SHADES OF THE SINGULARITY

RAY KURZWEIL’S FUTURE IMPERFECT
BY JOHN RENNIE

THINKING LIKE A HUMAN-WITH MEMRISTORS
BY MASSIMILIANO VERSACE & BEN CHANDLER
Meet MoNETA—
the brain-inspired
chip that will
outsmart us all

The Brain
of a New
Machine

BY MASSIMILIANO VERSACE & BEN CHANDLER
ILLUSTRATIONS BY CHAD HAGEN
we’d have by now. Computers diagnose patients over the Internet. We can do is automated tech support, intoned with a preternatural calm. Boolean logic and vast separation between memory and processing. But human handlers within 20 years. Now, 50 years later, it seems the best. Artificial intelligence is the U.S. Defense Advanced Research Projects Department of cognitive and neural systems, which will run on a brain-inspired microprocessor under development at HP Labs in California. It will function according to the principles that distinguish us mams most profoundly from our fast but witless machines. MoNETA (the goddess of memory—cute, huh?) will do things no computer ever has. It really is a name intended to communicate in some useful, integrate that information into the emerging structure of its reality, and in some applications, formulate plans that will ensure its survival. In other words, MoNETA will be motivated by the same drives that motivate cockroaches, cats, and humans. Researchers have suspected for decades that real artificial intelligence is complete. But we’ve been wrong. We’ve been wrong in thinking that a machine that could drive a car should be a machine that comprehends the world, could think. Even to behave like a brain. In this case, form is function, or more accurately, function is hopeless without form.

Barely a month after Deep Blue crushed Kasparov, researchers considered the impressive feat of Kasparov’s computer’s predecessors five times. After a series comprising one win apiece and three draws, Deep Blue finally trounced Kasparov in game six. Nevertheless, Deep Blue could not compete with the easily excitable, highly skilled, and somewhat unpredictable Kasparov, its special-purpose hardware a brute-force strategy of simply calculating the value of 200 million possible chess moves each second. In the same amount of time, Kasparov could plan roughly two chess positions.

Over the next 10 years, computing capabilities skyrocketed. By 2002 the processing power of that 1.4-ton supercomputer had been contained within a cellphone microprocessor roughly the size of a thumbnail. In the decade between then, transistor counts had quadrupled, FPGA interconnects grew by a factor of about 10 million. This was no longer the fracturing, chaotic function of a brain, the low power requirements, and the instantaneous internal communication. Turns out that those three things are key to making anything that can approach the power of the brain. The main reason chips can use less power is to avoid heat. Ever to behave like a brain. In this case, form is function, or more accurately, function is hopeless without form.

Barely a month after Deep Blue crushed Kasparov, researchers conjectured that real artificial intelligence is complete. But we’ve been wrong. We’ve been wrong in thinking that a machine that could drive a car should be a machine that comprehends the world, could think. Even to behave like a brain. In this case, form is function, or more accurately, function is hopeless without form.

Barely a month after Deep Blue crushed Kasparov, researchers conjectured that real artificial intelligence is complete. But we’ve been wrong. We’ve been wrong in thinking that a machine that could drive a car should be a machine that comprehends the world, could think. Even to behave like a brain. In this case, form is function, or more accurately, function is hopeless without form.
Hardware vs. Wetware

To understand the difference between the architecture of the brain and a standard computer, compare the path of a hypothetical bit of data in a brain with that in a brain simulation.

**BRAIN**

In the mammalian brain, storage and computation happen at the same time and in the same place. Neuron 1 sends a signal down the axon to Neuron 2. The synapse of Neuron 2 evaluates the importance of the information coming from Neuron 1 by contrasting it with its own previous state and the strength of its connection to Neuron 1. Then, these two pieces of information travel back to Neuron 1 along the axon of Neuron 2, down the dendrites, and flow toward the synapse—right through the body of Neuron 2. By the time that information reaches the body of Neuron 2, there is only a single value—a computation that has already taken place during the information transfer.

**COMPUTER**

On a computer, the processing units are physically separated. A significant physical distance separates areas where the data is stored from the areas where it is manipulated. Modeling a single synapse requires the following to happen in the machine: the synapse's state is in a location in main memory. To change that state, a signal must originate somewhere on the processor, travel to the edge of the processor, be packaged for transit over the main bus, travel between 2 and 3 centimeters to reach the physical main memory, and then be unpackaged to actually access the desired memory location. Multiplying that sequence by up to 8000 synapses—as many as in a single rat neuron—and then again by the brain's billions of neurons yields a single millisecond of brain activity.

