The Free Energy Principle – Explained with Marbles!

Welcome, everyone, to my presentation of the Bayesian Brain and the Free Energy Principle! This is gonna be a bit of a journey, because rather than outright say what this theory is and then try to explain it, I'd like to try to take you gradually through the conceptual steps that lead to it.

I want to start with what it tries to achieve. In a nutshell, the goal is a framework that answers an enormous question: what makes life and what makes the mind? What should they be understood to be for? To tackle this mission, this theory attempts to unify concepts from a range of disciplines, all in order to formulate a principle for self-organizing systems that process information, generally ; it doesn't matter whether these systems are brains, cells, AI or even corporations.

If you look this up, you're gonna see the theoretical framework we're discussing today connected to the name Karl Friston, an extremely prolific researcher at University College London. Because he, a neuroscientist, is at the center of this conversation, research in this direction has been shaped heavily by his interests in psychiatric disorders. So when people talk about this, they mostly talk about the brain; but the overarching ambition here is larger than that.

To begin building this theory, let's start simple.

I want you to imagine a box with a divider in the middle, that contains four marbles. The divider has some holes in it, meaning that it's possible for marbles to bounce from one compartment to the other. Imagine you take the box, you shake it, and you look inside. What you're most likely to see is that there are two marbles in each compartment, or perhaps one in a compartment and three in the other. The reason for this has to do with some simple probabilities. There is one possible configuration where all marbles are on the left, four configurations where 3 marbles are on the left, six configurations where 2 marbles are on the left, four configurations where there is 1 marble, and a single configuration where there are none. This with a total of sixteen possible configurations.

That means that there is a roughly 38% probability that they'll be distributed 2 and 2, a 50% probability that they'll be distributed 1 and 3, and a 12% probability that they'll be distributed 4 and 0.

The reason I'm telling you about these marbles is that they neatly demonstrate the first step on the way to the Free Energy Principle, namely thermodynamic entropy. Entropy is known as this slippery measure of disorder from physics, and it is the core concept of the second law of thermodynamics: the entropy of the universe at large is always increasing. Ink diffuses into water, living things die and decay, everything falls into gradual disorder. The marbles disperse around the box equally.

As we've just seen, this can be understood as a matter of probabilities: over a sufficient time scale, left on their own, all things tend towards the equilibrium state, the most likely state. If the box starts out in the improbable state A, it will tend to move towards the center of the curve. That's the second law: tendency towards the most probable.

Entropy is not inevitable, though. What is necessary to fight it is to expend energy. For example, if you use your finger to push the marbles to one side of the box every time, you expend energy to maintain low entropy, to push the box back towards the long tail of the probability curve, back towards order.

The insight has been around since the 50s that what living things do is harvest energy from the environment to maintain themselves in an ordered, low-entropy state. To push all their marbles in a very specific place, and no other. They have a little wiggle room, but if their entropy increases too much, they die.

Evolutionary mechanisms select for the organisms best at maintaining low entropy in a given environment; that means that the traits of the organism reflect the environment in a certain way. Otherwise put, the organism can be seen as a model of its own environment.

There's another way to look at this. Disordered systems contain little information. Ordered systems, however, encode information. TV static doesn't tell us much, but an ordered image conveys words. A recording of a hundred people talking over each other at equal volume tells us nothing, but a recording of 99 people talking quietly and 1 person talking loudly conveys information. Ants scattered all over the garden doesn't convey much, but ants walking a purposeful line between the anthill and a bush tells the location of and route to food.

Energy expenditure is necessary to maintain the order that encodes information. Like in a computer's RAM, where continuous input of electricity is necessary to hold the memory in place.

So if we could "read" an amoeba, which is also highly ordered, what information would *it* convey? It is proposed that *it* is a predictive model of its environment. This makes sense from an evolutionary perspective, because an organism that is good at predicting its environment would be good at adapting and perpetuating itself.

What does that look like? Let's suppose we have two amoebae whose environment is our box of marbles from earlier. Let's say that the red amoeba's predictive model of the box is such that it predicts the three main outcomes of a shake (0-4, 1-3 or 2-2 arrangement) to have equal likelihood; let's say the blue amoeba predicts that 0-4 has 20% odds of occurring, 1-3 has 50% odds and 2-2 has 30% odds. Everytime the box is shaken, they make a prediction according to their model (e.g. "the box will be in a

0-4 configuration this time"), and if the prediction turns out right, they get to eat. So... which amoeba would survive longest? Well, when it's put like this, it seems pretty obvious it's the blue one.

What we see here is that an organism will prevail longer if its model predictions match more closely the real probability distribution of outcomes in its environment. This means that each organism has inside it, metaphorically, its own box of marbles to represent the outside one. Physically, this means some structure that matches the probability distribution of the states of the environment, *a representation*. So the organism prevails longer, the closer its inner box of marbles matches the real one.

Information theory defines a quantity that is helpful here: surprise. It's defined as the negative logarithm of the probability assigned to an event. You can view it as a score that says how surprised a system is when an event occurs, given that the system predicted the event with probability p. If p is large, then surprise when it occurs is nearly zero. If p is small, then the surprise is large.

