# The Genius Neuroscientist Who Might Hold the Key to True AI

wired.com/story/karl-friston-free-energy-principle-artificial-intelligence

Shaun Raviv

When King George III of England began to show signs of acute mania toward the end of his reign, rumors about the royal madness multiplied quickly in the public mind. One legend had it that George tried to shake hands with a tree, believing it to be the King of Prussia. Another described how he was whisked away to a house on Queen Square, in the Bloomsbury district of London, to receive treatment among his subjects. The tale goes on that George's wife, Queen Charlotte, hired out the cellar of a local pub to stock provisions for the king's meals while he stayed under his doctor's care.

More than two centuries later, this story about Queen Square is still popular in London guidebooks. And whether or not it's true, the neighborhood has evolved over the years as if to conform to it. A metal statue of Charlotte stands over the northern end of the square; the corner pub is called the Queen's Larder; and the square's quiet rectangular garden is now all but surrounded by people who work on brains and people whose brains need work. The National Hospital for Neurology and Neurosurgery—where a modern-day royal might well seek

treatment—dominates one corner of Queen Square, and the world-renowned neuroscience research facilities of University College London round out its perimeter. During a week of perfect weather last July, dozens of neurological patients and their families passed silent time on wooden benches at the outer edges of the grass.

*Axis of Strength*

On a typical Monday, Karl Friston arrives on Queen Square at 12:25 pm and smokes a cigarette in the garden by the statue of Queen Charlotte. A slightly bent, solitary figure with thick gray hair, Friston is the scientific director of University College London's storied Functional Imaging Laboratory, known to everyone who works there as the FIL. After finishing his cigarette, Friston walks to the western side of the square, enters a brick and limestone building, and heads to a seminar room on the fourth floor, where anywhere from two to two dozen people might be facing a blank white wall waiting for him. Friston likes to arrive five minutes late, so everyone else is already there.

His greeting to the group is liable to be his first substantial utterance of the day, as Friston prefers not to speak with other human beings before noon. (At home, he will have conversed with his wife and three sons via an agreed-upon series of smiles and grunts.) He also rarely meets people one-on-one. Instead, he prefers to hold open meetings like this one, where students, postdocs, and members of the public who desire Friston's expertise—a category of person that has become almost comically broad in recent years—can seek his knowledge. "He believes that if one person has an idea or a question or project going on, the best way to learn about it is for the whole group to come together, hear the person, and then everybody gets a chance to ask questions and discuss. And so one person's learning becomes everybody's learning," says David Benrimoh, a psychiatry resident at McGill University who studied under Friston for a year. "It's very unique. As many things are with Karl."

At the start of each Monday meeting, everyone goes around and states their questions at the outset. Friston walks in slow, deliberate circles as he listens, his glasses perched at the end of his nose, so that he is always lowering his head to see the person who is speaking. He then spends the next few hours answering the questions in turn. "A Victorian gentleman, with Victorian manners and tastes," as one friend describes Friston, he responds to even the most confused questions with courtesy and rapid reformulation. The Q&A sessions—which I started calling "Ask Karl" meetings—are remarkable feats of endurance, memory, breadth of knowledge, and creative thinking. They often end when it is time for Friston to retreat to the minuscule metal balcony hanging off his office for another smoke.

Friston first became a heroic figure in academia for devising many of the most important tools that have made human brains legible to science. In 1990 he invented statistical parametric mapping, a computational technique that helps—as one neuroscientist put it—"squash and squish" brain images into a consistent shape so that researchers can do apples-to-apples comparisons of activity within different crania. Out of statistical parametric mapping came a

corollary called voxel-based morphometry, an imaging technique that was used in one famous study to show that the rear side of the hippocampus of London taxi drivers grew as they learned “the knowledge.”[1]

[1] To earn a London taxi license, drivers must memorize 320 routes and many landmarks within 6 miles of Charing Cross. The grueling process includes a written test as well as a series of one-on-one meetings with an examiner.

A study published in *Science* in 2011 used yet a third brain-imaging-analysis software invented by Friston—dynamic causal modeling—to determine if people with severe brain damage were minimally conscious or simply vegetative.

