The Plausibility of General Intelligence being a Weakly Emergent Property from Narrow Intelligence through the Meta-model Machine Learning Architectures

Andrew H. Zhang

University of Toronto St. George

COG250Y1: Introduction to Cognitive Science

Dr. J. Vervaeke

Last Revised: April 1, 2022

Abstract

Machine learning models that integrate numerous sub-models in composition with each other are referred to as “meta-models” (Hartmann et al., 2019). This essay examines the dispute between the materialist and property dualist theories of mind on whether human-level general intelligence is a weakly emergent or strongly emergent phenomenon from the human brain through insights gained from the recent advancements in the field of artificial intelligence with particular focus on models that implement the meta-model architecture. There is an abundance of empirical and theoretical evidence showing behaviors indicative of significantly increased levels of comparative general intelligence — such as complex categorization, certain degrees of relevance realization, and creativity — weakly emerge from AI systems that employ the meta-model architecture. Thus corresponding to the materialist theory of mind, it is highly plausible that general intelligence is a weakly emergent property from the interactions between narrowly intelligent agents through the meta-model architecture in AI systems.

Introduction

Emergent properties and behaviors of an entity are those that its constituent parts do not possess or display, such as "turbulence" as a property of water bodies that individual H_2O molecules cannot be said to possess, or assuming human mental processes are carried out in the brain, general intelligence as a property that emerges from individual neurons firing. Given general intelligence as a defining feature of the human mind, the debate between the materialist and the property dualist theories of mind give rise to two opposing classes of emergence. The materialist theory of mind argues that general intelligence is "weakly emergent" from the biological activities in the brain, where cognition and intelligence are completely reducible to the the firing of each individual neurons following a chain of physical causality just as the emergence of turbulence from the physical interaction of individual H_2O molecules, which by extension theoretically permits multiple realizability of the mind in non-biological systems that emulate the firing of neurons, and therefore allows AI with equal levels of general intelligence as humans to exist. On the other hand, the property dualist theory of mind argues for the "strong emergence" of general intelligence from the brain, where general intelligence is irreducible to the brain's biological activities as the materialist chain of physical causality from intelligence to neurons firing could never be formally established, which by extension implies that simulation of the human mind is impossible, and that there could never exist a human-level AGI.

The disagreement between whether general intelligence is strongly or weakly emergent from the biological brain endows the project to create human-level AGI with a profound philosophical significance as a means to settle the more than four century long debate between materialists and property dualists. From the victory of AlphaGo over Lee Sedol (Silver et al., 2016), it is evident

that narrow intelligence equal to and even exceeding humans in certain combinatorially explosive problem domains such as playing chess and go is weakly emergent from artificial neural networks. If in turn behaviors that could be considered generally intelligent can weakly emerge from the interaction between narrowly intelligent AI systems, it can be induced that it is more plausible for human-level general intelligence to be a weakly emergent phenomenon from the same processes that narrow intelligence emerges from. Such interactions between narrow AI systems can be established through the meta-model machine learning architecture (Hartmann et al., 2019). The meta-model machine learning architecture refers to when machine learning algorithms are applied in composition of the output of other machine learning algorithms designed to operate on distinct problem domains in order to operate on a more general problem domain.

To examine the plausibility of general intelligence being a weakly emergent property from narrow intelligence through the meta-model machine learning architectures, this essay will first establish a definition of "general intelligence" that is conducive to such examination, and then propose a possible mechanism for general intelligence defined as such to weakly emerge from the interactions between narrowly intelligent agents. Furthermore, with the frame problem and achieving relevance realization arguably being the most challenging aspect of the project to construct human-level AGI, we will examine how the mechanism can mitigate the frame problem and the extent to which the meta-model architecture helps achieve relevance realization in some problem domains. Moreover, with creativity in terms of the ability to produce reasonable solution output given minimal or vague problem formulation being considered as the only

achievable by generally intelligent agents, this essay will also examine how artificial creativity can weakly emerge from narrowly intelligent AI systems through the meta-model architecture.