Now, what a living system will want to do to perfect its predictions is minimize average surprise; not the surprise when the box is shaken once, but the average surprise over many shakes. Information theory has a name for this too: average surprise is called informational entropy. This is a distinct concept from thermodynamic entropy, but it is related. Thermodynamic entropy regards the order of physical systems ; informational entropy regards prediction accuracy.

So, the living system will want, in fact, to both keep its thermodynamic entropy low, and minimize its informational entropy.

Now comes the part where I tell you that I lied about one thing: the idea isn't exactly that the organism predicts what the environment will be like, at least not mainly; rather it predicts what the environment *is* like.

Let me explain. An information processing system doesn't have direct access to its outside world. Single-cell organisms are walled in by membranes; brains have a skull in the way. What they have at their disposal is a small number of sensors and effectors that relay to the system information about the environment and act on it. This is what they have to work with to minimize average surprise.

This also means that the probabilistic model they have about the outside world is built as a function of perceptions. Not the probability of an event p(e), but the probability of an event e coinciding with sensory input s: p(s, e). As an example: if I hear a loud bang in the house (the sensation), I will infer that most probably the event was that a window slammed because of strong winds, and I'll assign much lower probabilities to possibilities such as somebody having fired a gun.

This means that the equation for surprise from earlier gets rewritten as the negative logarithm of the probability of s coinciding with e. Bayesian probability says that that equals to the probability of sensation s occurring as a result of causing event e, times the probability of e occurring.

Which leads us to the problem. If an organism must minimize average surprise, it must evaluate it. To evaluate it, according to this equation, it must evaluate the probability of sensation conditional on event, which can be an artificial inner model; and it must evaluate the real, objective probability of event e in the world. It doesn't have access to that information, it's not God. So it can't evaluate surprise directly.

This leads us, finally, to the Free Energy Principle. Information theory shows, via some rather complicated mathematical artifice, that one can define an abstract quantity called Free Energy that represents an upper bound on surprise. Otherwise put, the surprise regarding an event cannot be larger than the free energy regarding that event. This matters, because the way it is defined free energy is a function of two parameters internal to the system: its own sensation recognition and its prediction-generating model of the world. So, the free energy, unlike surprise, can actually be evaluated. Since surprise cannot be larger than free energy, what you can try in order to make sure you minimize surprise, is to minimize free energy. So can be formulated, as a proxy to minimizing surprise, the Free Energy Principle: any adaptive change of a self-organizing system that processes information will minimize free-energy. This potentially applies to everything from single cell organisms to the brain to ecosystems. The adaptive changes we're talking about can be anywhere from evolutionary time to fractions of a second.

To summarize, such a self-organizing system aims to be surprised as rarely as possible, because surprise is bad for its survival ; it cannot evaluate surprise in its probabilistic model of the world directly; so, it evaluates a quantity that is always larger than surprise, namely free energy, and as a way to minimize surprise, it pushes down free energy as much as possible.

That's the Principle at its most abstract. Here, the brain is simply another step in life's journey to become better and better at modelling its own environment; it's seen as a complex arrangement of tissue whose goal is to accurately predict the world.

Now, to make it a bit clearer, I want to focus on the brain and give a series of examples.

In this framework, the brain is seen as an entity that produces inferences about the world, or what we can call beliefs. These are presented mathematically as probability distributions. So for instance the probability distribution for what is above my head when I wake up in the morning will say that most likely it's the ceiling of my room. The model, however, also leaves a small probability that what is actually above my head is a spider lowering itself towards my face on a thread.

The brain then takes perceptual input from my eyes and uses it to calculate the error of my belief. If the error's 0, it's all good. If it computes a large error, that is, a conflict between my belief and perceptual input, then it can take one of two actions to diminish this error, to minimize surprise. The first it can do is update my belief, to say that in fact, most likely above my head is a spider. The second thing it can do instead of modifying the model, is to act on the world to make it in line with my belief. In this case, I jerk my head to the left so that the spider is no longer directly above me.

To reiterate, when confronted with a conflict between belief and sensory input, the system can either change the belief, or act on the world to make it closer to the belief, or act on the world to look for evidence for the belief.

At the highest level of abstraction, what the brain does, what any organism does, is gather evidence for its own existence, for its own adequacy, confirmation of itself as a model. If in itself it is an inadequate model, it cannot maintain its low entropy structure, it breaks down. The marbles scatter.

And through this lens, sentience is, at least in one interpretation, a function of temporal depth of the system that generates beliefs. In translation, that means that sentience isn't something binary. You don't either have it, or not have it. Instead, the higher a system's capacity to ask "what if?" questions about the future and about the past, the more sentient.

Right. So why does this matter, because while sometimes it feels like a deep insight, the FEP sometimes seems like stating the obvious in a complicated way. The crucial reason it matters is that this theory makes mathematically operationalizable models and predictions regarding brain function, brain circuitry and regarding the origins of life. In neuroscience, in medicine, in AI, in evolutionary biology, that is a big deal. But right now it is early days for this theory; experimental evidence is trickling in only slowly, because the math involved is very hard, because of the interdisciplinarity needed, and because of the sheer novelty of it all.

The best case scenario is that the theory and the experimental side mature enough to give us some of those deep insights. To understand how the marbles come together in ordered, thermodynamically improbable configurations to represent information about their environment such that they self-update and self-maintain. To get us closer to understanding life, to build an artificial mind, and to tell us why some people lose their marbles.