When Friston was inducted into the Royal Society of Fellows in 2006, the academy described his impact on studies of the brain as “revolutionary” and said that more than 90 percent of papers published in brain imaging used his methods. Two years ago, the Allen Institute for Artificial Intelligence, a research outfit led by AI pioneer Oren Etzioni, calculated that Friston is the world’s most frequently cited neuroscientist. He has an h-index—a metric used to measure the impact of a researcher’s publications—nearly twice the size of Albert Einstein’s. Last year Clarivate Analytics, which over more than two decades has successfully predicted 46 Nobel Prize winners in the sciences, ranked Friston among the three most likely winners in the physiology or medicine category.

What’s remarkable, however, is that few of the researchers who make the pilgrimage to see Friston these days have come to talk about brain imaging at all. Over a 10-day period this summer, Friston advised an astrophysicist, several philosophers, a computer engineer working on a more personable competitor to the Amazon Echo, the head of artificial intelligence for one of the world’s largest insurance companies, a neuroscientist seeking to build better hearing aids, and a psychiatrist with a startup that applies machine learning to help treat depression. And most of them had come because they were desperate to understand something else entirely.



For the past decade or so, Friston has devoted much of his time and effort to developing an idea he calls the free energy principle. (Friston refers to his neuroimaging research as a day job, the way a jazz musician might refer to his shift at the local public library.) With this idea, Friston believes he has identified nothing less than the organizing principle of all life, and all intelligence as well. “If you are alive,” he sets out to answer, “what sorts of behaviors *must* you show?”

First the bad news: The free energy principle is maddeningly difficult to understand. So difficult, in fact, that entire rooms of very, very smart people have tried and failed to grasp it. A Twitter account[2] with 3,000 followers exists simply to mock its opacity, and nearly every person I spoke with about it, including researchers whose work depends on it, told me they didn't fully comprehend it.

[2] The account is called @FarlKriston. Sample tweet: "Life is an inevitable & emergent property of any (ergodic) random dynamical system that possesses a Markov blanket. Don't leave with out it!"

But often those same people hastened to add that the free energy principle, at its heart, tells a simple story and solves a basic puzzle. The second law of thermodynamics tells us that the universe tends toward entropy, toward dissolution; but living things fiercely resist it. We wake up every morning nearly the same person we were the day before, with clear separations between our cells and organs, and between us and the world without. How? Friston's free energy principle says that all life, at every scale of organization—from single cells to the human brain, with its billions of neurons—is driven by the same universal imperative, which can be reduced to a mathematical function. To be alive, he says, is to act in ways that reduce the gulf between your expectations and your sensory inputs. Or, in Fristonian terms, it is to *minimize free energy*.

To get a sense of the potential implications of this theory, all you have to do is look at the array of people who darken the FIL's doorstep on Monday mornings. Some are here because they want to use the free energy principle to unify theories of the mind, provide a new foundation for biology, and explain life as we know it. Others hope the free energy principle will finally ground psychiatry in a functional understanding of the brain. And still others come because they want to use Friston's ideas to break through the roadblocks in artificial intelligence research. But they all have one reason in common for being here, which is that the only person who truly understands Karl Friston's free energy principle may be Karl Friston himself.


Friston's office. A friend describes him as "a Victorian gentleman, with Victorian manners and tastes."

*Kate Peters*

Friston isn't just one of the most influential scholars in his field; he's also among the most prolific in any discipline. He is 59 years old, works every night and weekend, and has published more than 1,000 academic papers since the turn of the millennium. In 2017 alone, he was a lead or coauthor of 85 publications[3]—which amounts to approximately one every four days.

[3] A 2018 article in Nature analyzed the phenomenon of "hyperprolific" scholars, which the authors defined as anyone with more than 72 publications in a year.

But if you ask him, this output isn't just the fruit of an ambitious work ethic; it's a mark of his tendency toward a kind of rigorous escapism.

Friston draws a carefully regulated boundary around his inner life, guarding against intrusions, many of which seem to consist of “worrying about other people.” He prefers being onstage, with other people at a comfortable distance, to being in private conversations. He does not have a mobile phone. He always wears navy-blue suits, which he buys two at a time at a closeout shop. He finds disruptions to his weekly routine on Queen Square “rather nerve-racking” and so tends to avoid other human beings at, say, international conferences. He does not enjoy advocating for his own ideas.

