The Theory of Active Inference, the Free-Energy Principle and Affective Phenomena

By

Damien Bérubé

A thesis submitted to the Faculty of Arts and Social Sciences

in partial fulfilment of the requirements for the degree of

Bachelor of Cognitive Science with Honours

Concentration in Cognition and Computation

Thesis Supervisor: Dr. Mary Kelly

Department of Cognitive Science
Carleton University

© 2024

Damien Bérubé

Contents

Abstract **5**

Introduction **6**

1 A Conceptual Guide to the Basics of Active Inference **10**

1.1 A Unified Brain Theory 10

1.2 Not Physical Quantities 12

1.3 Minimization of Prediction Errors 14

1.3.1 It's All About Predictions 14

1.3.2 Two Types of Predictions, Two Types of Creatures 15

1.4 Perception and Action United 20

1.5 Homeostasis Through Free-Energy Minimization 22

1.5.1 Theoretical Overview 22

1.5.2 Meaningful Examples 23

2 A Mathematical Guide to the Basics of Active Inference **25**

2.1 Information Theory 25

2.1.1 An Intuition for the Notion of Surprise 26

2.1.2 The Formal Definition of Shannon Information 27

2.1.3 The Shannon, the Bit and the Nat 31

2.1.4 Information Entropy 33

2.2 Bayesian Inference 37

2.2.1 The Makings of Bayes' Theorem 38

2.2.2 From Data to Hypothesis and From Hypothesis to Data 45

2.2.3 The Generative Model in Active Inference 50

2.3 Variational Inference and Free Energy 53

2.3.1 The Kullback-Leibler Divergence 54

2.3.2 Free Energy as Inherent and Excess Surprise 59

2.3.3 Derivations of the Free Energy Equations 62

3 Active Inference and Affective Phenomena **70**

3.1 Characterization of Affective Phenomena 71

3.1.1 Why Are There Affective Phenonomena 72

3.1.2 What Are Affective Phenonomena 72

3.1.3 Allostasis 75

3.2 An Active Inference Account of Affects 76
3.2.1 Could Free Energy Be Felt? 76
3.2.2 Affects as Free Energy Minimization Mechanism 81
3.3 The Cortex as a False Positives and Free Energy Reductor 82
3.3.1 True/False Positives Asymmetry and Free Energy 84
3.3.2 The Fast Track and the Slow Tack to the Amygdala 85
3.3.3 Frontal Cortex Regulation of the Limbic System—Two Cases 88

Discussion 92
Counterfactuals 92
The Brain's Philosophy of Science 94

References 97

Appendices 106
Appendix 1 | Variational Free Energy from Jensen's Inequality 106
Appendix 2 | Modulation of Perception Through Utility-Based Prior Probabilities 110

List of Figures

1.1 *Visual Representation of the Generative Model and the Generative Process* . 16

1.2 *Discrepancy Minimization Via the Perception and Action Loops* 21

2.1 *The Graph of Two Candidate Functions for Surprise* 28

2.2 *Coin Entropy as a Function of Bias* 35

2.3 *Probability Distribution of a Fair Die and a Biased Die* 36

2.4 *Event A and Event B in the Universe U Before an Observation Is Made* . . 39

2.5 *The Reduced Universe U_r, Given that Event A Was Observed* 40

2.6 *Event A and not Event B in the Universe U* 43

2.7 *Reduced Universe of Hidden States s_i Intersecting with Observation o* 52

2.8 *The Target Probability Distribution $P_t(X)$ and Two $Q(X)$ Distributions from the $\mathcal{Q}_f$ Family.* 58

2.9 *Variational Free Energy as an Upper Bound on Surprise* 61

2.10 *Free Energy Minimization Through the Perception and Action Loops* 63

3.1 *Affective Concerns and The Free-Energy Principle* 78

3.2 *Affective Features and Variation in Free Energy (First Derivative)* 79

3.3 *The Fast and the Slow Road to the Amygdala* 87

3.4 *Characteristics (Data) of a Rope and a Snake (Hidden Causes) in the Universe U* 110

List of Tables

2.1 *Matrix Representation of Event A and Event B in Universe U* 41

3.1 *Confusion Matrix for the Detection of a Snake* 84

3.2 *Summary of the Values for Key Bayesian Terms of the Snake and Rope Example* 113

Abstract

To achieve completeness, cognitive (or mind) architectures must integrate emotions—a dimension of the mind that has historically resisted formalization. Active inference and the free energy principle offer a unified, neurologically and mathematically grounded theory of the mind and the brain, presenting a promising framework for addressing the complexity of emotions. This honours thesis begins by providing a conceptual and mathematical introduction to the fundamentals of active inference and the free energy principle, including derivations of the three key free energy equations. It then explores affective phenomena as characterized by the Human Affectome project, before analyzing how they can be understood as manifestations of the free-energy minimization process. It proposes that affects may be the subjective experience of free energy variation within the brain. Finally, the thesis examines the interaction between cognition and emotion, suggesting that cognition may serve to reduce the brain's rate of false positives, thereby optimizing the its predictions.

Introduction

Studying the ways various species fly might have been helpful in designing flying machines but flying machines do not require flapping wings to successfully defy gravity. This is an often-cited example when it comes to describing the relationship between artificial intelligence and biology. If the goal is intelligent behaviours, the means by which this goal is reached is irrelevant. Nevertheless, when failing to create intelligent behaviours, it makes sense to draw inspiration from systems that achieve intelligence—most preeminent among these systems are animals.

In deep learning, for instance, Goodfellow, Bengio and Courville (2016) note that artificial neural networks, at the heart of deep learning, did emerge from efforts to model biological learning. They also point out that while some "[. . .] researchers cite neuroscience as an important source of inspiration, others are not concerned with neuroscience at all." (p. 16) And one of the reasons for that lack of interest can be found in the current state of knowledge in neuroscience, as "[. . .] we simply do not have enough information about the brain to use it as a guide." (p. 15)

By contrast with deep learning, in other fields of research that intersect with artificial intelligence, like computational neuroscience and cognitive science, the primary goal is to model neurobiological systems and to uncover the algorithms that generate them. But the boundaries between these areas are not hermetic at all, and many researchers move back and forth between them (Goodfellow, Bengio & Courville, 2016). A good example of this fluidity is the recent effort to seek an alternative to the backpropagation algorithm—the algorithm at the heart of deep learning's success. One of the motivations behind this quest is that some of the limitations of backpropagation, such as the problem of catastrophic interference, might be mitigated through a more biologically valid algorithm because animals do not suffer from catastrophic interference.

Prominent in this search for neurologically valid learning algorithms is the neural coding framework which proposes various neural generative models of neural networks inspired by the theory of predictive processing and predictive coding (Ororbia & Kifer, 2022; Salvatori et al., 2023), where the brain is viewed as a Bayesian inferential organ that constantly makes top-down predictions about the stimuli it expects to collect (more on this later).

These biologically valid approaches to learning, as well as to memory, are of crucial importance in cognitive science, especially for cognitive scientists developing cognitive architectures. A cognitive architecture constitutes a proposal or hypothesis concerning the structures and processes that produce intelligence (Frankish & Ramsey, 2012); such a proposal often takes the form of a computational model that aims to simulate functions of the mind (e.g., memory, learning, problem-solving, etc.). In recent years, one cognitive architecture that seeks to build on cutting-edge biologically plausible computational models of memory and learning is the CogNGen, developed by Ororbia and Kelly (2022a, 2022b, 2023). The memory modules of the CogNGen are powered by a vector symbolic architecture that uses hyperdimensional computing. The learning algorithms are based on neural generative units powered by predictive coding neural networks. Together, the merging of vector symbolic architecture and neural generative coding makes the CogNGen a very stimulating project. Yet, in its current state, the CogNGen falls within the vast category of cognitive architectures that implement cold cognition, where affective processes are left out (LeDoux, 1998). Hot cognition, by contrast, describes approaches to cognition that include affective processes such as emotion (Abelson, 1963; LeDoux, 1998). Emotions often seem intangible and particularly difficult to study because of their subjective aspect. However, as pointed out by LeDoux (1998), "[t]here is really nothing more or less subjective about the experience of an emotion than about the experience of the redness of an apple or the memory of eating one." (p. 7) Even if such a thing was true, researchers in artificial intelligence probably do not feel incentivized to factor in emotions, especially given the fact that emotions have been the scapegoat for bad decision-making for centuries; the rational mind is supposedly the one responsible for sound decisions. But what if integrating affective-like processes in machine learning algorithms could make them more efficient? After all, our understanding of the reason-emotion dynamics is evolving in new directions (Barrett 2019; Damasio 1994; LeDoux 1998; Sapolsky 2017; Panksepp, 2012). But whether emotions have the potential to improve machine learning algorithms or not, in the end, the task of modelling affective phenomena appears to be more relevant to modellers of the mind, as it falls within their mandate.

