Numenta

“A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex”

A Companion Paper

The ideas in this piece are formally introduced in detail in a pre-print journal paper available on bioRxiv: “A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex” by Jeff Hawkins, Marcus Lewis, Scott Purdy, Mirko Klukas, and Subutai Ahmad, October 2018.

Numenta recently released a research paper titled "*A Framework for Intelligence and Cortical Function Based on Grid Cells in the Neocortex."* The paper introduces a new theory for how the brain works. Because the paper assumes prior knowledge of neuroscience concepts and terminology, people without a neuroscience background may find it challenging to read. The purpose of this companion article is to put into plain language the discoveries introduced in the paper. We've simplified some of the neuroscience, focusing less on the details and more on the ideas. We invite everyone to read the full paper, but depending on your experience with reading neuroscience papers, you might want to start with this piece.

Introduction – The Mystery of the Brain

The human brain is an amazing organ. Everything ever accomplished by a human, from the sciences to the arts, from engineering to philosophy, was created by the human brain. Over the past 100 years neuroscientists have amassed an enormous amount of detail about the brain, but, you may be surprised to learn, there is no accepted theory as to what intelligence is and how the brain produces it. As Francis Crick said in 1979, "In spite of the steady accumulation of detailed knowledge, how the brain works is still profoundly mysterious." He concluded that we lack a "theoretical framework," a framework in which we can interpret experimental findings and to which detailed theories can be applied. Nearly forty years after Crick wrote his essay, his observations are still largely valid.

Numenta has been working on understanding how the brain works since 2005. Jeff Hawkins, co-founder of Numenta, was previously the director of the Redwood Neuroscience Institute, which had a similar goal. The scientists at Numenta are theorists who study the data collected by experimental neuroscientists, create theories that explain the data, and predict new results that can be tested. In other words, while lab neuroscientists run experiments and then offer conclusions, theoretical neuroscientists come up with ideas that suggest further experimentation. Our area of focus is the neocortex, the part of the brain responsible for intelligence and perception. Our latest paper proposes a novel theory for understanding what the neocortex does and how it does it.

The theory contains new ideas and hypotheses, and describes new types of cells we propose exist in the brain and what we think they do. For the purposes of this piece, we won't go into detail about where exactly in the brain we believe these cells are, or cite the many neuroscience papers we draw on, but for those details, we refer you to the research paper.

Before delving into our proposal, let's first take a moment to cover some "brain basics" to lay the groundwork for the new theory.

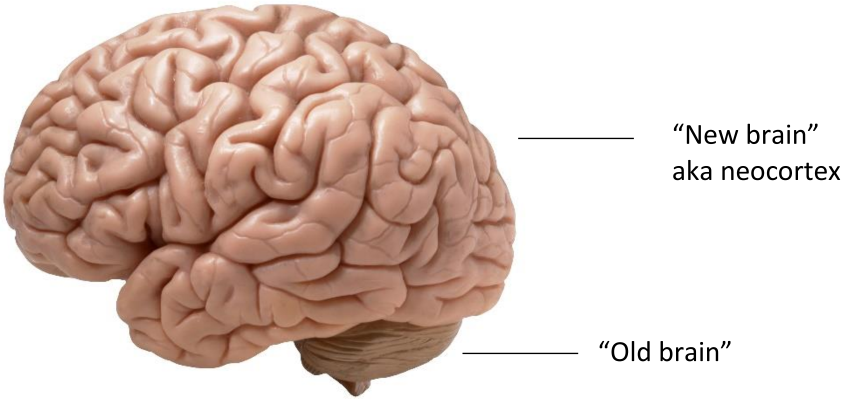

Figure 1. Brain anatomy showing a bit of the "old brain" with the "new brain" on top, covering the rest.