General Intelligence as a Variable and Comparable Quality

Meaningful discussion pertaining to artificial general intelligence is impossible without first defining general intelligence in the context of AI systems. It is initially tempting to use the Turing test or its many variants to define whether or not an AI system is as generally intelligent as a human is. However, such definitions that classify AI systems to be either generally intelligent or not is counterproductive when examining emergence. Even discounting the human element of the Turing test that inevitably leads to weak reproducibility (Schoenick et al., 2017), such binary definitions are not conducive to the analysis of the mechanism by which the emergence of "general intelligence" could possibly occur. For instance, consider the two newest generations of natural language generation AI systems GPT-2 and GPT-3 developed by OpenAI. Neither of these systems are capable of passing the Turing test (Oppy & Dowe, 2021) and therefore cannot be considered to be "generally intelligent" under a binary definition. However, empirical results show that the newer GPT-3 performs significantly better compared to GPT-2 across board and particularly in areas such as producing coherent fictional stories, which it is evident that GPT-3 is more generally intelligent compared to GPT-2. Therefore, in analysis of whether a determinable mechanism exists for comparatively greater "general intelligence", like that of GPT-3 compared to GPT-2, to emerge from the interactions between narrowly intelligent artificial agents demands a definition of emergence that is variable and comparable.

From a behavioral standpoint, an artificial problem solver that is more generally intelligent will require less human intervention to assist its problem formulation and to refine its solution, while also demonstrating the ability to adequately solve a greater domain of different problems. A variable and comparative definition of general intelligence for an artificial problem solver can thus focus on the comparative degree of autonomy with which it produces solutions, and the comparative degree of adaptability to varying problems domains.

A property dualist may vehemently object to this behavioral approach to defining general intelligence with the argument that behaviors appearing to emulate general intelligence does not imply that the agent displaying these behaviors is actually generally intelligent. Considering that the only means through which one could affirm the existence of general intelligence in another agent, including other human beings, beside themselves is through their behavior, a concession can be made that this is a completely valid argument as long as its proponent accepts solipsism as an equally valid philosophical outlook as that of artificial agents can only display but not possess some degree of general intelligence defined as above.

Function Composition in Meta-model Architecture as a Mechanism for Weak Emergence

As previously elucidated, the key distinguishing feature between strong and weak emergence is the existence of a clear mechanism through which the emergent behavior or property is given rise to by the interactions between the constituent components of the whole. If we make the abstraction of narrowly intelligent AI systems as mathematical functions that accepts a input vector from the problem domain that it is designed for and outputs a vector from a possible solution range, then if general intelligence is to weakly emerge through the meta-model architecture where the outputs of one or more AI systems is fed into another as inputs, it will be through function composition. Just from a purely mathematical standpoint, we can already observe weakly emergent behaviors that are essential to a general problem solver from the composition of very simple functions. First consider the linear function $f(x) = x$ in 2D space. If we compose this function as $f(f(x)) = x^2$, we obtain a parabola, where curvature is a weakly emergent property from this composition. Now further consider the task of drawing a boundary that completely separates the yellow data points from the purple ones on *Figure 1* and the parabolas $y(x) = x^2$, $x(y) = y^2$ in 2D space as well as the linear function $z(x, y) = x + y - 0.5$ in 3D space.

