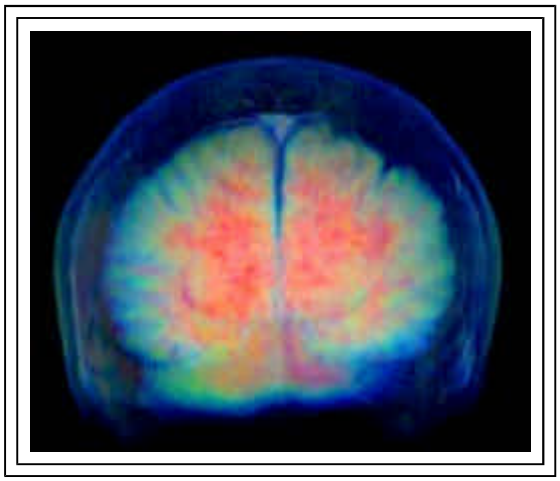
Is this a unified theory of the brain?

28 May 2008
From New Scientist Print Edition.
Gregory T. Huang



The quest to understand the most complex object in the known universe has been a long and fruitful one. These days we know a good deal about how the human brain works - how our senses translate into electrical signals, how different parts of the brain process these signals, how memories form and how muscles are controlled. We know which brain regions are active when we listen to speech, look at paintings or barter over money. We are even starting to understand the deeper neural processes behind learning and decision-making.

What we still don't have, though, is a way to bring all these pieces together to create an overarching theory of how the brain works. Despite decades of research, neuroscientists have never been able to produce their own equivalent of Schrödinger's equation in quantum mechanics or Einstein's *E=mc*2 - a powerful, concise, mathematical law that encapsulates how the brain works. Nor do they have a plausible road map towards a "theory of everything", like string theory in physics. Surely if we can get so close to explaining the universe, the human brain can't be that hard to crack?

Perhaps it is. The brain is much messier than a physical system. It is the product of half a billion years of evolution. It performs myriad functions - reasoning, memory, perception, learning, attention and emotion to name just a few - and uses a staggering number of different types of cells, connections and receptors. So it does not lend itself to being easily described by simple mathematical laws.

That hasn't stopped researchers in the growing field of computational neuroscience from trying. In recent years, they have sought to develop unifying ideas about how the brain processes information so that they can apply them to the design of intelligent machines.

Until now none of their ideas has been general or testable enough to arouse much excitement in straight neuroscience. But a group from University College London (UCL) may have broken the deadlock. Neuroscientist Karl Friston and his colleagues have proposed a mathematical law that some are claiming is the nearest thing yet to a grand unified theory of the brain. From this single law, Friston's group claims to be able to explain almost everything about our grey matter.

It's a controversial claim, but one that's starting to make people sit up and take notice. Friston's work has made Stanislas Dehaene, a noted neuroscientist and psychologist at the College of France in Paris, change his mind about whether a Schrödinger equation for the brain might exist. Like most neuroscientists, Dehaene had been pessimistic - but not any more. "It is the first time that we have had a theory of this strength, breadth and depth in cognitive neuroscience," he says.

Friston's ideas build on an existing theory known as the "Bayesian brain", which conceptualises the brain as a probability machine that constantly makes predictions about the world and then updates them based on what it senses.

The idea was born in 1983, when Geoffrey Hinton of the University of Toronto in Canada and Terry Sejnowski, then at Johns Hopkins University in Baltimore, Maryland, suggested that the brain could be seen as a machine that makes decisions based on the uncertainties of the outside world. In the 1990s, other researchers proposed that the brain represents knowledge of the world in terms of probabilities. Instead of estimating the distance to an object as a number, for instance, the brain would treat it as a range of possible values, some more likely than others.

A crucial element of the approach is that the probabilities are based on experience, but they change when relevant new information, such as visual information about the object's location, becomes available. "The brain is an inferential agent, optimising its models of what's going on at this moment and in the future," says Friston. In other words, the brain runs on Bayesian probability. Named after the 18th-century mathematician Thomas Bayes, this is a systematic way of calculating how the likelihood of an event changes as new information comes to light (see *New Scientist*, 10 May, p 44, for more on Bayesian theory).

