Response to the book: "*On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*", by Jeff Hawkins.

*"Prediction is not just one of the things your brain does.*
*It is the primary function of the neocortex,*
*and the foundation of intelligence."*

Many researchers in artificial intelligence (AI) today believe that, with enough computing memory and processing power, what is thought of as true AI will be achieved. The book "*On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*", by Jeff Hawkins is critical of heretofore models of intelligence that have guided such developments. The author convincingly argues that human brains and the intelligent machines we have as of today do inherently different things. He asserts that an ability to accurately predict outcomes is the foundation of intelligence, and reaches into neurobiology to build what he argues is the path to true AI: the "*memory-prediction*" framework. The present exercise should serve to unpack his premise and gauge its potential, placing it in the context of complementary and competing models, while considering some deeper questions: What observable changes would this model, if implemented, bring to AI applications? How does the understanding of "true" intelligence continue to shape the evolution of AI? What really drives forward research and developments in this field?

First of all, we will take a closer look at the book's thesis. The *memory-prediction* framework is elegantly assembled, presented as a layered but concise theory. It outlines the most important traits of the neocortex that enable the human brain to make predictions of future events, and claims that our unique form of intelligence –our predictive ability– exists in this region of the brain. In short, the neocortex is responsible for forming and storing patterns from sensory input, and retrieving them by detection of at least partial similarity in new ones. In the author's own words, *"The brain knows about the world through a set of senses, which can only*

*detect parts of the absolute world. The senses create patterns that are sent to the cortex, and processed by the same cortical algorithm to create a model of the world... It doesn't matter where the patterns come from; as long as they correlate over time in consistent ways, the brain can make sense of them" (p. 43).* This theory thus implies a very specific understanding of intelligence, and of the exact mechanism by which the brain achieves it.

One of the framework's core assumptions is that true understanding cannot be measured by observable behavior. While Alan Turing's work gave us the powerful Universal Turing Machines and the Turing Test, the author notes, their gift could be better described by effective information processing than AI. Turing's approach inherently defines intelligence in terms of the observable behavior of a program upon processing an input, and such behavior as its ultimate measuring stick. However, we could ask, if a program is only good at the one particular task for which it was designed, showing no flexibility or ability to generalize, then are we dealing with true intelligence? Naturally, Hawkins cites the old debate fueled by John Searle in 1980 when he proposed the Chinese room experiment: at its heart is the question of what intelligence really is.

Furthermore, integral to the book's premise is a criticism of the disconnect between research in neurobiology and AI, and a call to build a stronger bridge between the two. Those who oppose blending neurobiology with computer science may argue that the human brain, being the product of a hundred-million-year evolutionary process, is the equivalent of a cluttered program built in an unplanned fashion, full of inefficiencies and unnecessary complexities. According to this view, creating AI should emulate the invention of the wheel, possibly the most groundbreaking in transportation: it does not replicate the mechanisms of the fastest animals in nature, but still surpasses them in functionality. So, why devote limited resources to map the brain's circuits, if computer science can develop more efficient methods to achieve equivalent or even superior results? If we consider the costs for business and funding-dependent research to reconfigure the goal of machine intelligence and rebuild theories through trial and error, there are sufficient arguments for AI researchers to avoid diving into a study of the human brain altogether.

Hawkins highlights how his interest in the study of intelligence from a biological perspective was inspired by the work and ideas of theorists such as Francis Crick and Christoph

Koch, who two decades earlier began exploring the neurobiology of consciousness. In his 1995 book "*The Astonishing Hypothesis: The Scientific Search For The Soul*", Crick, a leading British molecular biologist, offered the view that everything that we experience, our sensations, emotions, even our sense of identity was the product of the behavior of our nerve cells. In this work, he specifically focused on visual awareness, and discussed the "binding" process of visual input in the brain. Crick suggested that "coherent oscillations" of neurons found across the cortex act as the binding mechanism of the many features of mental images (such as color, form, and motion), but doubted that the vivid pictures we host in our minds could be the result of messy neural patterns. He admitted that the evidence at the time was not strong enough to back up his proposal and suggested research efforts needed to explore this direction. Nonetheless, his work was highly influential in a different way: he has been called a "cross-worlds" influencer, because of his drive to help organize "un-disciplined" new research fields by building theoretical frameworks that would foster experimental research programs (Aicardi, 2016). He was especially motivated by questions that "seemed beyond the power of science to explain." And until then, the study of consciousness was not considered scientifically respectable. This background serves to reflect on the value that can be drawn from daring to bring together ideas and findings from different fields, to close evident gaps that, for a diversity of reasons, practicing scientists within a field might escape addressing.

All in all, a review of current AI models suggests that efforts in the evolution of the field have prioritized the use of Artificial Neural Networks (ANN), actually the closest extant framework to Hawkins'. ANN are structurally modeled from the human nervous system; they have been aptly described by the inventor of one of the first neurocomputers, Dr. Robert Hecht-Nielsen, as "a *computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs*." (Caudill, 1989). In simple terms, it works by first structuring a network for a particular application, which undergoes an iterative learning process where inputs are processed one at a time, and results are compared against desired outputs. Errors are then propagated through the node system, which makes necessary adjustments for the next iteration. (Interestingly, even though the neocortex, which is thought to have developed at later stages in human evolution, is considered

a more complex and high-level structure in our brains than the "old" brain, the more instinctive, vital functions, it is easier to replicate in programming.) The ANN model has been around since the 1960s, but mostly gained popularity from the 1980s on. Since then, it has been successfully implemented in software for image processing, natural language processing, handwriting detection, and prediction models for a diversity of applications. Its main strengths are the abilities to recognize patterns and identify errors in input-to-output mapping to refine its performance.