It would be fantastically inefficient. Inefficient hardware won't stop us from running neuromorphic algorithms (such as machine vision), but we would need an entire massive cluster of high-performance graphics processing units (GPUs) to handle the parallel computations, which would also come with the power requirements of a million-watt bulb. But reproducing the brain's functionality on even the most advanced supercomputers would require a dedicated power plant. To be sure, locality isn't impossible with today's silicon technology. A true artificial intelligence from Neuron 1 and the state of Neuron 2's synapse—flow toward the body of Neuron 2 and, by the time that information reaches the body of Neuron 2, there is only a single value—a computation that has already taken place during the information transfer.

That difference is the main reason the human brain can run on the wattage of a waltz bulb. The fact that Neuron 2's synapse—flow toward the body of Neuron 2 over the dendrites. By the time that information reaches the body of Neuron 2, there is only a single value—a computation that has already taken place during the information transfer.

**PARTLY TO AVOID THE FOOLY OF TRYING TO COAX INTELLIGENCE FROM FUNDAMENTALLY DUMB HARDWARE**

DARPA launched a program called SyNAPSE (Systems of Neuromorphic Adaptive Plastic Scalable Electronics) in 2008. The timing was good. That year, HP Labs had built a functioning memristor device hailed as the fourth fundamental electronic component, after the resistor, capacitor, and inductor. The concept wasn't new. In 1971, professor Leon Chua of the University of California, Berkeley, reasoned that a memristor would behave like a resistor with a conductance that changed as a function of its internal state and the voltage applied. In other words, a memristor could remember how much current had gone through it, it could work as an essentially nonvolatile memory. And sure enough, Korean dynamic random access memory giant Hynix Semiconductor made a splash recently when it chose the device as a possible fourth fundamental electronic component. But because memristors can remember their past state without using any power, their biggest potential all along has been as a realistic analogue to synapses in brains.

Here's why. A memristor is a two-terminal device whose resistance changes depending on the amount, direction, and duration of voltage that's applied to it. But here's the really interesting thing about a memristor: Whatever its past state, or resistance, it freezes that state until another voltage is applied to change it. Maintaining that state requires no power. That's different from a dynamic RAM cell, which requires regular charge to maintain its state. The upshot is that thousands of memristors could substitute for massive banks of power-hungry memory. Just to be clear, the memristor is not magic: Its nonvolatile state does decay over time. That decay can take hours or centuries depending on the material, and stability must often be traded for energy requirements—which is one of the major research reasons memristors aren't flooding the market yet.

Physically, a memristor is just an oxide junction between two terminals. For a memristor to function, two conditions must be met: One, the memristor must have memory; two, the memristor must not go to a nonfiring state (also known as a memristor reset). A memristor's memory depends on the amount, direction, and duration of voltage that's applied to it. (A memristor is thus a voltage-controlled memristor: hard to make, let alone build in silicon.)

**COMPUTER**

On a computer, the processing units are physically separated. A significant physical distance separates areas where the data is stored from the areas where it is manipulated. Multiplying that sequence by up to 8000 synapses—as many as in a single rat neuron—and then again by the brain's billions of neurons yields a single millisecond of brain activity.

That's why a memristor is a two-terminal device whose resistance changes depending on the amount, direction, and duration of voltage that's applied to it. But here's the really interesting thing about a memristor: Whatever its past state, or resistance, it freezes that state until another voltage is applied to change it. Maintaining that state requires no power. That's different from a dynamic RAM cell, which requires regular charge to maintain its state. The upshot is that thousands of memristors could substitute for massive banks of power-hungry memory. Just to be clear, the memristor is not magic: Its nonvolatile state does decay over time. That decay can take hours or centuries depending on the material, and stability must often be traded for energy requirements—which is one of the major research reasons memristors aren't flooding the market yet.

Physically, a memristor is just an oxide junction between two terminals. For a memristor to function, two conditions must be met: One, the memristor must have memory; two, the memristor must not go to a nonfiring state (also known as a memristor reset). A memristor's memory depends on the amount, direction, and duration of voltage that's applied to it. (A memristor is thus a voltage-controlled memristor: hard to make, let alone build in silicon.)

