A Gentle Introduction to the Free Energy Principle

Arthur Juliani · Follow

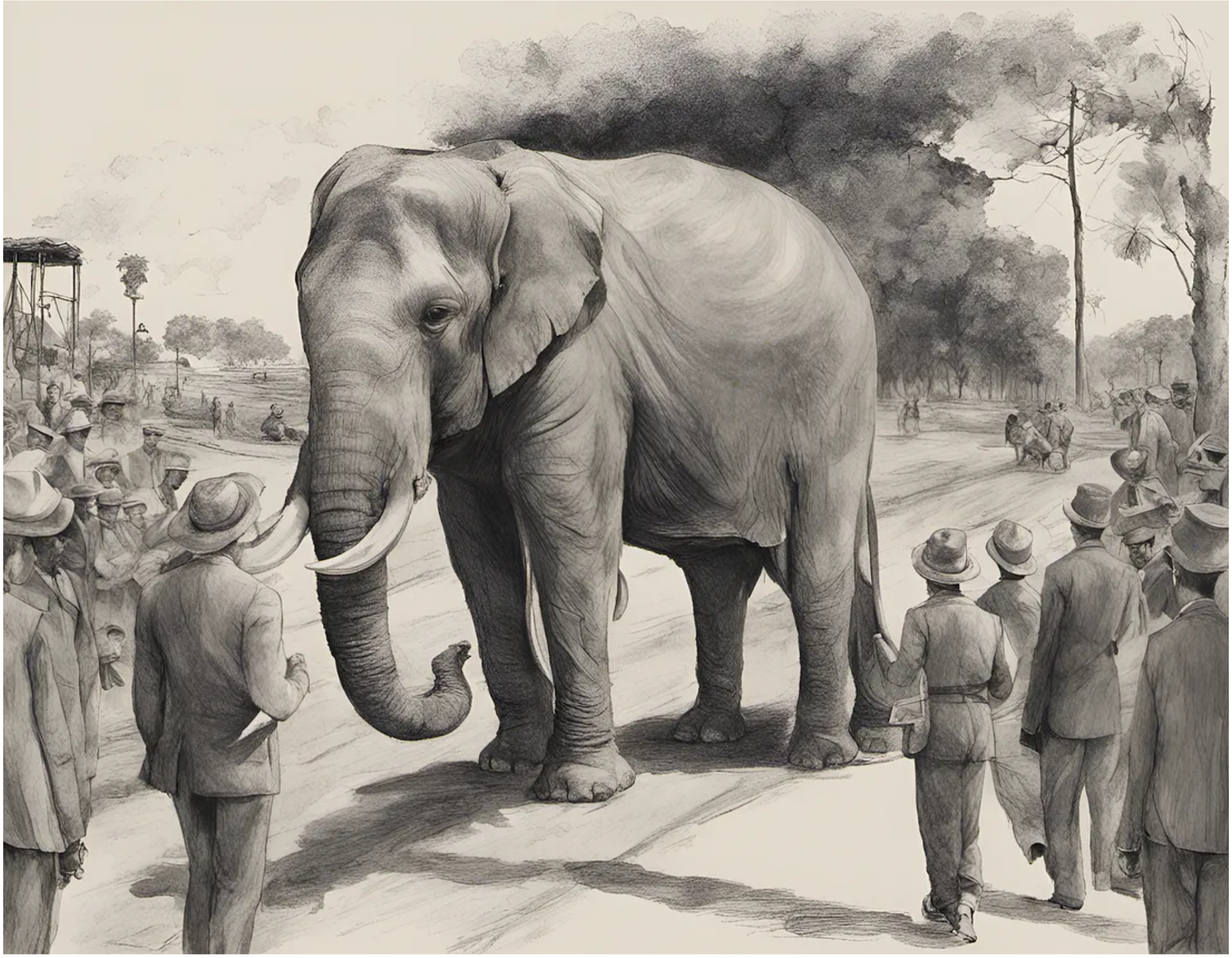
9 min read · Oct 4, 2023

Listen Share

What does it mean to be alive? The Free Energy Principle (FEP) is one of the more elegant theories to attempt an answer to this question. Developed by Karl Friston and others from the mid-2000s onward, the FEP builds on and integrates a number of key ideas from cybernetics, predictive coding, and bayesian inference into a single unified theory. At a basic level, FEP is a theory about what enables living organisms to stay alive. It is simple enough to account for single-celled bacteria and powerful enough to account for humans in all our complexity. In this article, I provide an overview of the theory (and its extension as Active Inference) from a completely non-technical perspective. If you are interested in a more mathematically rigorous treatment of the FEP, here are some useful resources (if you want a critique of the theory, there are those too).