Old brain vs. new brain

A simple way to think about the brain is that it has two parts: the "old brain" and the "new brain." The old brain, which comes from our reptilian ancestors and pre-dates dinosaurs, contains several different structures, such as the spinal cord and brainstem. It regulates your body (such as breathing), creates reflex behaviors (such as pulling your hand away from fire) and creates emotions (such as desire and anger). The new brain, or neocortex, is a single large organ. It sits on top of the old brain and is the brain's analytical engine. It's the part that can identify objects, learn a new language, or understand math. The neocortex takes in huge amounts of sensory input and makes sense of it all. To do this, the neocortex learns a model of the world and everything in it. For example, by seeing, touching and hearing a cat, the neocortex builds a "model" of cats. The model includes not only how a cat looks, feels and sounds, but how it behaves. The neocortex also builds models of abstract objects, things that we can't sense directly, such as "democracy" or "fairness."

Consistency of the cortex

The human neocortex, if you were to flatten it out, is roughly the same area as a large dinner napkin and a little bit thicker (2.5mm). Different areas, or "regions," of the neocortex have different functions. Some are responsible for vision, others for touch, others for language, and others for planning. There are dozens of regions in a human neocortex. If you were to cut into one of the regions, as experimental neuroscientists do, you would see extraordinary complexity, with dozens of different neuron types arranged in multiple horizontal layers stacked on top of each other, and intricate connectivity patterns. However, as complex as this circuitry is, it's relatively consistent across all regions, no matter what those regions do. Regardless of where you look, you'll see roughly the same cells and the same complex circuitry.

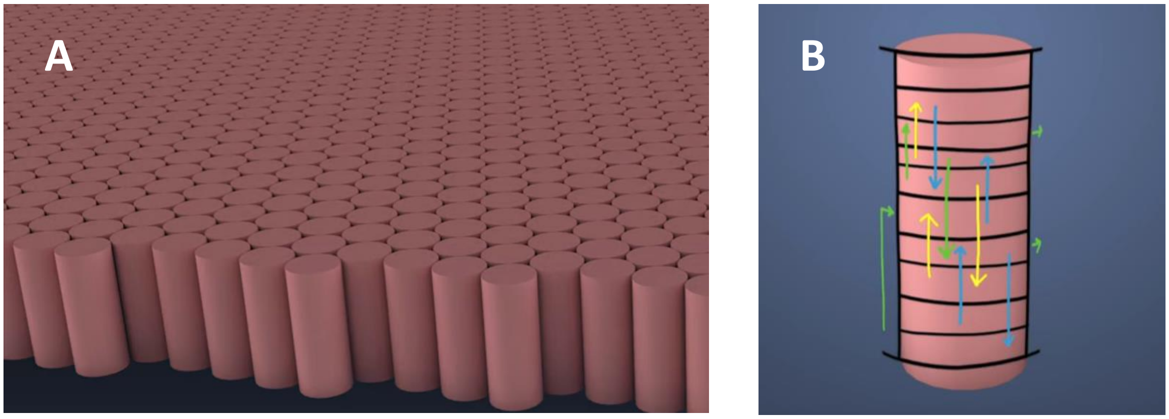

Figure 2A. Cartoon rendering of the neocortex spread flat, comprising cortical columns. **2B**. Cartoon rendering of the complex circuitry within a single cortical column.

The common cortical algorithm and the cortical column

In 1978, the neurophysiologist Vernon Mountcastle made a radical proposal. He suggested that the reason all regions of the neocortex have the same complex circuitry is that they are all doing the same thing. What makes one region of the neocortex a vision region is that it is connected to the eyes and what makes another area of the neocortex an auditory region is that it is connected to the ears. Mountcastle suggested that there is a **common cortical algorithm** that applies to everything the neocortex does. He further observed that a small area of neocortex, about 1mm square and 2.5mm deep, contains all the cell types and circuitry observed in the neocortex. Therefore, this small volume of neocortex, which he called a **“cortical column,”** must encompass the cortical algorithm. With a few exceptions, cortical columns are not physically demarked entities. A column is just a convenient way of referring to the building block of the neocortex that models objects in the world.