As of 2017, the Standard Model of Cognition (Laird, Lebiere & Rosenbloom, 2017) did not include emotions, although it is acknowledged that the model will remain incomplete until emotions—along with other aspects of the mind currently lacking—are integrated. While complaints about the absence of emotion in cognitive models are not new (Miller & Johnson-Laird, 1976; LeDoux 1998; Newell, Rosenbloom & Laird 1989; Simon, 1967), what is new is the emerging agreement that affective phenomena in general are essential to understanding a broad range of human behaviours (Dukes et al., 2021). Also new is the recent effort to elaborate computational models of affective phenomena, many of which are based on the theory of active inference, a theory that builds upon and integrates predictive coding (Barrett, 2017; Barrett & Satpute, 2013; Seth & Friston, 2016; Smith et al., 2019; Smith, Parr & Friston, 2019; Tschantz et al., 2022). Substantial research has also been deployed in the affectivization of reinforcement learning (Moerland, Broekens & Jonker, 2018). As it turns out, the CogNGen, with its neural generative units—anchored in predictive coding and the free-energy principle—falls within the domain of active inference. Therefore, active inference is a good place to initiate a reflection on the integration of affective processes in the CogNGen. One particularly important question becomes the following: How does active inference account for affective phenomena? To answer this, we need to present active inference, a theory notoriously difficult to approach (Lindsay, 2021). In what follows, we will provide a conceptual, and then a mathematical guide to the basics of active inference. Equipped with this entry-level understanding, we will try to make sense of affective phenomena, and then revisit them through the lens of active inference.

Chapter 1

A Conceptual Guide to the Basics of Active Inference

The main purpose of this section is to provide a high-level and conceptual explanation of the theory of active inference and of the free-energy principle. We will begin by briefly touching on the motivation behind active inference and on misconceptions about free energy. Then, after providing a non-technical definition of free energy, we will discuss perception as unconscious inference, state how active inference unites perception and action, and how, as free energy minimization mechanisms, they maintain an organism in homeostasis. In the next section (section 2), we will introduce concepts from probability and information theory to be able to give a formal definition of free energy and restate the principle in more intuitive terms.

1.1 A Unified Brain Theory

Active inference is a theory about how living organisms resist entropy, the natural tendency of matter and energy to disperse (Friston 2010; Parr, Pezzulo & Friston, 2022; Buckley et al., 2017). Living organisms find themselves in a very special state, a highly unlikely arrangement of molecules whose preservation, in the face of an ever-changing environment, requires energy.

For instance, the temperature of an organism's environment fluctuates. But outside of a certain temperature interval, the integrity of an organism's body is compromised, to a point where the organism stops functioning—when the molecules of that organism are rearranged into one of the many states that are not conducive to life. Naturally, inanimate objects that happen to be in a state that is far from thermodynamic equilibrium quickly settle back to equilibrium. A cup of hot coffee represents a system that is not in thermodynamical equilibrium with its environment—and that's why a cup of hot coffee inevitably reaches room temperature if no energy is injected into it. But the body of an organism needs to adapt, not to disperse or reach equilibrium with its environment. A living organism is composed of cells whose fluid membrane acts as customs control—deciding what flows in and out of the cell through various proteins. When the temperature falls below the comfort zone, the membrane fluidity decreases. Similarly, when the temperatures rises above the comfort zone, the membrane fluidity increases. In both cases, the new fluidity compromises the cell's functioning, and since living organisms are made of cells, their functioning as a whole is compromised as well. To stay alive, adaptation is required. A dog in the sun, for example, will deploy an array of strategies for temperature regulation, from panting to moving to a shadier spot. These actions help the dog keep its entropy low. Other living organisms are allowed to have their body temperature fluctuate with the environment temperature without losing control of their cell membrane fluidity. From plants to bacteria to snakes, such organisms cope with the variation of temperature by changing the composition of their cell membrane in such a way that fluidity remains constant (Audesirk, Audesirk & Byers, 2014). The same goal is achieved: the organism does not disperse but stays within a small set of low entropy states. And some creatures combine these two approaches. The caribou, a mammal whose environment often involves snow, needs to keep their core body temperature around 38 degrees Celsius, but the temperature of their legs (in the snow), can drop to near freezing point. Again, the cells in their legs change the molecular composition of their membrane to keep a working fluidity (Audesirk, Audesirk & Byers, 2014). Such is the creativity of life.

The British ecologist G. Evelyn Hutchinson poetically vocalized this unique characteristic of living organism: “Disorder spreads through the universe, and life alone battles against it.” (as cited in Audesirk, Audesirk & Byers, 2014, p. 96). The theory of active inference proposes that a simple principle underwrites this accomplishment. In the case of organisms endowed with a nervous system, it is mainly through the nervous system that the principle is at play and keeps the organism alive. Active inference is consequently also a theory of how the brain works at its most fundamental level. Pioneer of active inference and of the free-energy principle, Karl Friston (2009) asserts that when the brain is considered through the lens of the free-energy principle, “[...] nearly every aspect of its anatomy and physiology starts to make sense.” (p. 293) In that sense, the theory unifies all structures and functions of the brain; what layers of neurons do, what brain regions do, and what the whole brain does globally, follow an abstract operation that can be captured mathematically—the minimization of a quantity called free energy (Friston 2009, 2010; Parr, Pezzulo & Friston, 2022).

1.2 Not Physical Quantities

The principle that powers active inference is the minimization of free energy, which is called the free-energy principle. In the context of active inference, free energy refers to a *statistical quantity*, and not to physical energy that is measured in joules. This is a rather abrupt way to begin our presentation of active inference, but it is better to clear up any confusion upstream.

The concept of free energy originates in thermodynamics. The term has been imported into statistics and machine learning by Hinton and Zemel (1994) due to a similarity between the mathematical form that a statistics equation takes and the form the Helmholtz free energy takes (Buckley et al., 2017). In thermodynamics, Helmholtz free energy is understood to be the energy extractable from a system, and which can be used to do work. Mathematically, it is defined as follows:

$$F = U - TS = U - T\left(\frac{Q}{T}\right) = U - Q = Potential\ Energy - Heat \tag{1.1}$$

where F stands for the free energy of the closed system (in joules), U for the (potential) energy of the closed system (in joules), Q for heat, T for the absolute temperature of the surroundings (in kelvin), and S for the entropy of the system (in joules per kelvin). As the system reaches thermodynamic equilibrium with its environment, the free energy becomes zero—there is no more usable energy left to do work (Williams, 2018). By contrast, Hinton and Zemel's equation takes this form:

$$F = E - H = Energy - Entropy \tag{1.2}$$

where E represents energy (not in joules) and H stands for information entropy; by analogy with Helmholtz's equation, they called F free energy. But both E and H do not represent physical quantities.

The primary similarity between statistical (machine learning) free energy and Helmholtz free energy lies in their mathematical form (Buckley et al., 2017; Friston & Stephan, 2007). Therefore, it may be advisable to avoid drawing an analogy between the two, as it could lead to incorrect conclusions about active inference. From now on, when writing "free energy" without specifying which type, we will assume that we refer to the free energy of statistics and machine learning, a quantity that is not physical, but rather statistical.

The free-energy principle thus deals with a statistical quantity measurable in bits (when based 2 logarithm is used) or nats (when the natural logarithm is used), as defined in Shannon's information theory. Similarly, as just mentioned, it is important to note that, when discussing entropy in active inference, entropy sometimes refers to Shannon (or information) entropy and sometimes to thermodynamic entropy (measured in joules per kelvin). Although the two are intimately related (Stone, 2022), the type of entropy omnipresent in the conceptual and mathematical framework of active inference is the information-theoretic one. The nature of this abstract quantity will become more tangible later when we dive into the rudiments of information theory (see section 2). For now, suffice it to state that one consequence of the free-energy principle and the reduction of informational entropy is that it allows the organism to deal with "physical" entropy—think of maintaining one's body in a functioning state (i.e., not dead). In some sense, the free-energy principle constitutes a bridge between the informational and the physical. It is a very important point: the minimization of information entropy entails the minimization of physical entropy (Friston 2010; Parr, Pezzulo & Friston, 2022). This will make more sense later.

1.3 Minimization of Prediction Errors

1.3.1 It's All About Predictions

It is natural to expect free energy to be a good thing, something to maximize, which is misleading: free energy is a quantity to be minimized. If we think about Helmholtz free energy, a system that minimizes it is one heading towards thermodynamical equilibrium with its environment with zero joules of usable energy to do work. This is certainly not a state in which living organisms want to end up. In active inference, free energy can be understood in terms of *prediction error*: it is the gap between how an organism expects the world to be and how the world is. In other words, free energy is the prediction error that results from the comparison between the prediction a brain makes about its environment and the sensory information it samples from this same environment. As Lindsay (2021) sums up, "[...] everything the brain does can be understood as an attempt to minimize free energy—that is, to make the brain's predictions align as much as possible with reality." (p. 345) Now we can better appreciate why free energy, thought of as prediction error, would be a quantity that we want to have as little as possible of.