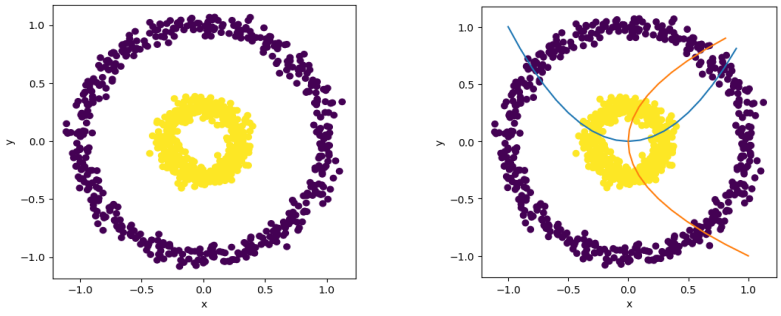

Figure 1. This figure is self-generated. See ***appendix*** *for the Python code that produced this figure.*

Left: purple and yellow data points distributed as concentric rings.
Right: Neither $y(x) = x^2$ (in blue) nor $x(y) = y^2$ (in orange) are able to separate the yellow and purple data points.

It is evident that in the initial 2D problem domain, neither $y(x) = x^2$ nor $x(y) = y^2$ could achieve this task. In fact, the injectivity of functions makes this task impossible to achieve for functions in 2D no matter what value of weights and biases that we add to the parabolas defined above. Even if we introduce a new dimension "z" to the problem domain, the task is still impossible for the linear function $z(x, y) = x + y - 0.5$ in 3D as can be seen on *Figure 2* since it has no curvature. However, if we compose the parabolas $y(x) = x^2$, $x(y) = y^2$ with $z(x, y) = x + y - 0.5$, we obtain the 3D paraboloid $z(x, y) = x^2 + y^2 - 0.5$, for which the ability to complete this task weakly emerges as illustrated on *Figure 2*.

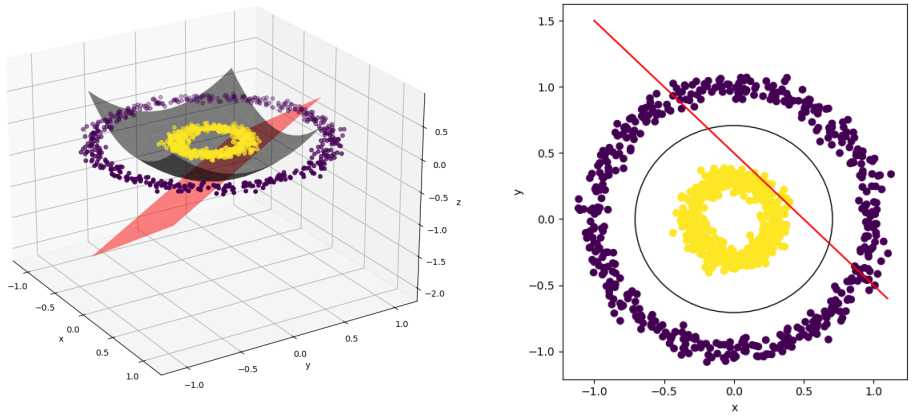

*Figure 2. This figure is self-generated. See **appendix** for the Python code that produced this figure.*

Left: It is impossible for $z(x, y) = x + y - 0.5$ (in red) to complete the task due to the lack of curvature, while the ability to complete this task emerges for $z(x, y) = x^2 + y^2 - 0.5$ (in black) after composition.
Right: Projection of the figure on the left onto the original 2D problem domain as a level set at $z = 0$.

While the above example illustrates the ability to separate different classes of data points organized in concentric rings weakly emerging from linear components that have no curvature and constrained by injectivity with just three levels of composition from linear functions to parabola to paraboloid, its implication is much more powerful if we consider the task of separating two sets of data as an abstraction of the cognitive task of categorization, which is a cornerstone of human general intelligence.