The FEP states that living organisms are dynamical systems that are separated from their environment by an interface. We will refer to these living organisms as 'agents' for the sake of convenience going forward, since all living organisms sense their environment and act in response to those sensations. This interface, referred to as a 'markov blanket,' whether it's a bacteria's cell membrane, the bark of a tree, or the skin of a mammal, separates the agent from its environment. Crucially, this interface separates both the environment from the agent and the agent from its surroundings, allowing it to exist as a self-contained and autonomous entity. Agents therefore always necessarily receive sensory information and act through this interface. They never have unmediated access to what is going on in the environment, and as a result always have to guess (or infer) what is the true state of the world based on the information they have access to.

According to the FEP, all agents act, as best as they can, to minimize the errors in their guesses about the environment. These errors can be quantified by a metric called 'free energy,' and the better an agent can minimize it over time, the better chance the agent has of staying alive and maintaining itself. To understand it intuitively, think of free energy as a measure of surprise or uncertainty. If an organism's guess about its environment is very off-mark, that results in high free energy, which signifies high surprise or unpredictability. On the other hand, when the organism's guess aligns closely with its environment, the free energy is low, indicating that things are as expected. Organisms, whether they're simple bacteria or complex beings like humans, always aim to minimize this surprise, ensuring they're not caught off-guard by their surroundings.



We can take the parable of the blind men and the elephant as an example. None had direct access to the elephant itself, and each had to make a best-guess based on their limited physical contact with the animal. The man who touched the tail and guessed that the animal is a broom would be able to locally minimize prediction errors while touching the tail, but would be faced with new prediction errors when touching the leg of the elephant. Likewise for the man who touched the trunk of the elephant and inferred he was dealing with a water hose. Inferring that the object is an elephant on the other hand is useful not only because it is 'correct' in some ontological sense, but it also ensures that when faced with future sensory evidence there won't be an increase in free energy.

As mentioned above, the minimization of free energy is not an end in itself. It is ultimately in service of the maintenance of the agent as an individual separated from its environment. Agents do this by attempting to remain in a state of homeostasis. All living agents have some sense of what that homeostasis for them should be, even if it is only represented implicitly within the structure of the agent itself. In the same way that sensory prediction errors increase free energy, so do deviations from the desired homeostasis of the agent, as the homeostatic set points of an agent itself are another kind of prediction about the world.

For a simple example of this homeostatic maintenance, we can consider the 'behavior' of a tree. It is obvious that trees do not possess either complex mental representations nor a complex behavior repertoire. Still, there is some agency which a tree is capable of exerting. We find that certain species of trees which live in windy regions are capable of growing in the direction of the wind. Here we have all the components of the FEP at work. The tree has a desired homeostasis that came about as the result of natural selection: to remain upright and firmly rooted. This expectation is violated when the wind blows against the tree and its branches, thus increasing the free energy of the system. In response, the tree ‘acts’ in order to minimize the free energy by slowly growing in the direction of the wind, thus reducing the strain put on the tree and bettering the chances of its survival.

This simple principle can be applied to agents much more complex than trees. Consider for example a lizard maintaining its body temperature by finding a sunny spot to rest in. Here we again have an expectation (the body temperature), the violation (shaded area reducing temperature), and the action to resolve it (moving to the sunny area). Of course, this raises the important question of how the agent knows where to find the sunny area. While it is straightforward for a tree to sense the direction of the wind, it is a more complex task for a lizard to find the sun. To understand this behavior, we need to expand the space of possible representations the agent can possess from internal ones of homeostasis to external beliefs about the state of the environment.

According to the FEP, all perception, cognition, affect, and behavior can all be understood through the lens of free energy minimization. This means that an agent’s visual system is engaged in the task of forming beliefs about what it will see, and then updating those beliefs when the expectations are violated. In the example of the lizard, it has learned to distinguish regions of its environment which are covered in light from the sun from those which aren’t through this process of free energy minimization. The same principle which applied to the men and the elephant and the tree and the wind applies here to the lizard and the sun, and indeed can be applied to all other animals as well.