1.3.2 Two Types of Predictions, Two Types of Creatures

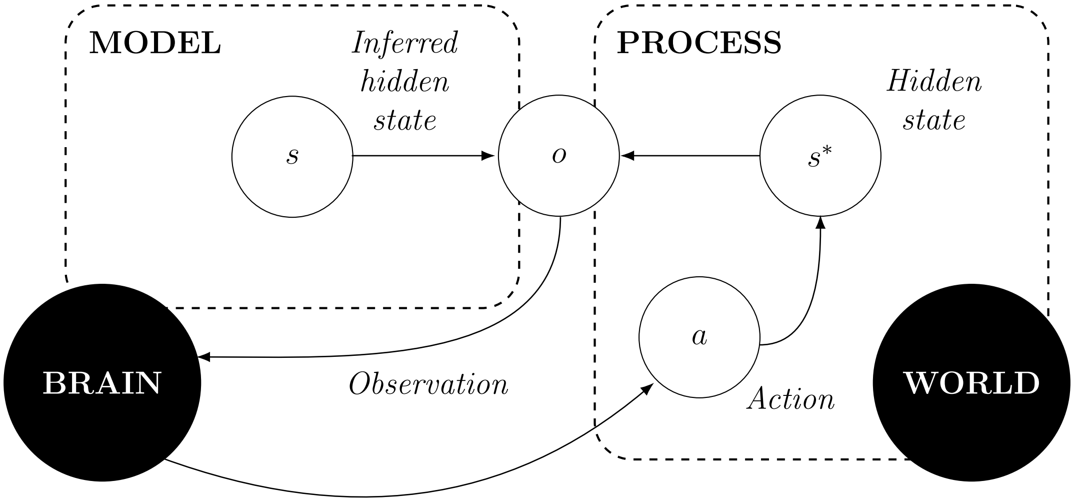
In active inference, there are two categories of predictions that the brain makes and it is important to distinguish the difference between them to avoid confusion. The first type of prediction is when the brain infers the cause of a sensation. The free energy related to predicting the hidden cause or state of an observation is called *variational free energy* (the term variational comes from the underlying mathematics that we will cover later). The second type of prediction is when the brain infers the next likely state of the world based on the current one. The free energy related to predicting what comes next is called *expected free energy*. The first type is the most fundamental one and is thought to be shared by all adaptive systems. The second type is believed to require more complex representations of the world and be implemented through hierarchical cognitive architectures. Interestingly, some consider the capacity to minimize expected free energy as marking the difference between simpler and more complex life forms (Parr, Pezzulo & Friston, 2022).

Predicting the Cause of a Sensation

Since the sensation comes after the cause, this prediction has a retrospective aspect to it. It is also anchored in the view that perception is inferential. This idea was first advanced by Helmholtz, who coined the term “unconscious inference” (1866). That is, there is a difference between sensation and perception and this difference takes center stage in active inference—and more generally in predictive coding theory (Friston, 2005, 2012; Rao & Ballard, 1999) and other approaches in the Bayesian brain hypothesis family (see Parr, Pezzulo and Friston (2022) for a discussion of the relationships between active inference and related theories like predictive coding). A sensation consists of the reception of sensory data; no interpretation is made. For instance, a visual sensation (or reception) consists of our visual sense organs receiving light stimuli from the world. Perception (or recognition) consists of inferring what caused the sensation—was it a bat or a leaf?

Figure 1.1

Visual Representation of the Generative Model and the Generative Process



Note. Using its inner model, the brain infers a hidden state s given incoming sensory data o (observation) and makes predictions about the sensory data that it expects to collect based on its generative model. The generative process belongs to the world, where reside the true hidden state s^* that generates observation o. The brain can also take action a to change the hidden state s^*, giving rise to different sensory data. (This figure is adapted from Figure 2.2 of Parr, Pezzulo and Friston (2022))

This is a good place, with the assistance of Figure 1.1, to introduce some of the terminology used in active inference. In this text, we will name the variables following Sajid et al. (2021). The predictions an organism makes are generated by an organism's *inner model* of the world. We call it the *generative model*. As Friston (2005) writes, "Recognition (i.e. inferring causes from sensation) is the inverse of generating sensory data from their causes. It follows that recognition rests on models, learned through experience, of how sensations are caused" (p. 815). It is the world that causes the observations an organism senses. We refer to the world as the *generative process*. These causes—whether real or inferred—are often called *hidden states*. Crucially, the organism does not have direct access to the causes of its sensory data—hence its model for it and the name "hidden states." This dynamic is nicely captured by Figure 1.1, where the generative process' hidden states are represented by the variable s^*; these hidden states give rise to observations, labelled o, which compose the interface between the brain and the world. Based on the observations o, the brain inferred and hypothesized a cause s, and by inverting its model, it generates a set of predicted observations—again, according to its inner model. The brain's predicted observations (top-down) collide with the world's actual observations (bottom-up). (We will turn to the action a momentarily.) If an organism has a bad generative model, it will constantly be in a state of surprise because its predictions will not match the output of the generative process. For instance, if an organism is presented with a lion, and sample sensations (e.g., visual and auditive), but instead of perceiving a lion it perceives something inoffensive like a plush toy, we would say that its model is bad—and dangerously so.

Our visual system is so effective that it might be difficult to feel that it requires an inference to label a sensation, but it does. If we are paying attention, we might notice moments in our daily life where we are presented with a visual sensation for which we have no explanation (inferred hidden state); it usually does not take long before we subsequently form a perception that makes sense of the puzzling sensation. Nevertheless, thinking about other senses or some forms of visual illusion usually makes this more intuitive. Humans are more often wrong about the cause of auditive stimuli than we are about visual ones. We often hear a sound without being too sure what caused it and usually seek the assistance of our eyes. A sound, like any other kind of stimulus, is often compatible with many different perceptions. In the language of philosophy of science, we could say that perception is often underdetermined by sensation (Stanford, 2023). This happens when a model takes in data, in the form of observations, and is unable to hypothesize a unique cause, because of the compatibility between the same data set and various causes. For instance, if the input data is "curviness," then rope, snake, and many more causes are valid candidates. The data is insufficient to determine a unique cause—it is underdetermined.

The Necker cube is known to lend itself to two different perceptions. One of the squares making the cube can be perceived as the front face of the cube, but also as the back interior. Given the same visual stimuli, we can go back and forth between two perceptions compatible with the sensory data but incompatible between them. The same goes for the famous painting titled My Wife and Mother-in-Law by W. E. Hill.

Let us consider one last example of perceptual illusion, but this time a multimodal case which involves two sensory systems. The McGurk effect is such a multimodal illusion which, as with most illusions, usually persists even once we know the trick (McGurk & MacDonald, 1976; Tiippana, 2014). In one version, the sound [ba] is played at the same time as a person pronounces [ga], but the sound produced is muted and instead, the sound of [ba] is played. Many people will perceive a third sound, [da], which results from the conflicting sensory inputs from the eyes and the ears. This third sound is inferred by the brain. In another version (Rosenblum (on BBC), 2010), the sound [ba] is paired with a person moving their lips in the way we do when pronouncing [fa]. Many people hear [ba] if they do not look at the person, but hear [fa] if they do. In this second case, the visual inference interferes with the auditive one and wins, such that we have the auditory experience of [fa].

Regardless of how insightful they are in revealing how the top-down activity of our brain teams up with the bottom-up sensory input in building a perception, in the case of the images, there is no correct interpretation. In the case of the McGurk effect, it is more ambiguous whether one inference is more accurate than another. But in the real world, there usually is a correct inference. For instance, a tree leaf swirling and a flying bat are two very different objects. Let us imagine we sample our environment and the sensory data we collect leads our brain to infer a bat. Each time we set our gaze on the swirling object, we collect more data against which we can test our inference. Given a certain angle, the object might really look and move in a way that is consistent with our model of a bat. However, when the object changes angle, we collect new sensory data in which the object appears flat and thin. At this point, the set of observations that we expect by inverting our model for the bat does not fit our new observations which are more consistent with a leaf. But before having this angle, the sensory data we had sampled agreed with at least two different perceptions—a bat and a leaf. At this earlier stage, our perception was underdetermined by our sensations. Then, through further observations, we could eliminate the bat candidate perception. We note that although the leaf hypothesis is the best candidate given our further observations, we still have no direct access to it—maybe it is neither a bat nor a leaf.

To sum up, for a system to be adaptive, it matters a lot that it is able to make appropriate inferences about its sensation, otherwise, it is doomed to misapprehend its environment, constantly decreasing its chances of survival—it is hard to adapt to a situation we do not perceive correctly.

Predicting What Comes Next

The type of free energy discussed so far was variational free energy; with its retrospective aspect, variational free energy deals with the past and the present; the past cause of a present sensation is inferred. By contrast, expected free energy is more prospective as it is concerned with future data. Planning and decision-making rest on the capacity to anticipate the future. This type of cognition is thought to require more network complexity, thus marking a smooth transition between adaptive systems lacking planning and those endowed with this capacity at different degrees (Parr, Pezzulo & Friston, 2022). Planning can go from a few seconds to years into the future.

In active inference, planning and decision-making are processes grounded in inference. The generative model makes these prospective inferences based on a policy. In this context, a policy is simply defined to be a sequence of actions (Sajid et al., 2021). The task of the inner model is to select the best possible policy—the best sequence of actions. As before, by best, we mean the option that minimizes free energy, but now it is the variational free energy we expect from anticipated future observations.