At the same time, Friston is exceptionally lucid and forthcoming about what drives him as a scholar. He finds it incredibly soothing—not unlike disappearing for a smoke—to lose himself in a difficult problem that takes weeks to resolve. And he has written eloquently about his own obsession, dating back to childhood, with finding ways to integrate, unify, and make simple the apparent noise of the world.

Friston traces his path to the free energy principle back to a hot summer day when he was 8 years old. He and his family were living in the walled English city of Chester, near Liverpool, and his mother had told him to go play in the garden. He turned over an old log and spotted several wood lice—small bugs with armadillo-shaped exoskeletons—moving about, he initially assumed, in a frantic search for shelter and darkness. After staring at them for half an hour, he deduced that they were not actually seeking the shade. “That was an illusion,” Friston says. “A fantasy that I brought to the table.”

He realized that the movement of the wood lice had no larger purpose, at least not in the sense that a human has a purpose when getting in a car to run an errand. The creatures’ movement was random; they simply moved faster in the warmth[4] of the sun.

[4] Young Friston was probably right. Many species of wood lice will dry out in direct sunlight, and some respond to a rise in temperature with kinesis, an increased level of random movement.

Friston calls this his first scientific insight, a moment when “all these contrived, anthropomorphized explanations of purpose and survival and the like all seemed to just peel away,” he says. “And the thing you were observing just *was*. In the sense that it could be no other way.”

Friston’s father was a civil engineer who worked on bridges all around England, and his family moved around with him. In just his first decade, the young Friston attended six different schools. His teachers often didn’t know what to do with him, and he drew most of his fragile self-esteem from solitary problem solving. At age 10 he designed a self-righting robot that could, in theory, traverse uneven ground while carrying a glass of water, using self-correcting feedback actuators and mercury levels. At school, a psychologist was brought in to ask him how he came up with it. “You’re very intelligent, Karl,” Friston’s mother reassured him, not for the last time. “Don’t let them tell you you’re not.” He says he didn’t believe her.

When Friston was in his mid-teens, he had another wood-lice moment. He had just come up to his bedroom from watching TV and noticed the cherry trees in bloom outside the window. He

suddenly became possessed by a thought that has never let go of him since. “There must be a way of understanding everything by starting from nothing,” he thought. “If I’m only allowed to start off with one point in the entire universe, can I derive everything else I need from that?” He stayed there on his bed for hours, making his first attempt. “I failed completely, obviously,” he says.

Toward the end of secondary school, Friston and his classmates were the subjects of an early experiment in computer-assisted advising. They were asked a series of questions, and their answers were punched into cards and run through a machine to extrapolate the perfect career choice. Friston had described how he enjoyed electronics design and being alone in nature, so the computer suggested he become a television antenna installer. That didn’t seem right, so he visited a school career counselor and said he’d like to study the brain in the context of mathematics and physics. The counselor told Friston he should become a psychiatrist, which meant, to Friston’s horror, that he had to study medicine.

Both Friston and the counselor had confused psychiatry with psychology, which is what he probably ought to have pursued as a future researcher. But it turned out to be a fortunate error, as it put Friston on a path toward studying both the mind and body,[5] and toward one of the most formative experiences of his life—one that got Friston out of his own head.

[5] Friston found time for other pursuits as well. At age 19, he spent an entire school vacation trying to squeeze all of physics on one page. He failed but did manage to fit all of quantum mechanics.

After completing his medical studies, Friston moved to Oxford and spent two years as a resident trainee at a Victorian-era hospital called Littlemore. Founded under the 1845 Lunacy Act, Littlemore had originally been instituted to help transfer all “pauper lunatics” from workhouses to hospitals. By the mid-1980s, when Friston arrived, it was one of the last of the old asylums on the outskirts of England’s cities.

Friston was assigned a group of 32 chronic schizophrenic patients, the worst-off residents of Littlemore, for whom treatment mostly meant containment. For Friston, who recalls his former patients with evident nostalgia, it was an introduction to the way that connections in the brain were easily broken. “It was a beautiful place to work,” he says. “This little community of intense and florid psychopathology.”