What do all the different cells, layers, and cell types in a cortical column do? It is difficult to imagine that a cortical column could do something that applies to everything humans do -- from vision, to touch, to language. Yet it does. As we discovered in our research, the cortical column is far more powerful than anyone imagined.

What Our Paper Proposes: The Thousand Brains Theory of Intelligence

Our paper proposes a basic explanation, or “framework,” for why cortical columns are complex and how they function. In brief, we propose that every cortical column learns models of complete objects through movement. We call this idea the **Thousand Brains Theory of Intelligence** because if every individual column is learning complete models, then our brain is not building one mega-model of an object, but rather thousands of models of an object in parallel. This section explains this theory further.

The idea of a column learning a model of a complete object is novel. Generally accepted wisdom in neuroscience describes objects being learned in a hierarchical system. Neuroscientists have understood for many years that the regions of the brain are connected in a hierarchical fashion. Sensors, such as your eyes or skin, stream input to a region. This region passes input to other regions, who pass it along to another region, until eventually it reaches the "top" of the hierarchy. The general belief is that the brain extracts more and more complex features as the data moves up the hierarchy. Imagine seeing a dog. At the lowest levels of the vision hierarchy, our brain might recognize simple lines, colors or textures; at the next level it might recognize more complex features such as the curve of an ear, the next level might detect the dog's face and at the highest level, the brain declares the result, "dog."

But this hierarchical model has always been incomplete. For one thing, there are many neurons that project horizontally to each other, between layers, and across the regions, rather than just vertically to their daughter or mother region. In a pure hierarchical model, these connections would not be necessary. In fact, more than 95% of the synapses in the neocortex are not explained by the pure hierarchical model. Additionally, many AI and deep learning networks are built on this hierarchical model. They typically require dozens of levels and millions of examples to learn something, whereas a human can learn something new in just a few exposures. The brain is clearly doing something different, and the general hierarchical view explains some of the story, but not all, so there must be something missing.

We believe we have uncovered the missing piece: all cortical columns have a signal representing **location**. This location signal represents a location relative to the object being sensed, not relative to the person sensing the object. Every column combines its sensory input with the location signal. In other words, a column knows not only what feature is being sensed, but where that feature is on the object. Each small patch of the retina or part of the finger learns a slightly different model. As we move our sensors, the "features at locations" input is integrated over time so that a single column can learn and recognize complete objects.

To illustrate this concept, imagine touching a coffee cup with one finger. As you move your finger over the cup, you sense different parts of it. You might feel the lip, then the curve of the handle, then the flatness of the bottom. Each sensation you receive is paired with its location relative to the cup. The curved handle of the mug is always in the same relative position on the cup, it is not a feature relative to you. At one moment it might be on your left and another moment on your right, but it is always in the same location on the cup.

With a location signal, each cortical column is now able to learn complete models of objects. Thus, in our theory, there isn't one single model of an object; there are thousands of models – some based on seeing the object and some based on touching it, and so on with all your senses (see Figure 3). The columns "vote" to resolve ambiguity. To illustrate how voting works, let's say you are petting your cat and one finger has rested on her collar and the other fingers on her back. Each finger will try to understand what it might be touching, based on what it's feeling at that particular location, while your eyes are informing their own cortical columns. Even though the finger resting on the collar might first guess "leather belt," its second guess might be "cat" because you've felt the collar when you've pet your cat before. In the meantime, your visual cortical region is guessing "cat" as well. One column has a different first choice, belt, but through the voting process, the columns quickly land on the common answer among them and agree it is a cat with a collar. The many horizontal connections across neocortical regions seem to be the logical place where voting occurs. The Thousand Brains Theory elegantly solves the problem of how the brain integrates information from multiple senses – each sense votes to arrive at a unified perception of an object.

It's important to note that movement is critical to learning. It's how we build models. For example, picture a baby encountering a rattle for the first time. What does she do? She touches it, grabs it, shakes it, puts it in her mouth, sets it down, picks it up again. With all her senses, she explores the object through movement. Without moving, she would not learn that the rattle makes a sound when she shakes it, or that one end is much easier to grab, or that it's not something she should eat.

