better model performance, by veering away from such biological constraints and taking advantage of the large computing power that is now available.

Approaching theories of intelligence

Rajan and Saxe both raised questions of how we can gain a coherent understanding of intelligence. Rajan suggested that instead of a unifying theory of intelligence across AI and neuroscience, there are likely "a pile of models" that will inform our holistic understanding. Saxe identified three key advances that he believes can lead to a "comprehensive quantitative framework of intelligence." First, deep learning systems have been applied to solve real-world problems. One of the lessons emerging from AI is convergence across domains, whereby a focus on developing algorithms that learn effectively allow for superior performance not only in the target domain, e.g., natural language processing, but also in other domains like vision and audition. Rajan discussed an approach called curriculum learning: similar to biological agents, AI models are trained with easier to increasingly complex data. This approach allows you to address both the forward problem, which is how can you optimize task performance, as well as the reverse problem of which learning rule was used. Second, deep learning systems are complicated and viewed as black boxes, but recent theoretical advances are unravelling their mathematical principles. Third, Saxe mentioned neuroscience methods that enable recording of large neuronal populations over long-term learning that can provide data about learning trajectories.

One theory of intelligence that Rajan highlighted is how multiple brain areas ("modules") interact to produce cohesive behaviour. Multi-area dynamical system models are gathering interest in the field, constrained by the kinds of data available to neuroscientists, such as calcium recordings of neurons that provide information of neural dynamics in a specific region, like the mouse primary visual cortex, or even the entire brain, like in the zebrafish. This data is often captured alongside behaviour. According to Rajan, this modelling enables researchers to look for convergence across species to "reveal the minimal multi-circuit model for a certain behaviour in a small model that can be used as a roadmap" for a broader understanding of how modules work together to produce behaviour in artificial and biological neural networks: Is modularity a function of task demands and the behavioural repertoire of the agent? How are multimodal measurements integrated in multiscale neural networks? What learning trajectories occur, and how can we optimize for them to boost learning?

"We now have useful approaches from both computational neuroscience and AI, and principled theoretical frameworks to be able to interpret our findings," Rajan said. Saxe emphasized that these advances in AI and neuroscience provide an opportunity to perform theoretical modelling, observe learning trajectories, and construct theories of intelligence. Central to theoretical neuroscience approaches is which facets of biology should we incorporate versus ignore when modelling intelligence and designing algorithms? While our current knowledge of genetics, the geometry of cell types, local dendritic computations, and long-range connectivity are likely important considerations, the biological aspects we incorporate into a model of intelligence depend on the scientific questions being asked.

Continued advances in AI and neuroscience

Ongoing studies in AI and neuroscience shape our current understanding of intelligent systems. Uncovering areas where humans outperform AI, such as tasks requiring compositionality, creativity, and flexibility, offer insight into how to build better algorithms. Conversely, tasks in which AI surpasses human performance reveal principles of learning crucial to intelligent behaviour, advancing our understanding of the brain and unravelling which features would contribute to better AI systems. Continued efforts to bridge the gaps between the performance of artificial and biological agents, as well as to develop theories of intelligence, will enable advances in both the AI and neuroscience fields.

Acknowledgements

This report would not be possible without the generous funding from Tianqiao and Chrissy Chen Institute, as well as the support of the Federation of European Neuroscience Societies (FENS), for the Chen-FENS Science Writer Fellowship. Thank you also to Arna Ghosh and Irum Chorghay for their feedback.