Twice a week he led 90-minute group therapy sessions in which the patients explored their ailments together, reminiscent of the Ask Karl meetings today. The group included colorful characters who still inspire Friston’s thinking more than 30 years later. There was Hillary,[6] who looked like she could play the senior cook on *Downton Abbey* but who, before coming to Littlemore, had decapitated her neighbor with a kitchen knife, convinced he had become an evil, human-sized crow.

[6] The names of Friston’s patients at Littlemore have been changed in this story.

There was Ernest, who had a penchant for pastel Marks & Spencer cardigans and matching plimsoll shoes, and who was “as rampant and incorrigible a pedophile as you could ever imagine,” Friston says.

And then there was Robert, an articulate young man who might have been a university student had he not suffered severe schizophrenia. Robert ruminated obsessively about, of all things, angel shit; he pondered whether the stuff was a blessing or a curse and whether it was ever visible to the eye, and he seemed perplexed that these questions had not occurred to others. To Friston, the very concept of angel shit was a miracle. It spoke to the ability of people with schizophrenia to assemble concepts that someone with a more regularly functioning brain couldn’t easily access. “It’s extremely difficult to come up with something like angel shit,” Friston says with something like admiration. “I couldn’t do it.”

After Littlemore, Friston spent much of the early 1990s using a relatively new technology—PET scans—to try to understand what was going on inside the brains of people with schizophrenia. He invented statistical parametric mapping along the way. Unusually for the time, Friston was adamant that the technique should be freely shared rather than patented and commercialized, which largely explains how it became so widespread. Friston would fly across the world—to the National Institutes of Health in Bethesda, Maryland, for example—to give it to other researchers. “It was me, literally, with a quarter of biometric tape, getting on an airplane, taking it over there, downloading it, spending a day getting it to work, teaching somebody how to use it, then going home for a rest,” Friston says. “This is how open source software worked in those days.”

Friston came to Queen Square in 1994, and for a few years his office at the FIL sat just a few doors down from the Gatsby Computational Neuroscience Unit. The Gatsby—where researchers study theories of perception and learning in both living and machine systems—was then run by its founder, the cognitive psychologist and computer scientist Geoffrey Hinton. While the FIL was establishing itself as one of the premier labs for neuroimaging, the Gatsby was becoming a training ground for neuroscientists interested in applying mathematical models to the nervous system.

Friston, like many others, became enthralled by Hinton’s “childlike enthusiasm” for the most unchildlike of statistical models, and the two men became friends.[7]

[7] At the time, Hinton was living in a particularly noisy building in Camden. The neighbors’ water pipes were so loud that he built a soundproof box in a basement bedroom out of rubber and ¾-inch drywall where he and his wife could sleep.

Over time, Hinton convinced Friston that the best way to think of the brain was as a Bayesian probability machine. The idea, which goes back to the 19th century and the work of Hermann von Helmholtz, is that brains compute and perceive in a probabilistic manner, constantly making predictions and adjusting beliefs based on what the senses contribute. According to the most popular modern Bayesian account, the brain is an “inference engine” that seeks to minimize “prediction error.”

In 2001, Hinton left London for the University of Toronto, where he became one of the most important figures in artificial intelligence, laying the groundwork[8] for much of today's research in deep learning.

[8] In 2012, Hinton won the ImageNet Challenge, a competition to identify objects in a 15-million-image database built by Fei-Fei Li. ImageNet helped propel neural networks—and Hinton—to the forefront of AI.

Before Hinton left, however, Friston visited his friend at the Gatsby one last time. Hinton described a new technique he'd devised to allow computer programs to emulate human decisionmaking more efficiently—a process for integrating the input of many different probabilistic models, now known in machine learning as a "product of experts."

The meeting left Friston's head spinning. Inspired by Hinton's ideas, and in a spirit of intellectual reciprocity, Friston sent Hinton a set of notes about an idea he had for connecting several seemingly "unrelated anatomical, physiological, and psychophysical attributes of the brain." Friston published those notes in 2005—the first of many dozens of papers he would go on to write about the free energy principle.

The Markov blanket in Karl Friston's office—"keeping your internal states warm since 1856."

*Kate Peters*

Even Friston has a hard time deciding where to start when he describes the free energy principle. He often sends people to its Wikipedia page. But for my part, it seems apt to begin with the blanket draped over the futon in Friston's office.

