What implications do recent advances in artificial neural network technology have for the computational theory of mind? (2500 words)

The implications that recent advances in artificial neural network technology have for the computational theory of mind, while valuable as broader heuristic information, are limited in the algorithmic or implementational sense. To demonstrate this, I will begin by outlining the relationship of artificial neural network technology to the computational theory of mind. Next, I will use the example of image recognition in computer vision to show how the results of artificial neural networks seem to match actual brain processes and where they fail, as well as how these failures provide valuable implications for the computational theory of mind. Finally, I will then use AlphaGo to exemplify how the different constraints acting on artificial neural networks as opposed to actual neural networks in the brain mean that their implementations, on an algorithmic level, should be assumed to have significant procedural differences unless there is explicit evidence to indicate otherwise.

Artificial neural network technology is central to the connectionist approach to computational theory of mind, based on idealizations of real neural networks in the brain (Stinson 2020). Instead of neurons, with all the complex structures which make them up, artificial neural networks use interconnected “units”, organized in layers, with variable weights attaching the units of each layer. The neuron’s electrical output is represented as a single number at a time, which it gets by applying the specified input-output function to the weighted numbers received by the previous layer (Hinton 1992). Instead of a human programming an algorithm which solves the task the network is set to directly, the programmed algorithm instructs the network to adjust the weightings over many attempts until the network can

reliably succeed at the task on its own. This approach has proved to be relatively successful for many tasks too complex for other tactics to work, such as image recognition (Han et al. 2019) and game- playing (Halina 2021).

No matter how useful or convenient a model is, however, the burden of proof is still on the model to show that it pertains to reality. So far, there is not much evidence that artificial neural networks function by similar methods to actual neural networks on the algorithmic level, since the structure of neurons are so much more complex than the model and our understanding of actual neural networks and our understanding of which parts of the network effect the result are so incomplete (Hinton 1992, McClelland 2009). While simplification is necessary to make useful models, it comes with downsides regarding the accuracy of that model, particularly when we have no reliable way to know which details are relevant to the model and so should not be simplified (Stinson 2020). At the broader computational level, however, recent advances in artificial neural network technology can provide helpful insight, especially in where they fail to accurately model the behavior of the mind.

As Stinson explains, the degree of simplification which is appropriate in a model depends on what it is trying to explain. In this case, the model of artificial neural networks has been abstracted away from its basis in actual networks between neurons in many ways. To some extent, this simplification seems to work well enough for artificial neural networks: For example, an artificial deep neural network can successfully solve complex, high-level visual tasks at a very good success rate for normal images (Han et al. 2019). Since the model is at least loosely based on actual neurological processes, this provides some weak evidence that there may be similarities between what the model is doing and what reality is doing. If nothing else,

the abstractions made in this model for this task do not seem to impinge much on the particular task of image recognition.

On the other hand, there are also portions of the task for which it behaves differently: Most notably, Han et al found that when presented with adversarial images, images which have been overlaid with noise imitating the texture of a category to which the image does not belong, the network misattributes images to the category of the texture, rather than the category of the object in the image (Hermann & Kornblith 2019). This particular error persists even when the tiling is so faint that humans barely notice it (Han et al. 2019). The DNNs used are also much more susceptible to image distortions in general than humans are, showing that they are much less robust than image recognition processes used in humans, particularly in settings with impoverished input (Han et al. 2019)—which, in the real world, is often the case. Similar textures blend into one another, shape outlines are broken up by camouflage or by occluded sections, or texture and shape alike can be obscured by inconvenient lighting, all of which occurs more often in the real world than it does in the ImageNet database used to train this artificial neural network (Hermann and Kornblith 2019).

Many such differences between human performance and artificial neural network performance on a given task are due more to insufficient variety in the training data than some inherent inability of the artificial neural network to recognize shape. When trained on much more distorted images, for example, artificial neural networks can be made to stop relying so strictly on texture, though that does not necessarily make them more accurate overall (Hermann & Kornblith 2019). This implies one of two things: Either humans process the images differently to begin with, preventing us from running into this texture bias issue in the first

place, or we might simply have some process to correct for adversarial textures at a higher level such that we don't notice them consciously to begin with (Han et al 2019). No matter which option is the case, this discrepancy challenges the assumption that just because artificial neural networks can reach similar results as humans in most cases, the mind performs exactly the same set of computations as these artificial neural networks. It does not, however, rule out that a similar process may take place in the brain, even if the model is most likely missing some pieces.

This work with computer vision is a good example of the sorts of problems that often arise when trying to make positive inferences about the computational theory of mind from artificial neural network technology, and where such models provide valuable insight. On the one hand, the brain is very complex, and it must be simplified in order to make any sort of useful model. On the other hand, we do not know enough about the brain to reliably guess what parts can be abstracted away, or even to guess what all the necessary steps for a given process will be, such as needing to correct for the bias toward recognizing texture before shape. When put together, this lack of knowledge means that it is difficult to make many positive inferences about the computational theory of mind based on artificial neural network technology. Especially on the algorithmic level, which we have very little access to so far, this work with computer vision at best only provides weak evidence that human vision *might* work in a similar way. If any given cognitive task has multiple realizability, after all, there is no particular reason to assume that any specific implementation which succeeds at the task happens to be the same implementation present in the brain, especially not when it is far from a perfect mimic of the corresponding behavior in the brain.