Over the past decade, neuroscientists have found that real brains seem to work in this way. In perception and learning experiments, for example, people tend to make estimates - of the location or speed of a moving object, say - in a way that fits with Bayesian probability theory. There's also evidence that the brain makes internal predictions and updates them in a Bayesian manner. When you listen to someone talking, for example, your brain isn't simply receiving information, it also predicts what it expects to hear and constantly revises its predictions based on what information comes next. These predictions strongly influence what you actually hear, allowing you, for instance, to make sense of distorted or partially obscured speech.

In fact, making predictions and re-evaluating them seems to be a universal feature of the brain. At all times your brain is weighing its inputs and comparing them with internal predictions in order to make sense of the world. "It's a general computational principle that can explain how the brain handles problems ranging from low-level perception to high-level cognition," says Alex Pouget, a computational neuroscientist at the University of Rochester in New York (*Trends in Neurosciences*, vol 27, p 712).

However, the Bayesian brain is not quite a general law. It is a collection of related approaches that each use Bayesian probability theory to understand one aspect of brain function, such as parsing speech, recognising objects or learning words. No one has been able to pull all these disparate approaches together, nor explain why the brain works like this in the first place. An overarching law, if one exists, should attempt to do this.

This is where Friston's work comes in. In the 1990s he was working next door to Hinton at UCL. At that time Hinton was beginning to explore the concept of "free energy" as it applies to artificial neural networks. Free energy originates from thermodynamics and statistical mechanics, where it is defined as the amount of useful work that can be extracted from a system, such as a steam engine. It is roughly equivalent to the difference between the total energy in the system and its "useless energy", or entropy.

Hinton realised that free energy was mathematically equivalent to a problem he was familiar with: the difference between the predictions made by an artificial neural network and what it actually senses. He showed that you could solve some tough problems in machine learning by treating this "prediction error" as free energy, and then minimising it.

Friston spent the next few years working out whether the same concept could underlie the workings of real brains. His insight was that the constant updating of the brain's probabilities could also be expressed in terms of minimising free energy. Around 2005 he proposed that a "free energy principle" explains at least one aspect of brain function - sensory perception.

As a simple example, take what happens when you glimpse an object in your peripheral vision. At first it is not clear what it is - or, as Friston would put it, there's a big error between your brain's prediction and what it senses. To reduce this prediction error, Friston reasoned that one of two things can happen: the brain can either change its prediction or change the way it gathers data from the environment (*Journal of Physiology* - Paris, vol 100, p 70). If your brain takes the second option you will instinctively turn your head and centre the object in your field of view. "It's about minimising surprise," he explains. "Mathematically, free energy is always bigger than surprise, therefore if you can minimise free energy you can avoid surprising encounters with the world."

Friston developed the free-energy principle to explain perception, but he now thinks it can be generalised to other kinds of brain processes as well. He claims that everything the brain does is designed to minimise free energy or prediction error (*Synthese*, vol 159, p 417). "In short, everything that can change in the brain will change to suppress prediction errors, from the firing of neurons to the wiring between them, and from the movements of our eyes to the choices we make in daily life," he says.

Take neural plasticity, the well-established idea that the brain alters its internal pathways and connections with experience. First proposed by Canadian psychologist Donald Hebb in the 1940s, it is thought to be the basic mechanism behind learning and memory.

Friston's principle accounts for the process by describing how individual neurons interact after encountering a

novel stimulus. Neuron A "predicts" that neuron B will respond to the stimulus in a certain way. If the prediction is wrong, neuron A changes the strength of its connection to neuron B to decrease the prediction error. In this case the brain changes its internal predictions until it minimises its error, and learning or memory forming is the result.

All well and good in theory, but how can we know whether real brains actually work this way? To answer this question, Friston and others have focused on the cortex, the 3-millimetre-thick mass of convoluted folds that forms the brain's outer surface. This is the seat of "higher" functions such as cognition, learning, perception and language. It has a distinctive anatomy: a hierarchy of neuronal layers, each of which has connections to neurons in the other levels.