We are not proposing that the longstanding view of the cortex built upon a hierarchy of regions is incorrect; rather, it's incomplete. We think the hierarchy still adds tremendous depth to perception and to understanding objects and ideas at different scales, but it is not the principle organizing mechanism of intelligence. Our proposal that every cortical column is capable of learning complete models means that object recognition does not occur solely at the top of the hierarchy. It occurs in every column, at all levels.

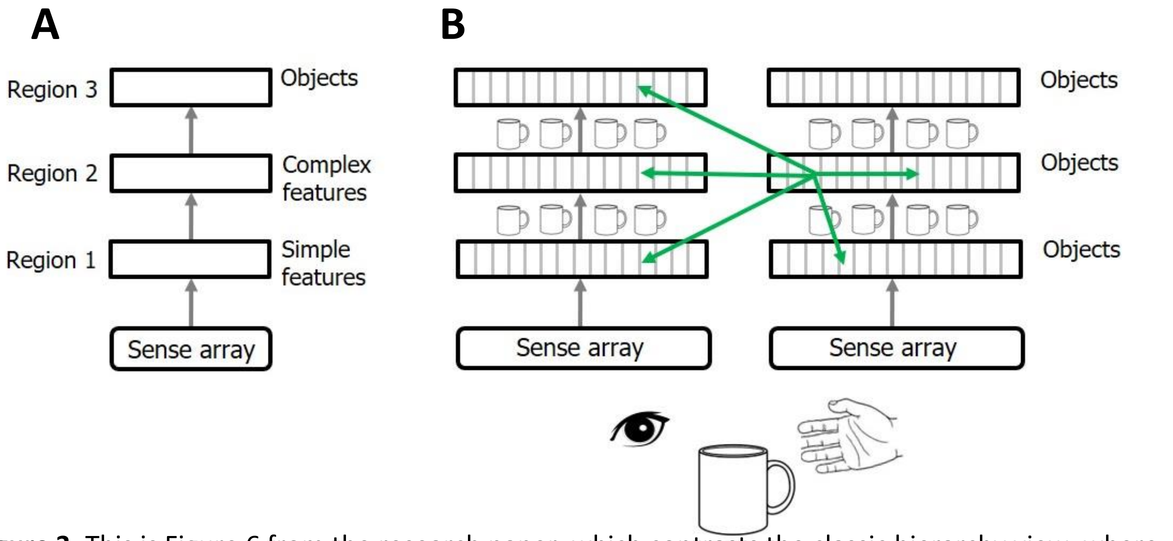

Figure 3. This is Figure 6 from the research paper, which contrasts the classic hierarchy view, where complete objects are recognized only at the top (on the left), with our Thousand Brains view, where every level and every column learns complete models of objects (on the right).

Let's now turn our attention to more depth on the location signal.

Grid Cells

In the new paper, we hypothesize where and how the location signal is generated: from grid cells in the neocortex. Grid cells are a foundational component of our theory, but unless you're a neuroscientist, you've probably never heard of them, so what are they exactly? A grid cell is a type of cell, or neuron, that exists in an older part of the brain involved in navigation. Grid cells are well studied, and over the last few decades neuroscientists have made great progress in understanding their role. Grid cells represent the location of a body as it moves. They are what enable a rat to find its way back to its nest or a dog to learn the regular route home from its daily walk. Animals have been tracking their location and navigating environments for millions of years, thanks to grid cells.

Recent experimental evidence suggests that grid cells also are present in the neocortex. In the paper, we refer to them as "cortical grid cells" in order to differentiate them from grid cells in the older part of the brain. **We propose that the neocortex uses cortical grid cells to learn the structure of objects in the same way that the older part of the brain uses grid cells to learn the structure of environments.** We learn objects as a set of locations, just as a rat learns an environment as a set of locations (see Figure 4).