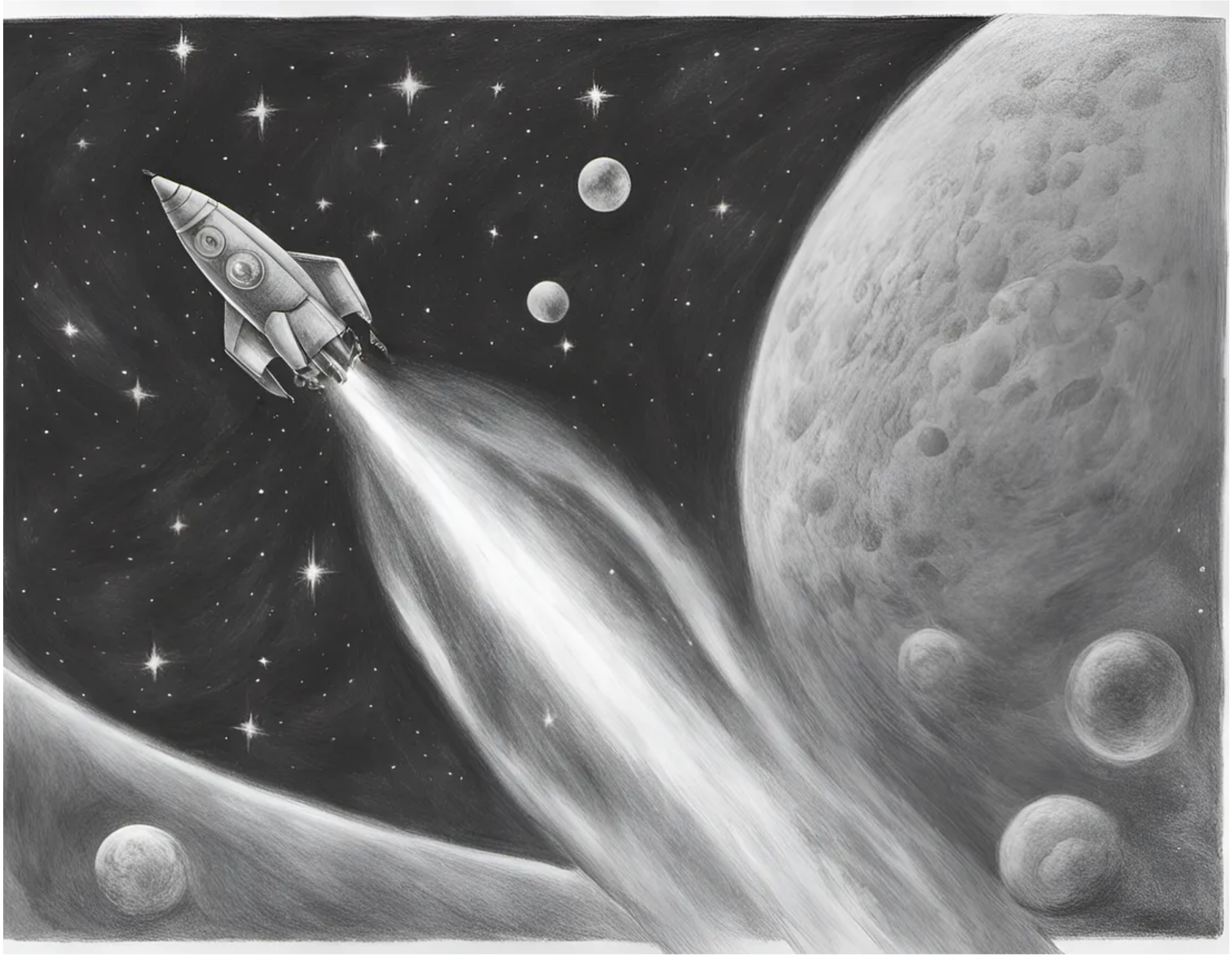
In making predictions about the world, agents form beliefs about the world. These beliefs are the guesses that we make, and not all guesses are made equal. We can take as an example the process of making sense of the world while walking around at night. Perhaps you are alone at night and suddenly see a person off in the distance. This event will send a cascade of prediction errors through the brain until at some point the discrepancy is resolved by a high level belief determining that what is being seen is a human. Once the high-level belief about the object's identity has been updated to match the sensory information, there are no longer prediction errors, and the free energy has been minimized. As you get closer, you might discover that rather than a human, it was in fact a shrub. Here again there is a cascade of prediction errors that eventually update the high-level belief about the nature of the object.