Consider deep neural networks as a machine learning algorithm. While in the modern machine learning lexicon, the deep neural network is often referred to as a monolithic machine learning algorithm that stands alone on it's own (Balas et al., 2021), historically it is developed from the multi-layered composition of a simpler structure named the "perceptron" that is its own narrowly intelligent machine learning algorithm (Freund & Schapire, 1998). This makes the deep neural network arguably the first machine learning algorithm to implement the meta-model architecture. Each array of perceptrons organized as the layers of "neurons" in a deep neural network essentially acts as a multivariable function that composes with the preceding layer by taking their output as input and passing its output to the following layer as input. As we have already illustrated how complex behavior can weakly emerge from the composition of simple functions, it is not surprising that the universal approximator theorem for artificial neural networks (Hornik et al., 1989) states that given a sufficiently large number of layers of function composition, a deep neural network is able to approximate virtually any function, and be able to use this approximation to solve virtually any categorization problem. *Figure 3* illustrates an example of how a neural network is able to classify highly entangled data points organized in concentric spirals with remarkable accuracy.

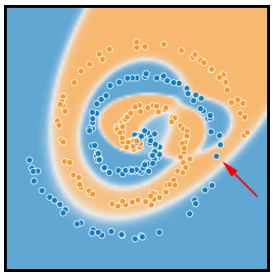

Figure 3. A neural network separating data points organized in concentric spirals. (Mantini & Shah, 2021)

There is an abundance of evidence demonstrating the weak emergence of more generally intelligent behaviors from greater depth of the layers of function composition in AI systems, the most relevant of which in the context of artificial neural networks and categorization is the comparison between neural networks AlexNet (Krizhevsky et al., 2017) and LeNet (LeCun et al., 1998). Both designed for the purpose of image recognition and employ similar architecture, LeNet with 5 layers of composition is able to classify a maximum of 10 different image classes with a dimension of 28 by 28 pixels within a reasonable error rate, while AlexNet with 8 layers of composition is able to do so for 1000 different image classes with a dimension of 224 pixels by 224 pixels, and therefore by our comparative definition of general intelligence, AlexNet is irrefutably more generally intelligent than LeNet.

A property dualist may reject the possibility of abstracting the human mind as mathematical functions like we did with AI systems in the above analysis to undermine the generalizability of this analysis to the general intelligence of human-beings. As a response to this counter-argument, one should consider from a behavioral perspective the nature of functions. A function in essence takes in a set of inputs and through some internal mechanism returns a set of outputs. If we view the firing of each one of our sensory neurons as belonging to this set of inputs and the firing of every one of our motor neurons as belonging to this of outputs, the possibility of representing the human mind as a function becomes very tangible.

Meta-model Architecture and Mitigating the Frame Problem

With logical reasoning as an essential capability of human-level general intelligence, the frame problem refers to the problem where there is no apparently obvious way for an artificially created intelligent to bypass the need to explicitly logically represent all logically representable statements about its input, which is combinatorially explosive, in order for logical reasoning to take place. (Shanahan, 2016). Due to the frame problem, it is almost always required that a human frame the problem such that only relevant information is present in the input to a narrowly intelligent agent for any meaningful processing by the narrow AI to take place. The frame problem is arguably the most challenging obstacle faced by the project to create AGI.

As an ideal method to solve the frame problem, relevance realization in essence is the ability of an intelligent agent to realize the relevance of a logical statement to the reasoning that needs to take place, and therefore be able to decide if it has gathered sufficient information to make a logical conclusion. (Vervaeke et al., 2009). Consider again the impressive feat of AlexNet as an artificial agent successfully categorizing unseen images into 1000 different classes within an acceptable margin of error. With the ability to distinguish between image classes such as “leopards”, “jaguars”, and “cheetahs” in its training set that even some humans might have trouble with, AlexNet is evidently capable of framing and solving its own sub-problem domain in order to achieve its main problem domain of image classification given that its program is able to execute within a finite amount of memory and running time. Consider AlexNet’s classification output for an image of a leopard and a motor scooter as illustrated on *Figure 4* on the next page. The fact that AlexNet’s top 5 choices for the most probable class that the image on the right belongs are all felines suggests that AlexNet would first recognize that the image on the right is

that of a feline before framing the sub-problem domain of identifying which species of feline is in the image by focusing its attention on features that are relevant to this identification such as the pattern on the feline's fur and the shape of its body. However, focusing on these features would have been useless when it comes to identifying the type of vehicle in the image on the left, as they are not relevant. Given that AlexNet's top 5 choices for the image on the left are all that of vehicles, it is evident that the same process is at work to focus on features that are relevant to identifying vehicles, and that AlexNet's architecture certainly does mitigate the frame problem by achieving some degree of relevance realization.