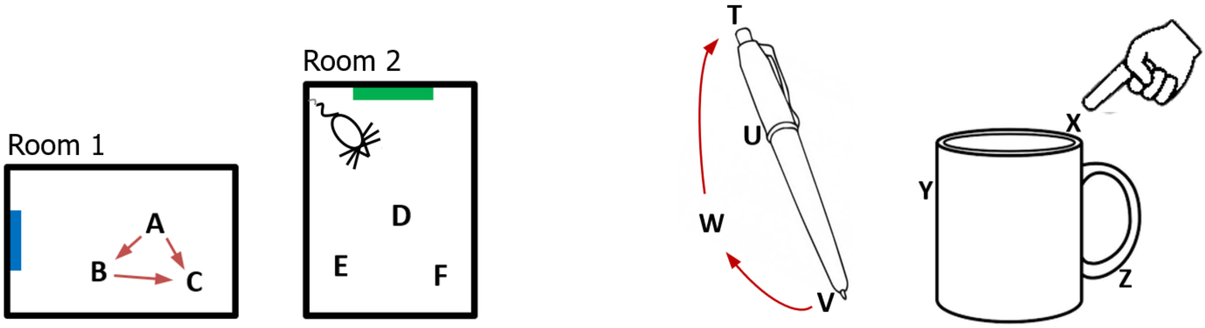

Figure 4. This is Figure 2 from the research paper, showing grid cells in the older part of the brain learn locations in an environment (left), and cortical grid cells learn locations on an object (right).

While grid cells in the older part of the brain track the location of one thing – your body – grid cells in the neocortex track the location of many things – your sensors – simultaneously. Think of your fingers typing on a keyboard, each one on a different key at any given time. Your brain has to know where each of your fingers are. It's as if each finger has its own GPS signal, but the signal is not relative to the finger, it's relative to the keyboard. While perhaps not as obvious, the same holds true for vision. If you're looking at someone's face, it's not like the brain takes a picture of that face as a whole. Different parts of your retina are looking at different parts of the face – maybe one on an eyelash, one on the tip of the nose and one at a freckle on a cheek – and each part knows where it's looking on the face.

The idea that cortical grid cells do something similar to grid cells in the older part of the brain seems logical in the context of evolution. It is easy to imagine natural selection leveraging this powerful mechanism in the neocortex. It's much harder to imagine evolution coming up with a completely different mechanism.

Object Compositionality

Our hypothesis for cortical grid cells suggested a solution to another problem that we've known we had to solve for years, but didn't know how: **object compositionality**. Almost everything we know is composed of other things we already have learned, which allows the brain to learn new things efficiently without constantly having to learn everything from scratch. For example, think of how many individual objects you understand in your car. You know how a steering wheel works, how to turn an ignition key, how to buckle your seatbelt, and so on. When you see a different kind of vehicle for a first time, say a large truck, you find all the car objects you have previously learned. You have some new things to learn, such as different mirrors or a trailer hitch, but you do not need to relearn the components you know already.

The paper uses an example of a coffee cup, which is composed of a cylinder, a handle, and a logo. The logo is a separate object that was learned previously. The idea that objects are learned as a set of locations led to a proposal of a new type of cell that exists in your brain. In the paper, we call them **displacement cells.** They complement grid cells and elegantly solve the composition problem. Displacement cells describe the relationship of objects, like the logo to the mug.

Our theory predicts that displacement cells must exist to enable us to shift back and forth between the cup and the logo with no confusion. Every time you learn a new object, displacement cells enable your brain to represent that object as a collection of objects you've previously learned, arranged in a particular way.