It's a white fleece throw, custom-printed with a black-and-white portrait of a stern, bearded Russian mathematician named Andrei Andreyevich Markov, who died in 1922. The blanket is a gag gift from Friston's son, a plush, polyester inside joke about an idea that has become central to the free energy principle. Markov is the eponym of a concept called a Markov blanket, which in machine learning is essentially a shield that separates one set of variables from others in a layered, hierarchical system. The psychologist Christopher Frith—who has an h-index on par with Friston's—once described a Markov blanket as "a cognitive version of a cell membrane, shielding states inside the blanket from states outside."

In Friston's mind, the universe is made up of Markov blankets inside of Markov blankets. Each of us has a Markov blanket that keeps us apart from what is not us. And within us are blankets separating organs, which contain blankets separating cells, which contain blankets separating their organelles. The blankets define how biological things exist over time and behave distinctly from one another. Without them, we're just hot gas dissipating into the ether.

"That's the Markov blanket you've read about. This is it. You can touch it," Friston said dryly when I first saw the throw in his office. I couldn't help myself; I did briefly reach out to feel it under my fingers. Ever since I first read about Markov blankets, I'd seen them everywhere. Markov blankets around a leaf and a tree and a mosquito. In London, I saw them around the postdocs at the FIL, around the black-clad protesters at an antifascist rally, and around the people living in boats in the canals. Invisible cloaks around everyone, and underneath each one a different living system that minimizes its own free energy.

The concept of free energy itself comes from physics, which means it's difficult to explain precisely without wading into mathematical formulas. In a sense that's what makes it powerful: It isn't a merely rhetorical concept. It's a measurable quantity that can be modeled, using much the same math that Friston has used to interpret brain images to such world--changing effect. But if you translate the concept from math into English, here's roughly what you get: Free energy is the difference between the states you expect to be in and the states your sensors tell you that you are in. Or, to put it another way, when you are minimizing free energy, you are minimizing *surprise*.

According to Friston, any biological system[9] that resists a tendency to disorder and dissolution will adhere to the free energy principle—whether it's a protozoan or a pro basketball team.

[9] In 2013, Friston ran a model that simulated a primordial soup full of floating molecules. He programmed it to obey both basic physics and the free energy principle. The model generated results that looked like organized life.

A single-celled organism has the same imperative to reduce surprise that a brain does.

The only difference is that, as self-organizing biological systems go, the human brain is

inordinately complex: It soaks in information from billions of sense receptors, and it needs to organize that information efficiently into an accurate model of the world. “It’s literally a fantastic organ in the sense that it generates hypotheses or fantasies that are appropriate for trying to explain these myriad patterns, this flux of sensory information that it is in receipt of,” Friston says. In seeking to predict what the next wave of sensations is going to tell it—and the next, and the next—the brain is constantly making inferences and updating its beliefs based on what the senses relay back, and trying to minimize prediction-error signals.

So far, as you might have noticed, this sounds a lot like the Bayesian idea of the brain as an “inference engine” that Hinton told Friston about in the 1990s. And indeed, Friston regards the Bayesian model as a foundation of the free energy principle (“free energy” is even a rough synonym for “prediction error”). But the limitation of the Bayesian model, for Friston, is that it only accounts for the interaction between beliefs and perceptions; it has nothing to say about the body or action. It can’t get you out of your chair.

This isn’t enough for Friston, who uses the term “active inference” to describe the way organisms minimize surprise while moving about the world. When the brain makes a prediction that isn’t immediately borne out by what the senses relay back, Friston believes, it can minimize free energy in one of two ways: It can revise its prediction—absorb the surprise, concede the error, update its model of the world—or it can *act* to make the prediction true. If I infer that I am touching my nose with my left index finger, but my proprioceptors tell me my arm is hanging at my side, I can minimize my brain’s raging prediction-error signals by raising that arm up and pressing a digit to the middle of my face.

And in fact, this is how the free energy principle accounts for *everything we do*: perception, action, planning, problem solving. When I get into the car to run an errand, I am minimizing free energy by confirming my hypothesis—my fantasy—through action.

For Friston, folding action and movement into the equation is immensely important. Even perception itself, he says, is “enslaved by action”: To gather information, the eye darts, the diaphragm draws air into the nose, the fingers generate friction against a surface. And all of this fine motor movement exists on a continuum with bigger plans, explorations,[10] and actions.