1.4 Perception and Action United

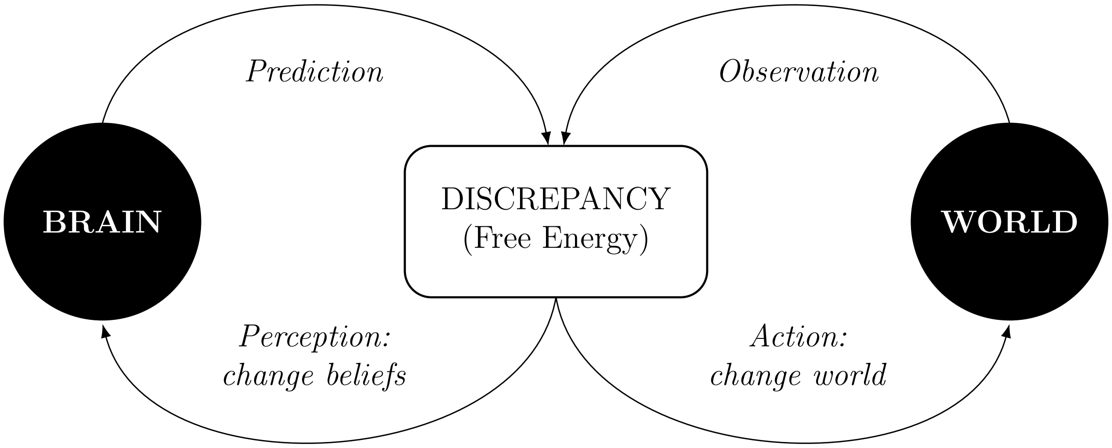
Although we just discussed the difference between the two types of predictions of active inference, in what follows, let us simply explore the general idea that the brain aims to minimize the discrepancy between the predictions it makes (whatever the type) and the data provided by its sensorium.

There are two ways to minimize free energy. We already discussed the first one, which is through *perception*: by updating its inner generative model based on the sensory inputs it samples from the world, an organism can reduce the discrepancy between its predictions about reality and reality itself. Perception is thus about changing one's beliefs. The general idea that the brain is a Bayesian inference machine is therefore subsumed into active inference. However, for technical reasons explored later, active inference doesn't directly make use of exact Bayesian updates but relies on approximations—the mathematics and algorithms used remain very much in the spirit of Bayesianism.

At the core of active inference is the idea that *action* is the second mechanism through which free energy can be minimized, thus unifying perception and action under a unique process. That is, both perception and action fundamentally do the same thing, minimize the same quantity (see Figure 1.2). Prediction error can be reduced by acting on the world, by changing it. There are two main ways to change the world: an organism can (1) act on its body to get more observations of the world and (2) act on its environment to make it align with its beliefs about the environment. In active inference, the adaptive system's body is conceptualized as part of the environment. In essence, everything that is not the inner model is the environment and can thus give rise to observations. Following the movement of a flying bat with our gaze is a form of action onto the world because we move our eyes, head, or whole body to do so. By doing so, we can collect more data and give our inner model a chance to improve. The second way to change the world is more direct. When there is a discrepancy between our model and our observations, we can alter our environment so that it produces the observations our model predicts. In Figure 1.1, action is represented by

Figure 1.2

Discrepancy Minimization Via the Perception and Action Loops



Note. When the brain prediction about observations does not match perfectly those observations, a discrepancy (free energy) results from this prediction error. The discrepancy can be reduced through perception (by changing beliefs about the world) or through action (by changing the world, such that the new observation matches the predictions. (This figure is adapted from Figure 2.3 of Parr, Pezzulo and Friston (2022))

the variable a, which acts on the hidden state s^* of the generative process. To give a crude example, if we believe we are drinking water (s), but we are not—that is, the sensory input (o) we receive from our immediate environment does not match the observations we expect from our belief "drinking water"—we can reach for the cup of water and start drinking from it (a), modifying the hidden state (s^*) of the world. In the words of Parr, Pezzula and Friston (2022), "[...] when the prior belief dominates [...], it is maintained even in the face of conflicting sensory evidence—and it induces a grasping action to resolve the conflict" (p. 202). Through the reach-the-cup-and-drink action, we have aligned the world with our model; both now generate the same predicted and actual observation. In active inference, actions are thus propelled by belief-observation contradictions. The most fundamental beliefs that often get contradicted are the ones an adaptive system has about the state in which it should be in order to persist in the face of environmental fluctuations. Examples are in order.

1.5 Homeostasis Through Free-Energy Minimization

1.5.1 Theoretical Overview

Homeostatic states are states conducive to the existence of an organism and consist of many parameters (e.g., body temperature, glucose level, oxygen level, acidity) that must stay around internal set points (e.g., body temperature around 37^oC). To illustrate how the free-energy principle works toward achieving homeostasis at the most elementary level, Friston and Stephan (2007) propose the (fictitious) adaptive snowflake—a snowflake with wings. In order to preserve its crystal shape, an adaptive snowflake must avoid phase transition. It can do so by maintaining an altitude that corresponds with a temperature at which it does not melt. In the adaptive snowflake case, if its perception mechanism is deficient, it might not notice that it is coming dangerously close to a zero-degree Celsius temperature region. In turn, if its action mechanism is deficient, it might not use its wings to regain an altitude with a colder temperature.

Of course, living organisms have many more parameters that must have values within a certain range for the organism to keep its integrity. Writing about adaptive systems in general, Friston and Stephan (2007) explain: "Systems which fail to minimize free energy will have sub-optimal representations or ineffective mechanisms for action and perception. These systems will not restrict themselves to specific domains of their milieu and may ultimately experience a phase-transition (e.g., death)" (p. 428-429). We can now make a first attempt at making the connection between the free-energy principle and homeostasis. Since we are still using prediction error as a proxy for free energy, this will not be the final formulation of the connection, but nevertheless a useful one. We already noted how having a sub-optimal representation of the world threatens the life-conducive conditions in which an organism needs to be. A sub-optimal representation means a high rate of prediction errors. However, in active inference, there are situations where even a good representation is accompanied by high prediction errors. We touched on this idea when describing how action is propelled by belief-observation contradictions. There is a category of beliefs that an adaptive system must not update. Instead, it must always change the world such that the world adapts to the beliefs in question. These are homeostatic set points, bodily states that are necessary conditions for the existence of an organism. The adaptive system will always predict that its observations match these beliefs. When they do not, an important level of prediction error (free energy) follows. To minimize it, action must be taken.

This is a good place to revisit a statement made earlier: the minimization of prediction error (related to information entropy—stay tuned) entails homeostasis (minimization of physical entropy).

1.5.2 Meaningful Examples

Let us explore two examples, both from Parr, Pezzulo and Friston (2022). The first example is about body temperature. If we find ourselves in a room where the temperature is rising, our thermoreceptors will provide us with new sensory data about our body temperature. If this temperature reading is outside of the range in which we expect our body to be, we will be confronted with a surprising discrepancy. This is because we have an extremely high expectation (i.e., confident belief) that our body temperature is in the comfort zone since it is a condition of our existence. These beliefs are not held consciously but encoded in the nervous system. Through perception, we arrive at the correct inference about the state of our body temperature. However, in this case, further adapting our beliefs to this new situation so that we do not find it surprising, is not exactly adaptive. Rather, the best way to resolve the discrepancy would be to change the world and make it agree with our inner model. We would make the world agree that our body temperature is in the homeostatic range, reducing the surprise. Hence, opening a window and letting colder air flow into the room might be such a free-energy-minimizing action. Dogs, for their part, might automatically start to pant and move their body to a cooler area.

In the second example, we consider glucose levels. Let us imagine that we receive sensory data from the glucose level in our blood and the report says that it is low. Based on this sensory data, we infer that we are hungry. At this stage, the adequation between our belief (we are hungry) and the state of the world (low glucose level) reduces the divergence part of free energy. However, we are still hungry—perception only helps us figure it out. To further deal with this surprising state of hunger and its accompanying sensory inputs, we must change the low-glucose blood level sensations; to do so, we can act—eat something.

This brings us to the end of this (partial) conceptual overview of active inference. In what follows, we will try to make sense of its basic mathematical expressions. We will only focus on variational free energy in its most simple form.

Chapter 2

A Mathematical Guide to the Basics of Active Inference

Viewing the brain as an inference machine leads to the question of how this inference is performed, both physically and computationally. Here, we are interested in the computational side of the question. As hinted earlier, Bayesian inference seems a suitable candidate. We will therefore explain why it is indeed the case, and why we need to approximate it. But first, let us define this elusive statistical quantity to which we alluded to and which is at the core of the computations of active inference. This quantity is information, which is the stuff of surprise, (information) entropy, and free energy.

2.1 Information Theory

One of the goals of information theory is to quantify information. The notion of information is conceptualized as a measure of surprise, a construct intimately related to probability. Information entropy, for its part, is built on surprise since it is defined as the average surprise associated with a system, comprising sets of events or states. In this section, we will begin by forging an intuitive sense of the relationship between probability and surprise. We will formally define surprise, also known as Shannon information, and apply it to different situations. We will then describe its units of measurement. Finally, we will define information entropy and provide a few examples.