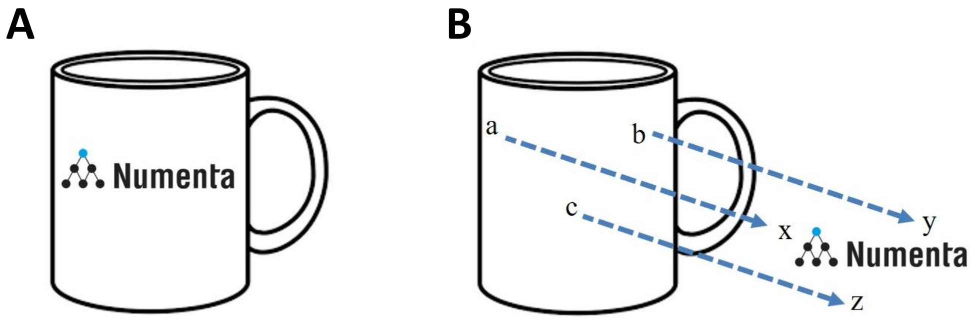

Figure 5. This is part of Figure 3 from the research paper. It shows how displacement cells, represented by the blue arrows, learn the relationship of the logo to the mug.

In Figure 5, displacement cells describe the relationship of the logo to the cup. Because the logo is fixed to the mug, the displacement cells do not change as you move the mug – the mug and the logo move together. But that is not always the case. Sometimes objects have parts that move independently, which leads us to an explanation for how objects behave.

Object Behaviors

Many objects we encounter in the world exhibit behaviors – sometimes simple, sometimes quite complex. Think of all the objects you have learned and whose behavior you understand, whether a toaster, a car, a zipper or your smartphone. When you unlock your phone, for example, the home screen appears; when you tap on an icon, that application opens; when you touch a text box, the keyboard pops up. Your brain has learned all of these behaviors and more, but how? Our theory suggests a solution.

We use a stapler as an example in the paper, which is simpler than a smartphone but still instructive. When you open a stapler to insert new staples, it has a different shape than when it was closed, even though it is the same object. The location of some parts of the stapler have changed relative to the other parts. The object has "behaved" in some way. Unlike the mug, where the logo was fixed in one spot relative to the mug, the top and the bottom of the stapler move relative to each other as you open and close it. But they don't move randomly; they move in a consistent and meaningful way.

The position of the top of the stapler relative to the bottom of the stapler is represented by **displacement cells** just as was the logo and cup. But now as the top moves, the displacement cells change. Therefore **learning a behavior is learning a sequence of displacements**. In other words, we learn behaviors as the sequence of how the relative positions of two locations change over time (in this case the top and bottom of the stapler).

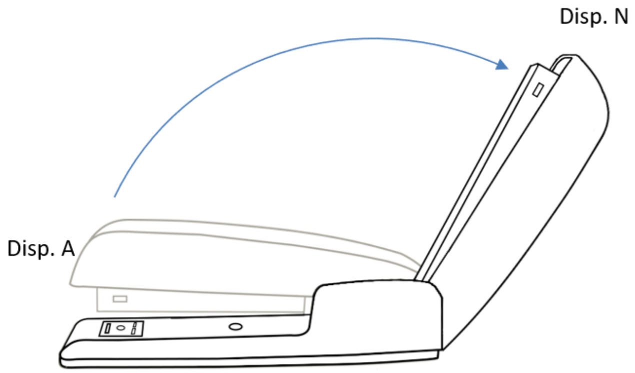

Figure 6. This is Figure 4 from the research paper. It shows how we learn behaviors of objects by learning how their relative positions change through movement.

Representing High-level Concepts

We've explained how we learn about an object such as a cat, mug or stapler -- how we identify it and how we learn its behavior. But what about other forms of intelligence? How do we learn physics or engineering? How do we learn concepts like "democracy" or "fairness?" Our theory suggests a provocative answer to this question. As stated earlier, Mountcastle proposed that all regions of the neocortex have the same complex circuitry and therefore they must be doing the same thing. If the regions of the neocortex that process sensory input use grid cells and locations to learn objects, then the **evidence strongly suggests that the regions of the neocortex that learn language or do math also use grid cells and locations to perform these functions**.