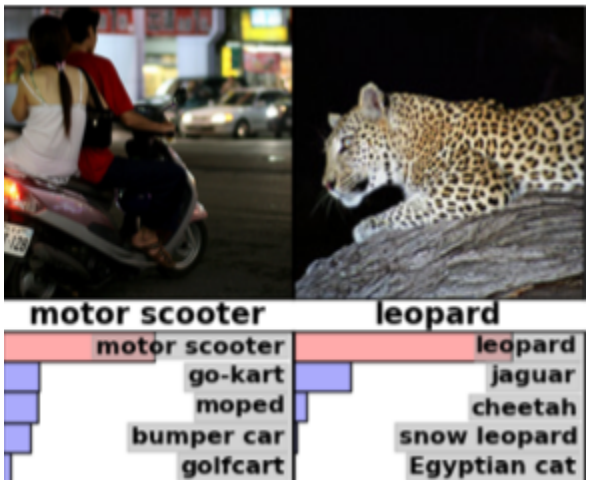

Figure 4. Classification example output from the original AlexNet paper. (Krizhevsky et al., 2017)

From top down are each of what AlexNet considers to be the five most probable image classes for the shown image, the bar represents this probability, and the red bar denotes the correct image class as labeled by humans.

Paying closer attention to the mechanism by which the ability to achieve a certain degree of relevance realization could weakly emerge from the meta-model architecture, it is very apparent that the ability to create abstractions of the input is integral in the input set. Each layer of perceptrons that compose with the preceding layer in a neural network creates a layer of abstraction of the input it receives and passes it on to the following layer, and this abstraction in the earlier layers is what allows AlexNet to first determine the general class that an image input

belongs to, such as felines or vehicles, and accordingly employ thee later layers to process features relevant to this abstraction.

Meta-model Architecture and Artificial Creativity

A property dualist may maintain that despite that the meta-model architecture enables more generally intelligent behaviors to weakly emerge in artificial agents, artificial agents will never be capable of creativity as represented by the ability to generate brand new meaningful outputs that they have never seen before. However, this is exactly what the generative adversarial network (GAN) is able to achieve by constructing a meta-model consisting of a generator deep neural network whose purpose is to artificially generate images of a fixed dimension based on the weights and biases of its neurons, and a discriminator neural network whose purpose is similar to that of AlexNet, namely to classify an image that it sees as either being a real image or one that had been artificially generated (Goodfellow et al., 2014). As the name of the model implies, the generator and discriminator neural networks in the GAN's meta-model works in an adversarial nature. As the generator attempts to create artificially generated new images, real images are also fed into the model as input. The generated image and the real image then are both randomly sent as input to the discriminator, which attempts to discern whether or not the image it received is artificially generated. Whether the discriminator could successfully discern an artificially generated image was then sent back to the generator as its cost function for it to generate more convincing images that could potentially deceive the discriminator in future iterations. As both the generator and the discriminator are trained on their respective inputs and become better at their designated tasks, the GAN as a whole will be able to artificially generate

more and more convincing original images that belong to the same image class that it received as inputs if we extract the images generated by the generator as the GAN's output. This process is illustrated on *Figure 5*, which is taken from the paper that first proposed the GAN meta-model.

Figure 5. Example output of a GAN from the original paper that proposed this model. (Goodfellow et al., 2014)

From left to right are the GAN's output with increasing iterations for each different image class.