Friston created a computer simulation of the cortex with layers of "neurons" passing signals back and forth. Signals going from higher to lower levels represent the brain's internal predictions, while signals going the other way represent sensory input. As new information comes in, the higher neurons adjust their predictions according to Bayesian theory. This may seem awfully abstract, but there's a concrete reason for doing it: it tells Friston what patterns of activity to look for in real brains.

Last year Friston's group used functional magnetic resonance imaging to examine what is going on in the cortex during a visual task (*NeuroImage*, vol 34, p 1199). Volunteers watched two sets of moving dots, which sometimes moved in synchrony and at others more randomly, to change the predictability of the stimulus. The patterns of brain activity matched Friston's model of the visual cortex reasonably well. He argues that this supports the idea that top-down signals are indeed sent downstream to reduce prediction errors.

More recently, Friston's team has shown that signals from higher levels of the auditory cortex are responsible for modifying brain activity in lower levels as people listen to repeated and predictable sounds (*Proceedings of the National Academy of Sciences*, vol 104, p 20961). This, too, fits with Friston's model of top-down minimisation of prediction error.

Despite these successes, some in the Bayesian brain camp aren't buying the grand theory just yet. They say it is hard to know whether Friston's results are ground-breaking or just repackaged old concepts - but they don't say he's wrong. Others say the free-energy principle is not falsifiable. "I do not think it is testable, and I am pretty sure it does not tell you how to build a machine which emulates some aspect of intelligence," says theoretical neuroscientist Tomaso Poggio of the Massachusetts Institute of Technology.

Friston disagrees, pointing out that there are experiments that would definitively test whether or not a given population of neurons is minimising prediction error. He proposes knocking out a higher region of the cortex - using transcranial magnetic stimulation, say - and seeing whether free-energy models can predict how the activity of a lower region of neurons would change in response.

Several groups are planning experiments along these lines, but they need to work out exactly which neurons to target. "This would, I think, be an aspect of the theory that could be proved or falsified," says Thomas Wennekers, a computational neuroscientist at the University of Plymouth in the UK.

Meanwhile, Friston claims that the free-energy principle also gives plausible explanations for other important features of the cortex. These include "adaptation" effects, in which neurons stop firing after prolonged exposure to a stimulus like a rattling fan, so after a while you don't hear it. It also explains other phenomena: patterns of mirror-neuron activation that reflect the brain's responses to watching someone else make a movement; basic communication patterns between neurons that might underlie how we think; and even the hierarchical anatomy of the cortex itself.

Friston's results have earned praise for bringing together so many disparate strands of neuroscience. "It is quite certainly the most advanced conceptual framework regarding an application of these ideas to brain function in general," says Wennekers. Marsel Mesulam, a cognitive neurologist from Northwestern University in Chicago, adds: "Friston's work is pivotal. It resonates entirely with the sort of model that I would like to see emerge."

So where will the search for a unified theory of the brain go from here? Friston's free-energy principle clearly isn't the ultimate theory yet it remains to be tested fully and needs to produce more predictions of how real brains behave. If all goes well, though, the outcome will be a concise mathematical law of brain function, perhaps something as brief and iconic as $E=mc2$. "The final equation you write on a T-shirt will be quite simple," Friston predicts.

On a more practical level, he says the approach will change our concepts of how the brain works and could help us understand the deeper mechanisms of psychological disorders, especially those thought to be caused by faulty connections in the cortex, such as schizophrenia. It could also shine a light on bigger questions such as the nature of human consciousness.

There's work still to be done, but for now Friston's is the most promising approach we've got. "It will take time to spin off all of the consequences of the theory - but I take that property as a sure sign that this is a very important theory," says Dehaene. "Most other models, including mine, are just models of one small aspect of the brain, very limited in their scope. This one falls much closer to a grand theory."

Gregory T. Huang is a journalist based in Seattle

From issue 2658 of New Scientist magazine, 28 May 2008, page 30-33