[10] Friston’s term for this kind of exploration is “epistemic foraging.” He is notorious among his colleagues for his coinages, known as Fristonese.

“We sample the world,” Friston writes, “to ensure our predictions become a self-fulfilling prophecy.”

So what happens when our prophecies are not self-fulfilling? What does it look like for a system to be overwhelmed by surprise? The free energy principle, it turns out, isn’t just a unified theory of action, perception, and planning; it’s also a theory of mental illness. When the brain assigns too little or too much weight to evidence pouring in from the senses, trouble occurs. Someone with schizophrenia, for example, may fail to update their model of the world to account for sensory input from the eyes. Where one person might see a friendly neighbor,

Hillary might see a giant, evil crow. “If you think about psychiatric conditions, and indeed most neurological conditions, they are just broken beliefs or false inference—hallucinations and delusions,” Friston says.

Over the past few years, Friston and a few other scientists have used the free energy principle to help explain anxiety, depression, and psychosis, along with certain symptoms of autism, Parkinson’s disease, and psychopathy. In many cases, scientists already know—thanks to Friston’s neuroimaging methods—which regions of the brain tend to malfunction in different disorders and which signals tend to be disrupted. But that alone isn’t enough to go on. “It’s not sufficient to understand which synapses, which brain connections, are working improperly,” he says. “You need to have a calculus that talks about beliefs.”

So: The free energy principle offers a unifying explanation for how the mind works and a unifying explanation for how the mind malfunctions. It stands to reason, then, that it might also put us on a path toward building a mind from scratch.

A few years ago, a team of British researchers decided to revisit the facts of King George III’s madness with a new analytic tool. They loaded some 500 letters written by the king into a machine-learning engine and laboriously trained the system to recognize various textual features: word repetition, sentence length, syntactical complexity, and the like. By the end of the training process, the system was able to predict whether a royal missive had been written during a period of mania or during a period of sanity.

This kind of pattern-matching technology—which is roughly similar to the techniques that have taught machines to recognize faces, images of cats, and speech patterns—has driven huge advances in computing over the past several years. But it requires a lot of up-front data and human supervision, and it can be brittle. Another approach to AI, called reinforcement learning, has shown incredible success at winning games: Go, chess, Atari’s *Breakout*. Reinforcement learning doesn’t require humans to label lots of training data; it just requires telling a neural network to seek a certain reward, often victory in a game. The neural network learns by playing the game over and over, optimizing for whatever moves might get it to the final screen, the way a dog might learn to perform certain tasks for a treat.

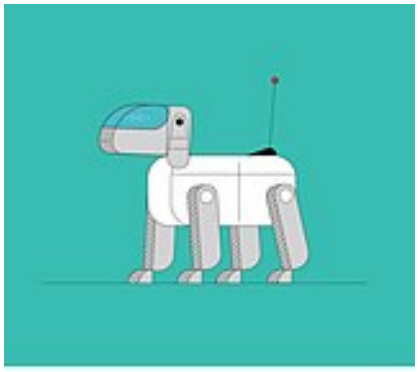
But reinforcement learning, too, has pretty major limitations. In the real world, most situations are not organized around a single, narrowly defined goal. (Sometimes you have to stop playing *Breakout* to go to the bathroom, put out a fire, or talk to your boss.) And most environments aren’t as stable and rule-bound as a game is. The conceit behind neural networks is that they are supposed to think the way we do; but reinforcement learning doesn’t really get us there.

To Friston and his enthusiasts, this failure makes complete sense. After all, according to the free energy principle, the fundamental drive of human thought isn’t to seek some arbitrary external reward. It’s to minimize prediction error. Clearly, neural networks ought to do the same. It helps that the Bayesian formulas behind the free energy principle—the ones that are so difficult to translate into English—are already written in the native language of machine learning.

Julie Pitt, head of machine-learning infrastructure at Netflix, discovered Friston and the free energy principle in 2014, and it transformed her thinking. (Pitt's Twitter bio reads, "I infer my own actions by way of Active Inference.") Outside of her work at Netflix, she's been exploring applications of the principle in a side project called Order of Magnitude Labs. Pitt says that the beauty of the free energy model is that it allows an artificial agent to act in any environment, even one that's new and unknown. Under the old reinforcement-learning model, you'd have to keep stipulating new rules and sub-rewards to get your agent to cope with a complex world. But a free energy agent always generates its own intrinsic reward: the minimization of surprise. And that reward, Pitt says, includes an imperative to go out and explore.