2.1.1 An Intuition for the Notion of Surprise

Intuitively, we can *feel* that an unlikely event is more surprising than a likely one. For instance, the probability of encountering an elephant in a public park in Ottawa (Canada) is much smaller than the probability of encountering a squirrel—there are many of them. We would also expect our level of surprise to be different in the two situations: the elephant will produce a higher level of surprise than the squirrel. Thus, probability and surprise appear to be proportional, but inversely so—small probability equals high surprise, and vice-versa. If we were to call a friend to share the news that we just saw a squirrel in a park, we can expect them to be somewhat baffled by this deeply uninteresting news. Everyone knows that there often are squirrels in parks. There is not much information conveyed here. By contrast, in the elephant scenario, our friend would likely understand the point of us calling, in part because they surely did not know that there was an elephant in the park. In this case, a lot of information is conveyed. We can compactly express our three observations with these three simple statements:

$$\text{Probability(``Squirrel in park'')} > \text{Probability(``Elephant in park'')}$$
$$\text{Surprise(``Squirrel in park'')} < \text{Surprise(``Elephant in park'')}$$
$$\text{Info(``Squirrel in park'')} < \text{Info(``Elephant in park'')}$$

Of course, the level of surprise felt may vary from person to person and be expressed inconsistently within the same individual. Although psychological surprise serves as a nice starting point and often overlaps with Shannon surprise, the latter has a much-needed formal definition. The ensuing presentation of Shannon information and information entropy is based on MacKay (2003) and Stone (2022).

2.1.2 The Formal Definition of Shannon Information

Although we just saw the surprise/information is inversely proportional to probability, the Shannon information is *not* defined as the inverse probability function, which we will call k, and express as follows:

$$k(x) = \frac{1}{P(x)} \tag{2.1}$$

where x is a possible outcome, and $P(x)$ is the probability of the outcome x. Instead, the Shannon information, denoted by the letter h, is formally defined as the logarithm of the inverse probability of observing an event (or outcome):

$$h(x) = \log\left(\frac{1}{P(x)}\right) \tag{2.2}$$

For reference, it is common to encounter this alternative expression:

$$\begin{aligned}
h(x) &= \log\left(\frac{1}{P(x)}\right) && \text{def. of surprise } h\text{, Eq (2.2)} \\
&= \log(1) - \log(P(x)) && \text{by log quotient rule} \qquad (2.3) \\
&= -\log(P(x)) && \text{log of 1 is 0 } (a^0 = 1) \qquad (2.4)
\end{aligned}$$

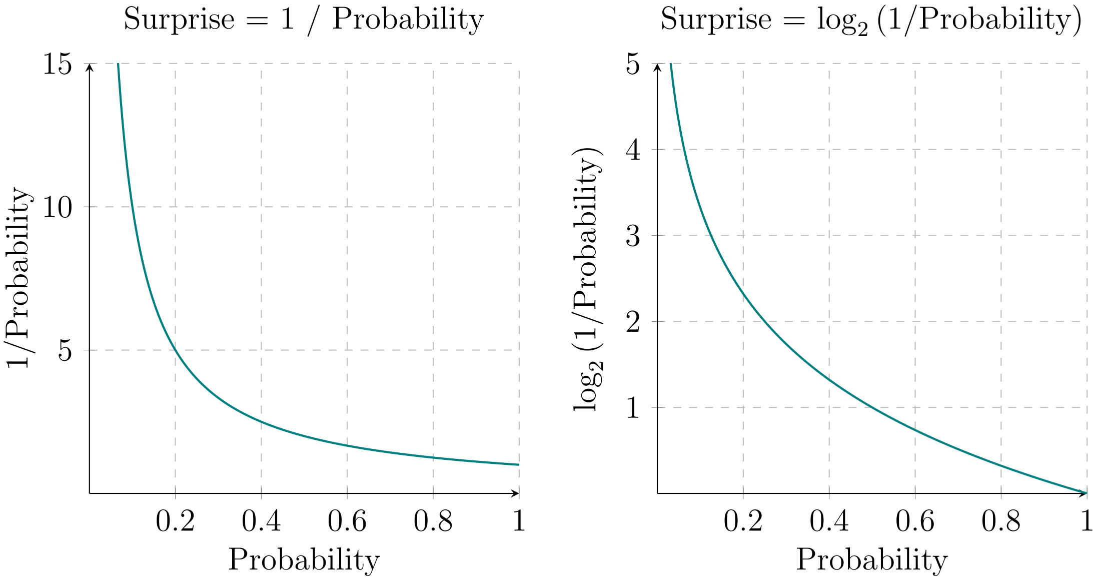
However, I personally prefer Eq. (2.2) because we can easily see the relationship with the inverse probability, which is somewhat lost in the manipulations that lead to Eq. (2.4). In this text, we will therefore mainly use Eq. (2.2).

One may wonder about the relevance of using the logarithm. There are a few reasons. For one thing, it is instructive to compare the graph of the inverse probability function and the graph of the logarithm of the inverse probability function (see Figure 2.1).

An important difference is that, in the case of the inverse probability function, the surprise of an outcome that occurs with a probability of 1 (complete certainty) is not zero. Let us consider the example of a coin to complement the graphs. A fair coin has the probability

Figure 2.1

The Graph of Two Candidate Functions for Surprise



Note. When we apply the logarithm function to the inverse probability function (on the left) we obtain the Shannon information function (on the right), outputting a surprise of zero when the probability of an event is 1.

distribution

$$P(x_h) = 0.5,\ P(x_t) = 0.5,$$

where x_h stands for the outcome "heads" and x_h stands for the outcome "tails." If the outcome of a coin flip experiment is certain, the coin is maximally biased——for instance, the coin always lands on heads or the two sides are heads. We get this different probability distribution:

$$P(x_h) = 1,\ P(x_t) = 0.$$

We would like the surprise of a completely certain outcome (e.g., heads) to be zero, and the surprise of an impossible outcome (e.g., tails) to go to infinity—this is the land of miracles. While both surprise functions make intuitive sense with respect to the latter property, only the logarithmic function provides the former. Indeed,

$$\lim_{P(x)\to 0^+} h(x) = \lim_{P(x)\to 0^+} \log\left(\frac{1}{P(x)}\right) = \infty \tag{2.5}$$

and

$$\lim_{P(x)\to 1^-} h(x) = \lim_{P(x)\to 1^-} \log\left(\frac{1}{P(x)}\right) = 0. \tag{2.6}$$

In Eq. (2.5), the limit as $P(x)$ approaches zero from the right goes to infinity (infinite surprise) since 1 gets divided by an infinitely small number, which is equivalent to multiplying 1 by an infinitely big number; the logarithm of a number that goes to infinity also goes to infinity. In Eq. (2.6), the limit as $P(x)$ approaches 1 from the left goes to zero (zero surprise), since 1 gets divided by a number that gets infinitely close to 1, and the logarithm of 1 is zero. These results correspond to our desired properties. (In the case of the function k, the limit as $P(x)$ approaches 1 from the left is 1, since $1/1 = 1$.)

Another reason to choose the logarithmic function of inverse probability to define information is that it possesses the property of additivity, which means, in this context, that the information coming from two independent events can be added. For instance, the outcome of flipping two (fair) coins is independent—the result of the first coin X does not influence the result of the second coin Y. The variables X and Y denote random variables. A random variable is a variable whose value is determined by the outcome of a random experiment like a coin toss. The independence of coins X and Y is defined as

$$P(X,Y) = P(X)P(Y). \tag{2.7}$$

For instance, the probability of the two coins landing heads is

$$P(x_h, y_h) = P(x_h)P(y_h) = (0.5)(0.5) = 0.25.$$

We can now verify that the logarithm of the inverse probability function is additive:

$$\begin{aligned}
h(X,Y) &= \log\left(\frac{1}{P(X,Y)}\right) && \text{def. of surprise } h\text{, Eq. (2.2)}\\
&= \log\left(\frac{1}{P(X)P(Y)}\right) && \text{by Eq. (2.7)}\\
&= \log(1) - \log(P(X)P(Y)) && \text{by log. quotient rule}\\
&= -\log(P(X)P(Y)) && \text{log of 1 is 0}\\
&= -\log(P(X)) - \log(P(Y)) && \text{by log multiplication rule}\\
&= \log\left(\frac{1}{P(X)}\right) + \log\left(\frac{1}{P(Y)}\right) && \log\left(\frac{1}{x}\right) = 0 - \log(x),\ \text{Eq. (2.4)}\\
&= h(X) + h(Y) && \text{def. of surprise } h\text{, Eq. (2.2)}
\end{aligned}$$

We can also verify that the inverse probability function k, is not additive:

$$\begin{aligned}
k(X,Y) &= \frac{1}{P(X,Y)} && \text{def. of surprise } k\text{, Eq. (2.1)}\\
&= \frac{1}{P(X)P(Y)} && \text{by Eq. (2.7)}\\
&\neq \frac{1}{P(X)} + \frac{1}{P(Y)} && \\
&= k(X) + k(Y) && \text{def. of surprise } k\text{, Eq. (2.1)}
\end{aligned}$$

Lastly, we note that when dealing with dynamics that scale exponentially, like coin tosses or particle arrangements, it is often useful to use logarithms. (After all, the logarithm undoes the exponential: $y(x) = a^x,\ log_a(y(x)) = x$.) For instance, the number of states in which n coins can be is exponential while the surprise is linear:

$$\begin{array}{lll}
1 \text{ coin: } \; 2^1 = 2 \text{ possible states,} & P(\text{all heads}) = 1/2, & h(\text{all heads}) = 1 \\
2 \text{ coins: } \; 2^2 = 4 \text{ possible states,} & P(\text{all heads}) = 1/4, & h(\text{all heads}) = 2 \\
3 \text{ coins: } \; 2^3 = 8 \text{ possible states,} & P(\text{all heads}) = 1/8, & h(\text{all heads}) = 3 \\
 & \ldots & \\
n \text{ coins: } \; 2^n \text{ possible states,} & P(\text{all heads}) = 1/2^n, & h(\text{all heads}) = n
\end{array}$$

We now turn to the question of units.