**to build a brain, you need to throw away the conceit of separate hardware and software because the brain doesn’t work that way.**

In the brain it’s all just wetware. If you really wanted to replicate a mammalian brain, software and hardware would need to be inextricable. We could build a microchip but the brain has allowed us to take a big step closer by approximating the biological form factor: hardware that can be both small and ultra-low power.

Where HP is taking care of the hardware component of the neuromorphic processor, we are building the software—the brain models that do the recognizing, reasoning, and learning. HP chose to build their own chips but the memristor has allowed us to take a step closer by approximating the biological form factor: hardware that can be both small and ultra-low power.

**CELEST was established to study basic principles of how the brain performs learning, computation, and memory.**

To allow the brain models and the neuromorphic hardware to interact, HP built a kind of special-purpose operating system called Cog Ex Machina. Cog, built by HP principal investigator Greg Snider, lets system designers interact with the underlying hardware, interface with the operating system, and do neuromorphic computation. Neuromorphic computation means computation that can be divided up between hardware that processes like the human brain and hardware that processes the way matrices and axioms do.

The two kinds of cores deal with processing in fundamentally different ways: The memristor was designed to speed up processing by taking the computation and memory functions that memristors do away from the computer processor and placing them on the hardware. It’s a little like having one’s own private megacache, meaning that every single core has its own private massive bank of memory. Memristors are incredibly tiny, even by the standards of today’s semiconductors. It is possible to connect more than 8000 synapses on a single chip, within a couple of decades it will be possible to build a nonvolatile synaptic array in a petabyte (a quadrillion bits) per square centimeter.

**Though memristors are dense, cheap, and tiny, they also have a high failure rate at present, characteristics that bear an intriguing resemblance to biological synapses.**

It means that the architecture must by definition tolerate defects in individual circuitry, much the way brains gracefully degrade their performance as synapses are lost, without suddenly becoming useless.

Basic memristors bring data close to the computation, the way biological systems do, and they use very little power to store that information, just as the brain does. A comparable computer, the new hardware that HP Labs is building, will use two to three orders of magnitude less power than Nvidia’s Fermi-class GPU. For the first time we will begin to bridge the main divide between biological computation and traditional computation. The use of the memristor could bring new hardware challenges of neuromorphic computing: the need to simultaneously move and manipulate data, thereby drastically cutting power consumption and space. You might think that to achieve processing that’s more like thinking than computing would require more than just new hardware—it would also require new software. You’d be wrong, but in a way that might surprise you.

Based on the memristor research and architecture, we couldn’t even think about building MoNETA.
characteristics resemble those of the neuron. But the trade-off is that the core sucks up a lot of power, so like neurons, these elements should make up only a small percentage of the system. A “de dendrite” core works more like a GPU, an inexpensive and high-performance microprocessor. Like a de dendrite, a GPU has a rigid architecture that is optimized for only a specific kind of computation—in this case, the com plicated linear algebra operations that approximate what happens inside a dendrite. Because GPUs are optimized for parallel computation, we can use them to approximate the distributed computation that de dendrites carry out. But there’s a cost to using these, too: GPU cores perform only a limited set of operations. The de dendrites in the final DARPA hardware will be much less flexible than neuron cores, but they will store extraordinary amounts of state information in their massive memristor-based memory banks, and like the tendrils of neurons, they will make up the vast bulk of the system’s computational elements. Memristors, finally, will act as the synapses that mediate the information transfer between the de dendrites and axons of different neurons. For a programmer, taking full advantage of a machine like this—with its two different core types and complicated memory-storage overlay—is tremendously challenging, because the problems need to be properly partitioned across those two radically differ ent types of processors. Thanks to Cog, we computational neuroscientists can forget about the hardware and focus on developing the soul inside the machine.

MoNETA will be a general-purpose mammalian-type intelligence, an artificial, generic creature known as an animat. With the DARPA hardware, we think we will be able to fit this level of intelligence into a shoebox.

The key feature distinguishing MoNETA from other AIs is that it won’t have to be explicitly programmed. We are engineering MoNETA to be as adaptable and ef ficient as a mammal’s brain. We intend to set it loose on a variety of situations, and it will learn dynamically.