## LEARN MORE

### The WIRED Guide to Artificial Intelligence



In late 2017, a group led by Rosalyn Moran, a neuroscientist and engineer at King's College London, pitted two AI players against one another in a version of the 3D shooter game *Doom*. The goal was to compare an agent driven by active inference to one driven by reward-maximization.

The reward-based agent's goal was to kill a monster inside the game, but the free-energy-driven agent only had to minimize surprise. The Fristonian agent started off slowly. But eventually it started to behave as if it had a model of the game, seeming to realize, for instance, that when the agent moved left the monster tended to move to the right.

After a while it became clear that, even in the toy environment of the game, the reward--maximizing agent was "demonstrably less robust"; the free energy agent had learned its environment better. "It outperformed the reinforcement-learning agent because it was exploring," Moran says. In another simulation that pitted the free-energy-minimizing agent against real human players, the story was similar. The Fristonian agent started slowly, actively exploring options—epistemically foraging, Friston would say—before quickly attaining humanlike performance.

Moran told me that active inference is starting to spread into more mainstream deep-learning research, albeit slowly. Some of Friston's students have gone on to work at DeepMind and Google Brain, and one of them founded Huawei's Artificial Intelligence Theory lab. "It's moving out of Queen Square," Moran says. But it's still not nearly as common as reinforcement learning, which even undergraduates learn. "You don't teach undergraduates the free energy principle—yet."

The first time I asked Friston about the connection between the free energy principle and artificial intelligence, he predicted that within five to 10 years, most machine learning would incorporate free energy minimization. The second time, his response was droll. "Think about why it's called active inference," he said. His straight, sparkly white teeth showed through his

smile as he waited for me to follow his wordplay. “Well, it’s AI,” Friston said. “So is active inference the new AI? Yes, it’s the acronym.” Not for the first time, a Fristonian joke had passed me by.

While I was in London, Friston gave a talk at a quantitative trading firm. About 60 baby-faced stock traders were in attendance, rounding out the end of their workday. Friston described how the free energy principle could model curiosity in artificial agents. About 15 minutes in, he asked his listeners to raise a hand if they understood what he was saying. He counted only three hands, so he reversed the question: “Can you put your hand up if that was complete nonsense and you don’t know what I was talking about?” This time, a lot of people raised their hands, and I got the feeling that the rest were being polite. With 45 minutes left, Friston turned to the organizer of the talk and looked at him as if to say, *What the hell*? The manager stammered a bit before saying, “Everybody here’s smart.” Friston graciously agreed and finished his presentation.

The next morning, I asked Friston if he thought the talk went well, considering that few of those bright young minds seemed to understand him. “There is going to be a substantial proportion of the audience who—it’s just not for them,” he said. “Sometimes they get upset because they’ve heard that it’s important and they don’t understand it. They think they have to think it’s rubbish and they leave. You get used to that.”

In 2010, Peter Freed, a psychiatrist at Columbia University, gathered together 15 brain researchers to discuss one of Friston’s papers. Freed described what happened in the journal *Neuropsychoanalysis*: “There was a lot of mathematical knowledge in the room: three statisticians, two physicists, a physical chemist, a nuclear physicist, and a large group of neuroimagers—but apparently we didn’t have what it took. I met with a Princeton physicist, a Stanford neurophysiologist, a Cold Springs Harbor neurobiologist to discuss the paper. Again blanks, one and all: too many equations, too many assumptions, too many moving parts, too global a theory, no opportunity for questions—and so people gave up.”

## Related Stories

But for all the people who are exasperated by Friston’s impenetrability, there are nearly as many who feel he has unlocked something huge, an idea every bit as expansive as Darwin’s theory of natural selection. When the Canadian philosopher Maxwell Ramstead first read Friston’s work in 2014, he had already been trying to find ways to connect complex living systems that exist at different scales—from cells to brains to individuals to cultures. In 2016 he met Friston, who told him that the same math that applies to cellular differentiation—the process by which generic cells become more specialized—can also be applied to cultural dynamics. “This was a life-changing conversation for me,” Ramstead says. “I almost had a nosebleed.”