2.1.3 The Shannon, the Bit and the Nat

The unit varies depending on the logarithm base used. The two most common bases in information theory are 2 ($log_2(x)$) and e ($ln(x)$, the natural logarithm $\log_e(x)$). When base 2 is used, the unit is called the *bit* (for binary digit) or the Sh (the *shannon*); when using base the e the unit is the *nat* (for natural unit of information). We will mostly use base 2. One shannon is defined as the amount of information content provided by an event that has a probability of 50%——like a fair coin flip. Perhaps more meaningfully, it also corresponds to the information provided by an answer to a yes-no question where both outcomes "yes" and "no" are equally probable (50% each). For instance,

$$h(x = \text{yes}) = \log_2\left(\frac{1}{P(\text{yes})}\right) = \log_2\left(\frac{1}{\frac{1}{2}}\right) = \log_2(2) = 1 \text{ Sh}.$$

To better understand how to tally information, let us consider the following scenario: we are on our way to a chocolate factory, and we do not know the itinerary, but we start in the right direction. If we reach a fork in the road, we can either go left or right. If someone would kindly indicate whether we should go left or right to stay on the path of the chocolate factory, that person would give us 1 shannon. If we then reached another fork, we would again need 1 shannon. At this point, 2 shannons would have helped us choose among 4 possible itineraries (left-left, left-right, right-left, right-right), with each itinerary having an equal probability of 1/4 to lead us to our destination.

$$h(x) = \log_2\left(\frac{1}{P(x)}\right) = \log_2\left(\frac{1}{\frac{1}{4}}\right) = \log_2(4) = 2 \text{ Sh}.$$

With 2 shannons, we can thus identify any of four equally likely options, like the season in which someone is born (Lindsay, 2022). Question 1: "Is your birthday during a solstice season?" If yes, then question 2 could be: "Is your birthday in summer?" More generally, n shannons of information allow selecting the correct answer among a total of 2^n equally plausible answers. This number becomes big very fast. For instance, in our trip scenario, with 20 left-or-right indications, we can discriminate between

$$2^{20} = 1\ 048\ 576$$

equally likely itineraries. With no assistance from a knowledgeable source, it would be extremely unlikely ($\frac{1}{1\ 048\ 576}$) that we get to the chocolate factory by choosing randomly. It would also be extremely surprising. We can see that 20 shannons correspond to quite a large amount of surprise.

In this road trip example, it would be correct to say that the person we asked for directions gave us one bit of information. However, as suggested by MacKay (2003), using the shannon as a unit eliminates the ambiguity that often lurks around when using the bit, resulting in confusion. The bit has two meanings. The first one is a binary digit, that is, a digit that takes the value of 0 or 1. So, if we map 0 onto "no" and 1 onto "yes", we can answer a yes-no question by handing a bit. But a bit of information is, as mentioned above, the information content of an event with 50% probability of occurring. To give an example that illustrates the difference between the two types of bits, let us suppose that the correct itinerary that leads to the chocolate factory is simply the one where we keep left at each fork. We could ask someone: Do we need to keep left all the way there? If that person says “yes”, they just gave 1 bit (binary digit), but 10 bits of information. The same asymmetry appears when the outcomes of an experiment are not equally likely. The result “heads” of a coin toss using a coin which is 90% biased toward “heads” only provides $log_2(\frac{1}{9/10}) \approx 0.152$ shannon of information. However, it requires one binary digit to communicate the result. The information is less than one shannon because we *expect* the coin to land on “heads.” Using the shannon, instead of the bit, clarifies the matter.

2.1.4 Information Entropy

We now have the building blocks required to define information entropy—also called Shannon entropy. Information entropy, denoted by the capital letter H, is the expected (average) surprise h (Shannon information) of a set of possible events. Formally,

$$\begin{aligned} H(X) &= \sum_{i=1}^{n} P(x_i)h(x_i) && \text{(2.8)} \\ &= \sum_{i=1}^{n} P(x_i)\log_2\left(\frac{1}{P(x_i)}\right) && \text{def. of surprise, Eq (2.2)} \quad \text{(2.9)} \\ &= \mathbb{E}_{P(X)}\left[\log_2\left(\frac{1}{P(X)}\right)\right] && \text{def. in expectation notation} \quad \text{(2.10)} \end{aligned}$$

where x_i is one of the n possible outcomes of random variable X. For each possible event x_i, we multiply (weigh) its surprise by its probability of occurring. For instance, let us revisit the fair coin. We have the fair coin random variable X with probability distribution

$$P(X) = \{p(x_1),\ p(x_2)\} = \left\{p(x_h) = p(x_t) = \frac{1}{2}\right\}$$

and with entropy

$$\begin{aligned} H(X) &= \sum_{i=1}^{2} P(x_i) \log_2 \left(\frac{1}{P(x_i)} \right) \\ &= P(x_1) \log_2 \left(\frac{1}{P(x_1)} \right) + P(x_2) \log_2 \left(\frac{1}{P(x_2)} \right) \\ &= \frac{1}{2} \log_2 \left(\frac{1}{1/2} \right) + \frac{1}{2} \log_2 \left(\frac{1}{1/2} \right) = 1 \text{ Sh.} \end{aligned}$$

Therefore, the information entropy of a fair coin is 1 shannon. But what if the coin is biased? Figure 2.2 shows a graph of the entropy of a coin with respect to its bias. At the two extremities, the coin is either completely biased toward heads or tails, reducing the entropy to zero. There is no uncertainty and consequently no information gained from the outcome of flipping such a coin. The peak entropy occurs when the coin is maximally random (i.e., fair). In this simple example, we could take the derivative of the entropy function and solve for when its slope is zero to find the peak entropy.

To give one last example and extend our insight into the concept of information entropy, we consider two dice, one fair and one biased, with values

$$V_X = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}\,. \tag{2.11}$$

The fair die has the probability distribution

$$P_f(X) = \left\{ \frac{1}{6},\ \frac{1}{6},\ \frac{1}{6},\ \frac{1}{6},\ \frac{1}{6},\ \frac{1}{6} \right\}, \tag{2.12}$$

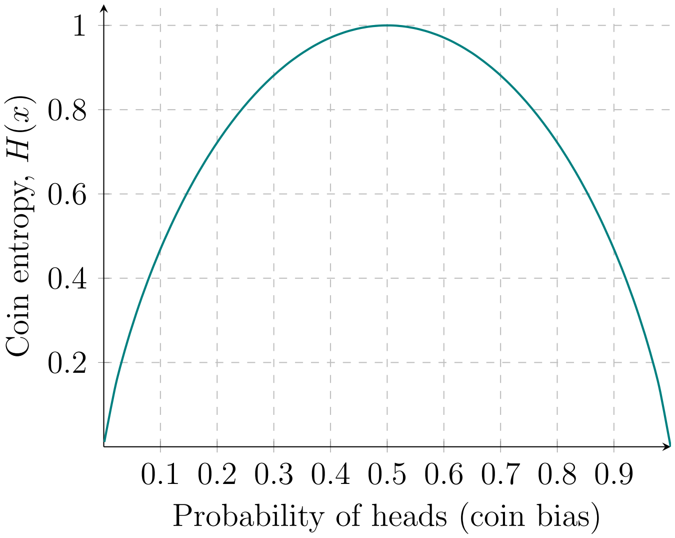
where the subscript f stands for “fair.” The probability distribution of the biased die is

$$P_b(X) = \left\{ \frac{1}{30},\ \frac{1}{30},\ \frac{1}{30},\ \frac{1}{30},\ \frac{5}{6},\ \frac{1}{30} \right\}. \tag{2.13}$$

where the subscript b stands for “biased.” We observe that both distributions sum up to 1,

Figure 2.2

Coin Entropy as a Function of Bias



Note. At the two extremities, when the coin is maximally biased, with the probability of heads being 0 or 1, the entropy is zero (no surprise). The entropy reaches its maximum when the coin is fair (at the 0.5 mark), which is to say maximally random.

as required by this rule of probability:

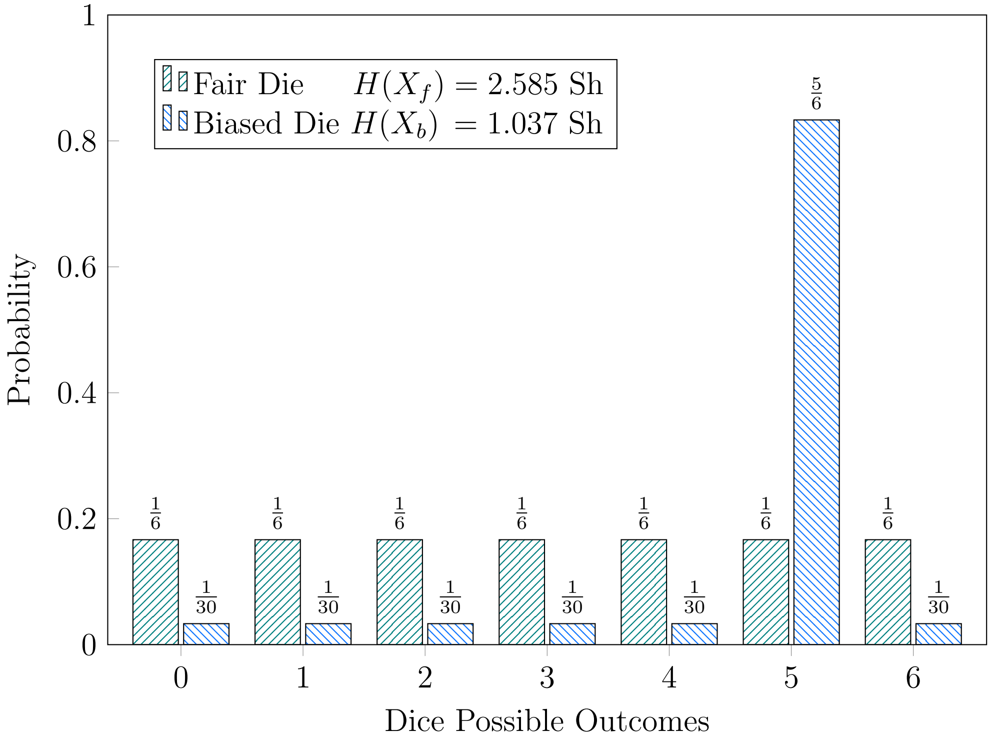
$$\sum_{i=1}^{n} P(x_i) = 1. \tag{2.14}$$

For completeness and visualization sake, the two entropy graphs of the dice are displayed in Figure 2.3.