What accounts for the threshold at which we switch our belief about the world from 'seeing a human' to 'seeing a shrub'? The FEP introduces the concept of the precision of beliefs, which describes how strongly encoded a given belief is. A low-precision belief can be more easily revised in the face of incoming prediction errors. In contrast, a high-precision belief is more likely to suppress those errors. In the case of the shrub at night, the precision of the belief in a human was relatively low due to the poor light. At the time it seemed like it could be a human, but you weren't certain, and that belief was malleable. In contrast, seeing a person in sunlight during the day is likely to result in a much higher precision belief about the existence of that person. It would take much stronger sensory evidence to change such a belief.

An interesting aspect of the FEP is that it can often produce behavior which paradoxically doesn't seem like error minimization, at least in the short-term. Let's imagine that someone moves to a new neighborhood they are unfamiliar with. They could choose to stay inside their new home, which would indeed allow them to minimize any prediction errors that would arise from the novel environment. On the other hand, they could go outside and explore their new surroundings. Doing the latter will produce many more short-term prediction errors. In the long run though it will allow them to avoid even bigger surprises in the future. It is this dynamic of long-term free energy minimization that can account for various forms of exploration, adventure, and industry which we are motivated to engage in. Not just us, but many other animals also engage in this kind of proactive free energy minimization. Anyone who has lived with a cat can attest to both their unique spatial curiosity as well as their general tendency to be easily frightened. This apparent paradox is resolved if we understand their behavior through the lens of free energy minimization. They do not simply oscillate between curiosity and fear, the former is rather in service of preventing the latter!

We humans distinguish ourselves from other living agents by our ability to not only react to the environment but also establish goals and then set out to accomplish them. This might seem to be a challenge for a theory which suggests that all behavior is in the service of entropy minimization, but the FEP accounts for this quite elegantly as well. In this view, when we set a goal for ourselves, we are making a special kind of prediction about the world. This prediction is one that we know goes against the current sensory evidence, thus increasing free energy. Instead of updating our beliefs to account for the world (and in essence give up on the goal), what we do instead is to set out to act in the world in order to ensure that the sensory evidence corresponds to our prediction, and minimize the free energy in the process.

A simple example of setting predictions and making them come true is the act of picking up a glass of water and drinking it. If someone is thirsty, they set the goal of drinking water for themselves. This violates the current sensory evidence, since they are not in fact drinking water at that exact moment. Thankfully, in most circumstances they are able to pick up the glass of water and resolve the discrepancy by bringing it to their mouth. The same basic logic applies to much more complex goals, such as traveling to a destination across town, or marrying, or getting a college degree. It even applies to loftier goals which involve bringing into the world things which have never existed before, such as rockets into space or any other technology which we humans have invented. In each case, we make a prediction about how we would like the world to be, and then act in order to make that prediction come true.



The Free Energy Principle offers an elegant and expansive explanation for the behavior of living organisms. From trees trying to stay upright to humans setting intricate life goals, FEP suggests that all actions are fundamentally driven by a desire to reduce uncertainty and maintain balance. Agents continuously endeavor to predict and understand their surroundings, and to minimize discrepancies between beliefs and sensory evidence. As I mentioned in the introduction, the FEP isn't universally accepted, and has its detractors. Still, I think it provides one of the more intuitive and compelling explanations for why we, and all living beings, do what we do. If you found this article interesting and would like to learn more about the FEP, here are some useful resources (and if you'd like to read some technical critiques of the theory, there are those too).

Psychology Life Neuroscience Machine Learning Philosophy

Follow

Written by Arthur Juliani

13.9K Followers · 57 Following

Interested in artificial intelligence, neuroscience, philosophy, psychedelics, and meditation. http://arthurjuliani.com/

Responses (21)

What are your thoughts? Respond