At this point, it is worth breaking down an essential component of the *memory-prediction* model: the four essential properties of the neocortex that make it unique. Specifically: it **stores sequences of patterns**, meaning that it can encode and process temporally ordered, sequential events; it **recalls patterns auto-associatively**, it feeds the output of each neuron back into the input, creating a feedback loop that associates patterns with itself; it **stores patterns in an invariant form,** it is able to receive incomplete or altered inputs through any of the five senses, and recognize and fill in the missing parts correctly. Artificial equivalents will normally seek exact matches before claiming recognition, and may miss a pattern that has been rotated, rescaled, or transformed in some way; and it **stores patterns in layers of hierarchy**, not physically but functionally defined hierarchies. Some regions receive input from the senses and process the information at its rawest, most basic level. They transmit a signal to others further up in the hierarchy; each of these is concerned with more specialized or abstract aspects of the information. Interestingly, information does not flow in a single direction, and actually more signals flow back down from the "higher" areas in the hierarchy.

Although ANN replicates the distributive nature of information in the brain, Hawkins would point out that it does ignore two important aspects of how the latter is organized: first, the crucial role of feedback coming in and out of the neural circuit, and second, the hierarchical organization of the neocortex. It follows that the neural networks approach isn't misguided, but essentially lacking in these respects. In addition, the author notes that it has been static in its development for decades, branching out into novel applications but mostly without any structural improvements. One small branch among neural theorists (or "connectionist" researchers) did come closer to realizing the missing elements. Their focus was on *auto-associative* memories, the product of neurons that fed their outputs back into their source of

input, a form of feedback. In this model, nodes would develop patterns out of these firings with themselves. The most crucial result of this process is that it allows retrieval of an entire pattern stored in memory by inputting incomplete pieces of the whole. This process proves key in the workings of pattern encoding and recognition, and to Hawkins' theory.

A final practical issue to ponder about Hawkins' proposal is: how would the implementation of a neural network based on its principles behave? This is a hard, but very interesting question to think about. Based on all of the above, a program's pattern memory would look very different. It would learn about its environments in a new way, especially when it comes to sensory, real-time data. For instance, self-driving cars might be able to make useful human-like predictions: if a car in front has its blinkers on for longer than normal, it is likely that it will not turn left or right after all. For a human, this would be quite an easy assumption to make, and in many cases an accurate one that would feed into the decision-making process in a useful way. Self-driving cars might be able to understand traffic and driving itself, in such a way that they can better blend in with every other human driver who is processing the information in this very unique way. By the same token, its driving ability would be adaptive to such inputs, and improve in this direction over time. After all, expecting and quickly responding to behavior outside of the norm is the mark of a "better", more experienced driver. Certainly, both the general public and scientific community will once again question if this approach might somehow "contaminate" the purer logic of a computer, adding an unnecessary "human error" component. One valid counterargument is that, as we progressively interact and rely more on machine intelligence, wouldn't it be desirable that we all spoke "the same language"?

Repeated resistance from the AI community to his ideas and a deep conviction of the value in this research direction, have led Hawkins to found his own research center *Numenta*, at the heart of Silicon Valley. It has been dedicated to the study of the human neocortex for the past decade. An article published just this month in the *New York Times* announced the researchers are "*well on their way to cracking the problem…which he* [Hawkins] *says explains the inner workings of cortical columns, a basic building block of brain function*", and adds that they "*decided that cortical columns did not just capture sensations. They captured the location of those sensations. They captured the world in three dimensions rather than two. Everything was seen in*

*relation to what was around it."* (Metz, 2018) From this more recent report, it appears that the theory in question has been evolving since the publication of this book in 2005, but its basic principles hold. Over time, we might get the opportunity to see its concrete results.

As a final thought, one interesting byproduct of this theory of intelligence is that not only does it provide a rationale for human behavior, it may also help explain dimensions of our identities. If "*all cortical predictions are predictions by analogy*" (p. 124), then we predict the future by analogy to the past. Our predictions thus define our expectations of what is possible and what is not in the world around us and in our own lives. Simply consider how a theory of intelligence can help trace what we understand as the personality tendencies of optimism and pessimism, back to the simpler mechanism of predicting future events. We ask ourselves: "*What does this situation resemble?; How is it likely to play out?; How should I choose to act?*" Sharing our unique intelligence quality with machines is consequential at a technical level, but also at a philosophical one.

References:


Aicardi, C. (2016) Francis Crick, cross-worlds influencer: A narrative model to historicize big bioscience. *Studies in History and Philosophy of Biological and Biomedical Sciences,* Vol. 55,
pp. 83-95

Caudill, M. (1989). Neural Network Primer: Part I. *AI Expert*, *Vol.* 2(12), pp. 46- 52.

Crick, F. (1994). *The astonishing hypothesis: The scientific search for the soul*. New York: Scribner.

Crick, F. & Koch, C. (1995) Are we aware of neural activity in primary visual cortex? *Nature*, Vol. 375(121-123).

Hagan, M.T., Demuth, H.B. & Beale, M. (1996) Neural Network Design. Boston, MA: PWS Publishing Co.

Hawkins, J. & Blakeslee, S. (2005) On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines. New York: Macmillan

Koch, C., & Davis, J. L. (1994). *Large-scale neuronal theories of the brain*. Cambridge, Mass: MIT Press.

Metz, C. (2018, October 14). *Jeff Hawkins Is Finally Ready to Explain His Brain Research.* Retrieved from https://www.nytimes.com/2018/10/14/technology/jeff-hawkins-brain-research.html on Oct. 15, 2018

TED Conferences, LLC (2009) *How brain science will change computing*. Retrieved from https://www.ted.com/talks/jeff_hawkins_on_how_brain_science_will_change_computing/discussion?en#t-1160625 on October 10, 2018