We don't fully understand the implications of this idea yet. In the research paper, we cite some intriguing work from other researchers that supports this assertion. For example, one such paper showed that people thinking about birds would mentally assign birds into a "space of bird attributes" and that imagining new birds was equivalent to moving in the space of bird attributes. Humans performing this task were not necessarily aware that they were thinking about birds in this way, but researchers detected it with brain imaging.

The “method of loci” is a memory enhancement technique that involves assigning things you want to remember to locations in a familiar space, like your house. When you want to recall the items in the list, you “move” through that space to retrieve them. The reason this method works is because **the brain naturally wants to associate things to locations**. Something similar is likely happening for everything you learn.

We don’t know yet how concepts such as democracy or fairness are represented in the neocortex, but we can be confident that locations will play a key part. This notion is an exciting area for further exploration.

Summary of the framework

Before we go further, let’s review the main points of the paper:

- Thousand Brains Theory of Intelligence – because every cortical column can learn complete models of objects, the brain creates thousands of models simultaneously, rather than one big model.
- Location, location, location – every cortical column has a location signal, which we propose is implemented by cortical grid cells.
- Object compositionality – we propose a new type of neuron exists throughout the brain, “displacement cells,” and they enable us to learn how objects are composed of other objects.
- Object behaviors – learning an object’s behavior is simply learning the sequence of movements tracked by displacement cells.
- Concepts and high-level thought – we learn conceptual ideas in the same way we learn physical objects.

Implications

Neuroscience

Understanding the complex circuitry in the neocortex is one of the key goals of neuroscientists, and our framework may provide a useful map with guideposts. It’s important that the framework fits within the known biological constraints. In the paper, we explain how the framework fits the anatomy of the neocortex, and we propose where grid cells and displacement cells are. We look forward to continued collaboration with experimental neuroscientists as they work to validate these hypotheses, fill in more details, and move to an even more complete understanding of how the brain is intelligent.

Artificial Intelligence (AI) and Robotics

Many AI researchers have concluded that today’s AI techniques are limited. Current AI networks are used today mainly for pattern recognition. If you train them on enough images, they can take a new image and tell you what it is, yet they don’t understand anything about an object beyond its label. They don’t know about the structure of an object or how it behaves. They’re also very brittle. The smallest change or occlusion of an image often makes an algorithm fail.

In order to create true AI, or what some refer to as Artificial General Intelligence (AGI), we will need models that can understand compositional structure and learn through movement. We'll need models that can apply learning in one modality, such as touch, and integrate it with another, such as vision. The Thousand Brains Theory, which is built on the common cortical algorithm, provides a single architecture that can learn anything – from objects, to behaviors of objects, to concepts.

We can't know all the applications our theories might yield, but one can imagine how they might be applied to robotics. Today's robots are highly constrained. They can execute programmed activities but are unable to explore and understand the world without an explicit set of instructions. Everything in the neocortex is built around movement. If our theories help to create robots who can learn robust models of the world from their sensory data, they are more likely to be useful in applications such as investigating dangerous situations or exploring remote planets.

We hope the Thousand Brains Theory can offer a path forward for AI and robotics to break through some of the limitations of current techniques. Having a framework for how the neocortex works has numerous potential implications across other disciplines as well, from philosophy to pedagogy to understanding neurological diseases.

Conclusion

In this companion paper, we've given an overview of the framework we are proposing, the problems we believe it addresses and the implications we'll likely see. We invite you to read the paper for a more thorough discussion.

For more than a hundred years, people have sought to understand how the brain works. From scientists to philosophers, from medical doctors to teachers, humankind has yearned for a deeper understanding of what it means to be intelligent. We believe that the Thousand Brains Theory is an exciting response to Dr. Crick's challenge to discover a theoretical framework for intelligence, one which we hope will enable the next generation of neuroscientists and computer scientists to make great further progress.

For more reading

There are several other non-technical resources that expound on these ideas.

Thousand Brains Model of Intelligence (blog)
What's Old is New: The evolutionary context of the neocortex (blog)
Method of Loci (blog)
Numenta Research: Key Discoveries in Understanding How the Brain Works (video)