From the mechanism with which GANs operate, it can be said the meta-model composition of the generator with the discriminator provides the GAN with a built-in ability to examine and evaluate its own outputs and adapt accordingly to it, giving it a sense of "understanding" of the quality of the image it is generating, and therefore arguably a degree of "self-awareness" that emerges from the meta-model.

Conclusions

All of the above analysis strongly suggests that general intelligence is strongly emergent from the composition of narrowly intelligent agents through the meta-model machine learning architecture. From the examples outlined, behaviors indicative of significant levels of increase in comparative general intelligence such as complex categorization, certain degrees of relevance realization, and creativity have been demonstrated to weakly emerge from AI systems that employ the meta-model architecture.

If we examine the evolution of our biological brains that gave rise to our general intelligence, we find a process that is comparable to the approach of creating AIs from AIs through the meta-model architecture. Different regions of our biological brain specialize in different narrowly defined tasks, such as the visual cortex for processing the sensory input from rod and cone cells and the auditory cortex for processing auditory input. In an abstract sense our very own brain is a meta-model consisting of each of its narrowly specialized regions, and it is highly plausible that our own general intelligence emerged from the composition of the inputs and outputs of these different brain regions. It is thus not an outlandish outlook that if we continue to recursively integrate existing AI systems to create new more powerful ones, a generally intelligent artificial agent with capabilities paralleling ours could one day evolve and weakly emerge from this process.

Therefore, until a specific aspect of human general intelligence could be proven to be fundamentally impossible to mechanize, the materialist theory of mind will remain the more empirically and philosophically substantiated one.

Word Count: 3342

References

Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. E. (2021, September 15). Multilayer Perceptron and Neural Networks. Wseas Transactions on Circuits and Systems. Retrieved April 1, 2022, from https://www.academia.edu/52398369/Multilayer_perceptron_and_neural_networks

Freund, Y., & Schapire, R. E. (1998). Large margin classification using the perceptron algorithm. Proceedings of the Eleventh Annual Conference on Computational Learning Theory - COLT' 98. https://doi.org/10.1145/279943.279985

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. Communications of the ACM, 63(11), 139–144. https://doi.org/10.1145/3422622

Hartmann, T., Moawad, A., Schockaert, C., Fouquet, F., & Le Traon, Y. (2019). Meta-modelling meta-learning. 2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems (MODELS). https://doi.org/10.1109/models.2019.00014

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional Neural Networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

Mantini, P., & Shah, S. K. (2021). CQNN: Convolutional Quadratic Neural Networks. 2020 25th International Conference on Pattern Recognition (ICPR). https://doi.org/10.1109/icpr48806.2021.9413207

Oppy, G., & Dowe, D. (2021, October 4). The turing test. Stanford Encyclopedia of Philosophy. Retrieved April 1, 2022, from http://seop.illc.uva.nl/entries/turing-test/

Schoenick, C., Clark, P., Tafjord, O., Turney, P., & Etzioni, O. (2017). Moving beyond the Turing test with the allen AI science challenge. Communications of the ACM, 60(9), 60–64. https://doi.org/10.1145/3122814

Shanahan, M. (2016, February 8). The frame problem. Stanford Encyclopedia of Philosophy. Retrieved April 1, 2022, from https://plato.stanford.edu/entries/frame-problem/

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and Tree Search. Nature, 529(7587), 484–489. https://doi.org/10.1038/nature16961

Vervaeke, J., Lillicrap, T. P., & Richards, B. A. (2009). Relevance realization and the emerging framework in Cognitive Science. Journal of Logic and Computation, 22(1), 79–99. https://doi.org/10.1093/logcom/exp067

Appendix

For the Python code that generated *Figures 1 & 2*, see GitHub repository:

https://github.com/a663E-36z1120/cognitive-science/tree/main/Weak%20Emergence%20of%20General%20Intelligence%20through%20Meta-model%20Architecture