Since the probability of the fair die is uniform, all possible outcomes have the same

Figure 2.3

Probability Distribution of a Fair Die and a Biased Die



Note. Comparison of the probability distribution of a fair and a biased dice showing the entropy value of each, namely $H(X_f) = 3$ Sh and $H(X_b) = 1.037$ Sh, respectively. The more a distribution is random (uniform), the higher its entropy.

weight and surprise, giving an entropy of

$$\begin{aligned} H(X_f) &= \sum_{i=1}^{6} P_f(x_i) \log_2\left(\frac{1}{P_f(x_i)}\right) \\ &= \sum_{i=1}^{6} \frac{1}{6} \log_2\left(\frac{1}{1/6}\right) \\ &= 6\left(\frac{1}{6}\log_2(6)\right) \\ &\approx 2.585 \text{ Sh.} \end{aligned}$$

The entropy of the biased die is much lower:

$$\begin{aligned}
H(X_b) &= \sum_{i=1}^{6} P_b(x_i) \log_2 \left(\frac{1}{P_b(x_i)} \right) \\
&= \sum_{i=1}^{5} \frac{1}{30} \log_2 \left(\frac{1}{1/30} \right) + \frac{5}{6} \log_2 \left(\frac{1}{5/6} \right) \\
&= \frac{5}{30} \log_2(30) + \frac{5}{6} \log_2 \left(\frac{6}{5} \right) \\
&\approx 1.037 \text{ Sh.}
\end{aligned}$$

Similarly to the coin example, it is the most random die, the one with the highest uncertainty, which has the highest entropy.

As a last remark, in both physics and information theory, one of the most favoured words to refer to entropy is *uncertainty* (Gould & Tobochnik, 2016; Lemons, 2013). The less a system is random, the less it is uncertain, and the lower its entropy. A living organism has a very non-random and certain (low entropy) structure, an organisation that is very far from the equilibrium we find in randomly arranged systems. We could say that its states are extremely biased toward a few arrangements conducive to life. As such, we highly expect to find its constituent parts in one of these arrangements. There is not a lot of uncertainty about how it is organized. Injecting randomness (entropy) into such an adaptive system severely threatens its integrity. And this is precisely the challenge that the environment poses.

2.2 Bayesian Inference

The Bayesian universe contains all the possible events (or characteristics) in which we are interested, and these events are divided into two categories: observable and unobservable. The goal of Bayesian inference is to revise a belief about an *unobservable* event given an *observation* (Bolstad & Curran, 2016). To see how this can be done, we first derive Bayes' theorem, following Bolstad and Curran's (2016) approach. We will use an example adapted from Clayton (2022) to illustrate its various components and, to keep things simple, we will only handle discrete probability distributions, leaving aside the continuous variables. Thereafter, we will apply the Bayesian inference to a made-up experiment that will help us understand the importance that each of the mathematical terms plays. Lastly, we will make salient the reasons why the Bayesian approach to statistics constitutes such a good candidate for providing the mathematical tools needed by active inference.

2.2.1 The Makings of Bayes' Theorem

We suppose that there is a group of 100 people and we are interested in two characteristics: a person has a brain design printed on their shirt (event A); a person is a neuroscientist (event B). In the whole group, 20 people are neuroscientists and 25 people have a brain on their shirt. Figure 2.4 shows a Venn diagram of events A and B (the diagram is not to scale).

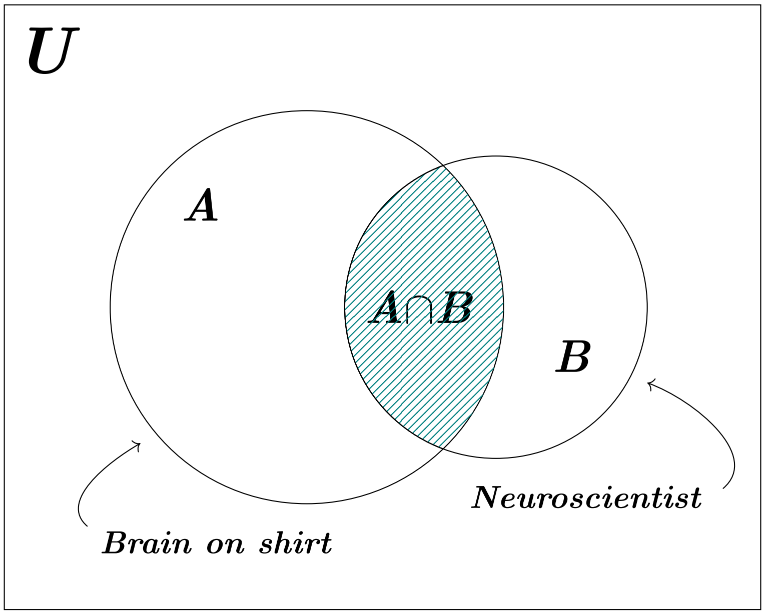
As we can see in the figure, events A and B intersect. This means that they sometimes happen together. In our example, the *observable* event is the shirt, while the profession of the person is *unobservable*. What we want to know is the new probability of B in the reduced universe where A is observed, that is, $P(B \,|\, A)$. Let us suppose that of the 20 neuroscientists, 16 have a brain on their shirt (all these values are shown in Table 2.1. To find the probability of B given A, we calculate a ratio: the joint probability of A and B divided by the probability of A:

$$P(B \,|\, A) = \frac{P(A \cap B)}{P(A)} = \frac{|\, A \cap B \,|}{|\, A \,|}. \tag{2.15}$$

We have just defined conditional probability. The probability of B given A is proportional to their intersection. The more they intersect (happen together), the more observing A increases the chances that event B is the case. But for this reduced universe of A to have a probability of 1, we need to rescale the intersection by $P(A)$ (see Figure 2.5). As for the

Figure 2.4

Event A and Event B in the Universe U Before an Observation Is Made



Note. A person who wears a shirt with a brain on it and who is a neuroscientist falls in the coloured intersection $A \cap B$. (This diagram is inspired by Figure 4.7 of Bolstad and Curran (2016). It is not to scale.)

vertical barres on the right-hand side, they denote the size of a set. Again, the size of A and B is 16, and the size of A is 25. Plugging the numbers in, we obtain

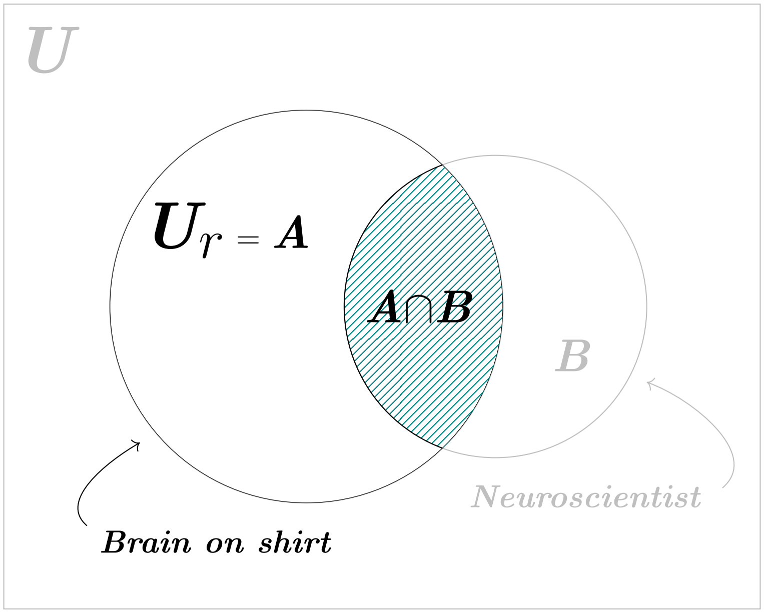
$$P(B \,|\, A) = \frac{|\, A \cap B \,|}{|\, A \,|} = \frac{16}{25} = 0.64.$$

Hence, the probability that a person with a brain on their shirt is also a neuroscientist is 64%. We have our answer.