“This is absolutely novel in history,” Ramstead told me as we sat on a bench in Queen Square, surrounded by patients and staff from the surrounding hospitals. Before Friston came along, “We were kind of condemned to forever wander in this multidisciplinary space without a

common currency," he continued. "The free energy principle gives you that currency."

In 2017, Ramstead and Friston coauthored a paper, with Paul Badcock of the University of Melbourne, in which they described all life in terms of Markov blankets. Just as a cell is a Markov-blanketed system that minimizes free energy in order to exist, so are tribes and religions and species.

After the publication of Ramstead's paper, Micah Allen, a cognitive neuroscientist then at the FIL, wrote that the free energy principle had evolved into a real-life version of Isaac Asimov's psychohistory,[11] a fictional system that reduced all of psychology, history, and physics down to a statistical science.

[11] In *Foundation*, published in 1951, one of Asimov's characters defines psychohistory as "that branch of mathematics which deals with the reactions of human conglomerates to fixed social and economic stimuli."

And it's true that the free energy principle does seem to have expanded to the point of being, if not a theory of everything, then nearly so. (Friston told me that cancer and tumors might be instances of false inference, when cells become deluded.) As Allen asked: Does a theory that explains everything run the risk of explaining nothing?

On the last day of my trip, I visited Friston in the town of Rickmansworth, where he lives in a house filled with taxidermied animals[12] that his wife prepares as a hobby.

[12] On a recent Saturday, a man came to the door asking if Friston's wife was home. When Friston said yes, the man said, "Good, because I got a dead cat here." He wanted it stuffed.

As it happens, Rickmansworth appears on the first page of *The Hitchhiker's Guide to the Galaxy*; it's the town where "a girl sitting on her own in a small café" suddenly discovers the secret to making the world "a good and happy place." But fate intervenes. "Before she could get to a phone to tell anyone about it, a terrible stupid catastrophe occurred, and the idea was lost forever."

It's unclear whether the free energy principle is the secret to making the world a good and happy place, as some of its believers almost seem to think it might be. Friston himself tended to take a more measured tone as our talks went on, suggesting only that active inference and its corollaries were quite promising. Several times he conceded that he might just be "talking rubbish." During the last group meeting I attended at the FIL, he told those in attendance that the free energy principle is an "as if" concept—it does not require that biological things minimize free energy in order to exist; it is merely sufficient as an explanation for biotic self-organization.

Friston's mother died a few years ago, but lately he has been thinking back to her frequent reassurances during his childhood: *You're very intelligent, Karl*. "I never quite believed her," he says. "And yet now I have found myself suddenly being seduced by her argument. Now I do

believe I'm actually quite bright." But this newfound self-esteem, he says, has also led him to examine his own egocentricity.

Friston says his work has two primary motivations. Sure, it would be nice to see the free energy principle lead to true artificial consciousness someday, he says, but that's not one of his top priorities. Rather, his first big desire is to advance schizophrenia research, to help repair the brains of patients like the ones he knew at the old asylum. And his second main motivation, he says, is "much more selfish." It goes back to that evening in his bedroom, as a teenager, looking at the cherry blossoms, wondering, "*Can I sort it all out in the simplest way possible? "*

"And that is a very self-indulgent thing. It has no altruistic clinical compassion behind it. It is just the selfish desire to try and understand things as completely and as rigorously and as simply as possible," he says. "I often reflect on the jokes that people make about me—sometimes maliciously, sometimes very amusingly—that I can't communicate. And I think: I didn't write it for you. I wrote it for me."

Friston told me he occasionally misses the last train home to Rickmansworth, lost in one of those problems that he drills into for weeks. So he'll sleep in his office, curled on the futon under his Markov blanket, safe and securely separated from the external world.

---

**Shaun Raviv** *(@ShaunRaviv) is a writer living in Atlanta, Georgia.*

*This article appears in the December issue. Subscribe now.*

*Listen to this story, and other WIRED features, on the Audm app.*

*Let us know what you think about this article. Submit a letter to the editor at mail@wired.com.*

---

## More Great WIRED Stories