The strongest implications of recent advancements in artificial neural network technology instead have to do with what is *missing* from existing models. For example, the tendency of deep learning computer vision networks to prioritize an object's texture over its color or shape unless trained on a very distorted set of images implies that since humans seem to prioritize shape over texture and color alike, one of the differences between human visual processing and this implementation of machine vision has to do with de-prioritizing textures, or at least has to do with correcting for adversarial textures. Whether this difference takes the form of a fundamental difference in how humans process images or whether it simply comes down to a more impoverished or distorted set of training data which forces us not to rely on texture alone, the discrepancy in accuracy between humans and deep neural networks does at least point out an area of interest which requires more investigation in order to make an accurate model of human vision.

Other implications of recent advances in artificial neural network technology arise not necessarily from differences in accuracy, but from differing constraints on artificial neural networks and actual neural networks in the brain. Not only is the "training data" present in the real world significantly more diverse than that in, say, the ImageNet database, but operating inside a living organism adds its own set of constraints. One of the most visible fundamental differences between human learning and weight adjustment in artificial neural networks is the training period: Humans can learn from a single instance of observation, whereas artificial neural networks such as the deep neural network used for image categorization in Han et al. 2019 was trained on 1.2 million labelled images and still suffered from a lack of diversity in training input (Lake & Tenenbaum 2015, Han et al. 2019). It could be argued that infancy, in

humans, takes the place of this training period—but even as adults, humans are adept at learning to recognize new categories from very few examples.

Furthermore, most artificial neural networks are only built to simulate one task at a time, since each task on its own is complex enough to begin with, whereas human skills tend to be domain-general wherever possible. For example, the deep artificial neural network-based program AlphaGo, which is considered to be "capable of creative problem-solving" according to the standards currently set in comparative psycology, can only apply these procedures when it comes to playing Go, the game it was designed to play (Halina 2021). In the brain, by contrast, skills tend to be domain-general as much as is possible, which certainly impacts how that skill is implemented. In the case of AlphaGo, this difference in implementation is clearly visible: AlphaGo's entire search tree is based on winning game board positions, something which only applies to the game Go, and consistently comes up with move sequences which human players do not expect (Halina 2021). Both of these aspects imply that clearly, while AlphaGo can play Go very well, it does not go about doing so in a similar way to a human, and its strategy is only applicable to this one game which it was built to play. In humans, by contrast, the same skills which are used to learn to play Go—mental planning and scenario-building, and so on—can be used both for other games and for variants of Go with slightly different rulesets, both of which cause AlphaGo to flounder (Halina 2021).

This same issue carries over, albeit less visibly, to most applications of artificial neural networks, even those which attempt, to some degree, to create models with positive implications for a computational theory of mind. Even if there are relatively few possible implementations of, for example, object recognition from an image, the way that task is

implemented in the brain is most likely useful for more than the one specific task, or is probably intertwined in some way with other processes in the brain. The same can not be said of current artificial implementations of object recognition. This is yet another difference which suggests that the implications of recent developments in artificial neural networks are mostly restricted to the broader computational level when it comes to modelling a computational theory of mind. We cannot rule out that due to multiple realizability and their differing constraints, the implementations which appear in artificial neural networks have significant algorithmic differences from the implementations of those same processes in the brain.

In conclusion, the implications of recent developments in artificial neural networks on the computational theory of mind are helpful in the heuristic sense, but ultimately limited to less direct, non-algorithmic insight. A model must show that it is related to reality before it can be assumed to give a positive account of that part of the world. While artificial neural networks are based on processes in the brain and can, in some scenarios, be adjusted to perform similarly to a human, it cannot be assumed that this is the only possible implementation of any given task. In fact, there is considerable evidence to support that while some of the differences in performance between humans and artificial neural networks likely come down to the diversity of the training data, there are also some more fundamental differences in how tasks are implemented on the algorithmic level. In particular, these differences seem to be mostly caused by the constraints surrounding a task: diversity of inputs, but also amount of training data needed to learn and the requirement that a skill be domain-general when implemented in the brain. Because of these differences in constraints, we cannot assume there are no major algorithmic differences in how a task is realized in the brain versus in an artificial neural

network. These developments in artificial neural network technology do, however, have valuable implications for the computational theory of mind on the broader, computational level: They show which areas of a model need more work, and through areas where the results of artificial neural networks differ from that of humans, they make visible constraints on the brain which are hard to notice otherwise.

Bibliography

Halina. (2021). Insightful artificial intelligence. *Mind & Language, 36*(2), 315–329. https://doi.org/10.1111/mila.12321

Han, Chihye, Wonjun Yoon, Gihyun Kwon, Seungkyu Nam, and Daeshik Kim. 2019. "Representation of White- and Black-Box Adversarial Examples in Deep Neural Networks and Humans: A Functional Magnetic Resonance Imaging Study." arXiv:1905.02422, Cornell University.

Hermann, Chen, T., & Kornblith, S. (2019). *The Origins and Prevalence of Texture Bias in Convolutional Neural Networks*.

Hinton. (1992). How neural networks learn from experience. *Scientific American*, *267*(3), 145–151.

Lake, Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science (American Association for the Advancement of Science)*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

McClelland, James L. (2009). The Place of Modeling in Cognitive Science. *Topics in Cognitive Science 1* (1): 11–38.

Schöner, G. (2008). Dynamical systems approaches to cognition. *In* R. Sun *(Ed.),* Cambridge handbook of computational psychology *(pp.* 101–126*). New York: Cambridge University Press.*

Stinson. (2020). From Implausible Artificial Neurons to Idealized Cognitive Models: Rebooting Philosophy of Artificial Intelligence. *Philosophy of Science*, *87*(4), 590–611. https://doi.org/10.1086/709730