DOCUMENT SUMMARY

This document is a highly influential review article by Karl Friston (2010) introducing the **free-energy principle** as a potential unified theory of the brain. The principle posits that any self-organizing system, including the brain, must minimize its free energy to resist a natural tendency toward disorder. This minimization is achieved through two primary mechanisms: **perception** (updating internal models of the world to better predict sensory input) and **action** (actively sampling the world to make sensory input match predictions). Friston argues that this single imperative—to minimize surprise (or prediction error)—can unify various major brain theories, including the **Bayesian brain hypothesis**, **predictive coding**, **neural Darwinism**, and **optimal control theory**.

FILENAME

Friston_2010_research_review_article_free_energy_principle

METADATA

- **Primary Category**: RESEARCH
- **Document Type**: review_article
- **Relevance**: Core
- **Update Frequency**: Static
- **Tags**: #free-energy-principle, #bayesian-brain, #predictive-coding, #unified-brain-theory, #neuroscience, #action-perception-loop, #karl-friston
- **Related Docs**: This paper provides a high-level theoretical framework that can integrate the findings from other papers like "The Neuroscience of Autism" and "Jensen_1997_research_article_adhd_disorder_of_adaptation."

FORMATTED CONTENT

The free-energy principle: a unified brain theory?

Karl Friston

Abstract

> A **free-energy principle** has been proposed recently that accounts for action, perception and learning. This Review looks at some key brain theories in the biological (for example, neural Darwinism) and physical (for example, information theory and optimal control theory) sciences from the free-energy perspective. Crucially, one key theme runs through each of these theories — optimization. Furthermore, if we look closely at what is optimized, the same quantity keeps emerging, namely value (expected reward, expected utility) or its complement, surprise (prediction error, expected cost). This is the quantity that is optimized under the **free-energy principle**, which suggests that several global brain theories might be unified within a free-energy framework.

The free-energy principle

The **free-energy principle** says that any self-organizing system that is at equilibrium with its environment must minimize its **free energy**. The principle is essentially a mathematical formulation of how adaptive systems (that is, biological agents, like animals or brains) resist a natural tendency to disorder.

Motivation: resisting a tendency to disorder. The defining characteristic of biological systems is that they maintain their states and form in the face of a constantly changing environment. This maintenance of order is seen at many levels and distinguishes biological from other self-organizing systems; indeed, the physiology of biological systems can be reduced almost entirely to their **homeostasis**. Mathematically, this means that the probability of these sensory states must have low **entropy**; in other words, there is a high probability that a system will be in any of a small number of states. Entropy is also the average self information or '**surprise**'. Biological agents must therefore minimize the long-term average of surprise to ensure that their sensory entropy remains low.

So how do they do this? A system cannot know whether its sensations are surprising and could not avoid them even if it did know. This is where free energy comes in: **free energy** is an upper bound on surprise, which means that if agents minimize free energy, they implicitly minimize surprise.

Implications: action and perception. Agents can suppress free energy by changing the two things it depends on: they can change sensory input by **acting** on the world or they can change their **recognition density** (a probabilistic representation of what caused a sensation) by changing their internal states. This distinction maps nicely onto **action** and **perception**.

In short, the agent will selectively sample the sensory inputs that it expects. This is known as **active inference**.

The Bayesian brain hypothesis

The **Bayesian brain hypothesis** uses Bayesian probability theory to formulate perception as a constructive process based on internal or **generative models**. The underlying idea is that the brain has a model of the world that it tries to optimize using sensory inputs. In this view, the brain is an inference machine that actively predicts and explains its sensations.

The free-energy formulation was developed to finesse the difficult problem of exact inference by converting it into an easier optimization problem. Minimizing the free energy effectively optimizes **empirical priors** (that is, the probability of causes at one level, given those in the level above).

Minimizing free energy then corresponds to explaining away **prediction errors**. This is known as **predictive coding** and has become a popular framework for understanding neuronal message passing among different levels of cortical hierarchies. In this scheme, prediction error units compare conditional expectations with top-down predictions to elaborate a prediction error. This prediction error is passed forward to drive the units in the level above... which optimize top-down predictions to explain away (reduce) prediction error in the level below.

The principle of efficient coding

The principle of efficient coding suggests that the brain optimizes the mutual information (that is, the mutual predictability) between the sensorium and its internal representation. At its simplest, the **infomax principle** says that neuronal activity should encode sensory information in an efficient and parsimonious fashion.

This principle becomes a special case of the **free-energy principle**. The infomax principle can be understood in terms of the decomposition of free energy into complexity and accuracy: mutual information is optimized when conditional expectations maximize accuracy (or minimize prediction error), and efficiency is assured by minimizing complexity.

The cell assembly and correlation theory

The **cell assembly theory** was proposed by Hebb and entails **Hebbian** or **associative plasticity**... "cells that fire together wire together." This enables the brain to distil statistical regularities from the sensorium.

A gradient descent on free energy (that is, changing connections to reduce free energy) is formally identical to **Hebbian plasticity**. This is because the parameters of the generative model determine how expected states (synaptic activity) are mixed to form predictions. Put simply, when the presynaptic predictions and postsynaptic prediction errors are highly correlated, the connection strength increases, so that predictions can suppress prediction errors more efficiently.

Neural Darwinism and value learning

In the **theory of neuronal group selection (Neural Darwinism)**, the emergence of neuronal assemblies is considered in the light of selective pressure. Plasticity rests on correlated pre- and postsynaptic activity, but here it is modulated by **value**. Value is signalled by ascending neuromodulatory transmitter systems and controls which neuronal groups are selected and which are not.

The beauty of **neural Darwinism** is that it nests distinct selective processes within each other. The answer is simple: **value is inversely proportional to surprise**, in the sense that the probability of a phenotype being in a particular state increases with the value of that state. This means that free energy is the complement of value, and its long-term average is the complement of adaptive fitness. Put simply, valuable states are just the states that the agent expects to frequent.

Optimal control theory and game theory

Value is central to theories of brain function that are based on **reinforcement learning** and **optimum control**. The basic notion that underpins these treatments is that the brain optimizes value, which is expected reward or utility (or its complement - expected loss or cost). This is formalized in **optimal control theory** as the **Bellman equation**.

If one assumes that the optimal policy performs a gradient ascent on value, then it is easy to show that value is inversely proportional to surprise. This means that free energy is (an upper bound on) expected cost, which makes sense as optimal control theory assumes that action minimizes expected cost, whereas the **free-energy principle** states that it minimizes free energy.

Conclusions and future directions

> Although contrived to highlight commonalities, this Review suggests that many global theories of brain function can be united under a Helmholtzian perceptive of the brain as a **generative model** of the world it inhabits; notable examples include the integration of the **Bayesian brain** and computational motor control theory, the objective functions shared by **predictive coding** and the **infomax principle**, hierarchical inference and theories of attention, the embedding of perception in **natural selection** and the link between **optimum control** and more exotic phenomena in dynamical systems theory.

The constant theme in all these theories is that the brain optimizes a (free-energy) bound on surprise or its complement, value. This manifests as perception (so as to change predictions) or action (so as to change the sensations that are predicted).