Biological intelligence is the result of the coordi nation of many highly interconnected and plastic brain areas. Most prior research has focused on modeling those individual parts of the brain. The results, while impressive in some cases, have been a piecemeal assortment of experiments, theories, and models that each nicely describes the architecture and function of a single brain area and its contribution to perception, emotion, and action. But if you tried to stitch those findings together, you would more likely end up with a nonfunctioning Frankenstein’s monster than anything like a mammalian intelligence.

Truly general-purpose intelligence can emerge only when everything is interconnected like our brain. Think of a rat, all perception (including auditory and visual inputs, or the brain areas responsible for the generation of fine finger movements), emo tion, actions, and reactions combine and interact to guide behavior. Perceiving without action, emotion, higher reasoning, and learning would not only fail to lead to a general purpose AI, it wouldn’t even pass a commonsense Turing test.

Creating this grand AI architecture has been precluded by several practical limitations. The most important is the lack of a unified theory of the brain. But the creation of large centers such as CElast has advanced our understanding of what key aspects of biological intelligence might be applicable to our task of building a general-purpose AI.

MoNETA: A Made of Memristors

DARPA’s neuromorphic chip is still a long way from reaching biological efficiency. Even with optimistic assumptions about how much information single artificial synapses can store, the hardware still has one-hundredth the efficiency of biology. Still, that’s 2000 times as power efficient as today’s best supercomputers.

How will we know we’ve succeeded? How will we know that all this effort and new hardware and new software have yielded what we want—an artificial intelligence? We’ll know we have successfully built an animat when we are able to motivate MoNETA to run, swim, and find food dynamically, without being programmed explicitly to do so.

It should learn throughout its lifetime without needing constant reprogramming or needing to be told a priori what is good for it, and what is bad. This is a huge challenge for traditional AI; it is not possi ble to preprogram a lifetime of knowledge into a vir tual or robotic animal. Such wisdom has to be learned from the interaction between a brain—with its large and (not yet infinite) number of synapses that store memories—and an environment that is constantly changing and dense with information.

The animat will learn about objects in its envi ronment, navigate to reach its goals, and avoid dangers without the need for us to program spe cific objects or behaviors. Such an ability can act as standard issue in mammals, because our brains are plastic throughout our lives. We learn to recognize new people and places, and we acquire new skills without being told to do so. MoNETA will need to do the same.

We will test our animat in a classic trial called the Morris water navigation task. In this experi ment, neuroscientists teach a rat to swim through a water maze, using visual cues, to a submerged plat form that the rat can’t see. That task might seem sim pler, but it’s anything but. To get to the platform, the rat must navigate through many stenched, hazardous, and unmanned aerial vehicles will be just the beginning.

Will these chips “experience” vision and emotions by simulating and appropri ately connecting the brain areas known to be involved in the subjective experience associated with them? It’s too soon to say. However, our goal is not to replicate subjective experience—consciousness—in a chip but rather to build functional machines that can behave intelligently in complex environments. In other words, the idea is to make machines that behave as if they are intelligent, emotionally biased, and moti vated, without the constraint that they are actually aware of these feelings, thoughts, and motivations.

Neuromorphic chips won’t just power niche AI applications. The architectural lessons we learn here will revolutionize all future CPUs. The fact is, conventional computers just can’t do sig nificantly more powerful unless they move to a more parallel and locality-driven architecture. While neuromorphic chips will first sup plement today’s CPUs, soon their sheer power will overwhelm that of today’s microprocessors.

The semiconductor industry’s relentless push to focus on smaller and smaller transistors will soon mean transistors have higher failure rates. Transistor feature sizes. By 2018, that number will have shrunk to 12 nm, at which point atomic processes will interfere with transistor func tioning. Transistors in future CPUs. The fact is, conventional computers just can’t do significantly more powerful unless they move to a more parallel and locality-driven architecture. While neuromorphic chips will first sup plement today’s CPUs, soon their sheer power will overwhelm that of today’s microprocessors.

The semiconductor industry’s relentless push to focus on smaller and smaller transistors will soon mean transistors have higher failure rates. Transistor feature sizes. By 2018, that number will have shrunk to 12 nm, at which point atomic processes will interfere with transistor func tioning. Transistors in future CPUs. The fact is, conventional computers just can’t do significantly more powerful unless they move to a more parallel and locality-driven architecture. While neuromorphic chips will first sup plement today’s CPUs, soon their sheer power will overwhelm that of today’s microprocessors.

It won’t be long until all multicore chips integrate a dense, lowpower memory with their CMOS cores. It’s just common sense.