But Bayes' theorem, although it will give the same answer, takes another form that involves *reverse* and *prior* probabilities, two terms that add considerable insight into what is happening when we compute this forward conditional. Mathematically, we can reverse B

Figure 2.5

The Reduced Universe U_r, Given that Event A Was Observed



Note. Once event A has occurred, the universe U reduces to A (thus, $U_r = A$). (This diagram is inspired by Figure 4.7 of Bolstad and Curran (2016). It is not to scale.)

given A, and write A given B:

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)} = \frac{|\, A \cap B \,|}{|\, B \,|}. \tag{2.16}$$

By formal symmetry, this equation holds. However, we must keep in mind that B is unobservable while A is not—in our example, we see what a person wears but not their profession. Hence, the meaning of this reserved expression differs from the first: it asks for how likely observation A is assuming the unobservable event B occurred. In our scenario, it asks: if a person is a neuroscientist, what is the probability that they have a brain on their shirt? Let us go ahead and calculate this reverse probability:

$$P(A \,|\, B) = \frac{|\, A \cap B \,|}{|\, B \,|} = \frac{16}{20} = 0.8.$$

The reverse probability that a neuroscientist wears a brain on their shirt is 80%. Importantly, this result differs from the 64% probability that a person is a neuroscientist given that there is a brain on their shirt. We will come back to this remark soon.

Table 2.1

Matrix Representation of Event A and Event B in Universe U

	A (*Brain*+)	$\widetilde{A}$ (*Brain*−)	Total
B (*Neuro*+)	**16**	4	**20**
$\widetilde{B}$ (*Neuro*−)	9	71	80
Total	25	75	100

Note. The entries with numbers 16, 4, 9, and 71, represent the intersections. In bold are the values used to compute the conditional probability using Eq. (2.16). The sum of the entries of the first column, divided by 100, gives the marginal probability ($P(A)$).

We can rearrange Eq. (2.16) to express the probability of the intersection as

$$P(A \cap B) = P(A \,|\, B)P(B). \tag{2.17}$$

This equation is known as the *multiplication rule*. By clearing the denominator in Eq. (2.15), we equivalently define the intersection as

$$P(A \cap B) = P(B \,|\, A)P(A). \tag{2.18}$$

Combining both expressions for the intersection, we obtain the equality

$$P(A \,|\, B)P(B) = P(A \cap B) = P(B \,|\, A)P(A) \tag{2.19}$$

$$P(A \,|\, B)P(B) = P(B \,|\, A)P(A). \tag{2.20}$$

This last expression is essentially Bayes' theorem. We get its conventional form by dividing both sides by the probability of B:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} \tag{2.21}$$

On the left-hand side of the equation, we have the *posterior* conditional probability, $P(B \mid A)$. The *posterior* probability represents our belief in B after observing A. The term $P(A \mid B)$ has many names, including the *likelihood*, the *reverse* probability and the sampling probability; it tells us how likely it is that we observe A assuming the occurrence of the unobservable ("latent") event B. We call $P(B)$ the *prior* probability, consisting in our belief in the unobservable event B prior to the new data that we incorporate into our reasoning in the form of the observation of event A. Once the posterior probability is calculated, this new confidence in B becomes our new prior—we update our belief in B. Lastly, the denominator $P(A)$ is often referred to as the *marginal* or the *pathway* probability; it represents the probability that we observe event A. Alternatively, we can call it the *evidence*, as it is the probability of the observation from which we make an inference—the evidence for a hypothesis.

As it can be helpful to temporarily reduce the abstraction created by the variables, let us rewrite Bayes' theorem in words:

$$P(\text{unobservable} \mid \text{observable}) = \frac{P(\text{observable} \mid \text{unobservable})P(\text{unobservable})}{P(\text{observable})}. \tag{2.22}$$

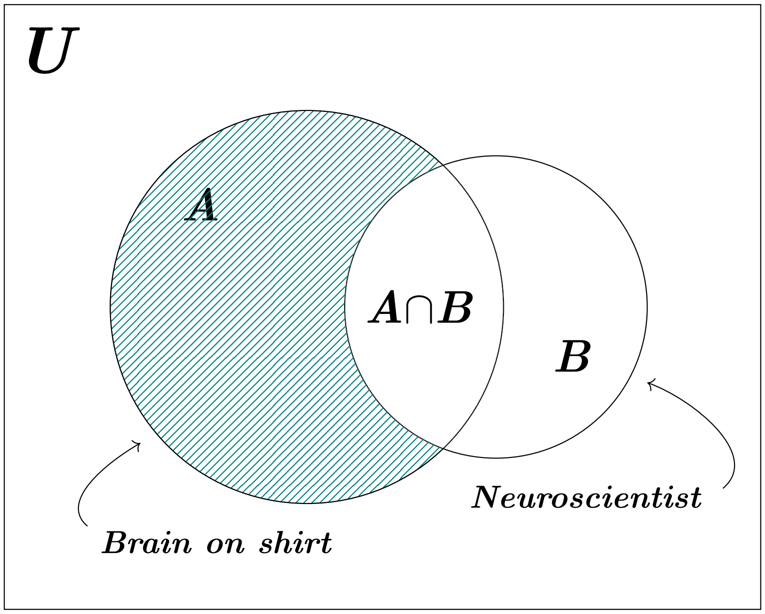
If we wish to give it a stronger explanatory flavour, as required in active inference, we can use the vocabulary of *data* and *hypothesis*:

$$P(\text{hypothesis} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis})P(\text{hypothesis})}{P(\text{data})}. \tag{2.23}$$

In our example, we were given the value of the marginal, $P(A)$. But often, we do not know this value and need to compute it. The number of people who wear a shirt with a brain design printed on it includes two groups, two intersections: the neuroscientists who have a brain on their shirt (A and B) and the non-neuroscientists who have a brain on their shirt (non-B (symbolically: $\widetilde{B}$) and A. (Referring to Table 2.1 as well as Figure 2.4 and Figure 2.6 might be helpful.)

Figure 2.6

Event A and not Event B in the Universe U



Note. A person who wears a shirt with a brain on it and who is not a neuroscientist falls in the coloured section $A \cap \widetilde{B} = A - B$.

Formally, we write the marginal as

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \widetilde{B}) && \text{by def. of marginal} && (2.24) \\ &= P(A \mid B)P(B) + P(A \mid \widetilde{B})P(\widetilde{B}) && \text{by Eq. (2.17)} && (2.25) \end{aligned}$$

Subbing Eq. (2.25) in Bayes' formula (Eq. (2.21)), we get

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \widetilde{B})P(\widetilde{B})}. \tag{2.26}$$

To better understand why we would need to compute the marginal through a sum of the intersections, we will consider another example. But before, let us conclude the neuroscientist scenario with a *surprise* or *Shannon information* computation. Basically, we have two hypotheses: the person is a neuroscientist (B), and the person is not a neuroscientist (not B). Which one minimizes surprise? Let S be the set of events that contain B and $\widetilde{B}$, and s is an event of that set. We write

$$I(S) \triangleq \operatorname*{arg\,min}_{s \in S} h(s \mid A), \tag{2.27}$$

where I is the inference function which takes the set of hypotheses S as arguments and outputs the hypothesis which minimizes the posterior probability given our data (event A, "has a brain on their shirt"). We already calculated the conditional probability for the first hypothesis and obtained 64%. We now compute the probability for the second hypothesis:

$$\begin{aligned}
P(\widetilde{B} \,|\, A) &= \frac{P(A \,|\, \widetilde{B}) P(\widetilde{B})}{P(A)} && \text{Bayes formula} && (2.28)\\
&= \frac{\frac{|\,A \cap \widetilde{B}\,|}{|\,\widetilde{B}\,|} P(\widetilde{B})}{P(A)} && \text{by Eq. (2.16)} && (2.29)\\
&= \frac{\frac{|\,A \cap \widetilde{B}\,|}{|\,\widetilde{B}\,|} (1 - P(B))}{P(A)} && \text{by def. of not in } B && (2.30)\\
&= \frac{\frac{25-16}{100-20}(1 - 0.20)}{0.25} = 0.36
\end{aligned}$$

Therefore, the probability that the person is not a neuroscientist given the brain on their shirt is 36%, which was the probability to expect (100% − 64% = 36%). Let us find the Shannon information of each using Eq. (2.2):

$$\begin{aligned}
h(B \,|\, A) &= \log_2 \left(\frac{1}{P(B \,|\, A)} \right) && (2.31)\\
&= \log_2 \left(\frac{1}{64/100} \right) \approx 0.644 \text{ Sh}
\end{aligned}$$

and

$$
\begin{aligned}
h(\widetilde{B} \mid A) &= \log_2 \left(\frac{1}{P(\widetilde{B} \mid A)} \right) \\
&= \log_2 \left(\frac{1}{36/100} \right) \approx 1.47 \text{ Sh.}
\end{aligned}
\tag{2.32}
$$

Plugging our values in Eq. (2.27), we get

$$
I(S) \triangleq \underset{s \in S}{\arg\min}\; h(s \mid A) = \underset{s \in S}{\arg\min}\; \{0.644,\, 1.47\} = B.
$$

The neuroscientist hypothesis yields a surprise of 0.644 shannon, which is less than the 1.47 shannons of surprise associated with the non-neuroscientist hypothesis. To be clear, we do not know whether the person is a neuroscientist; we simply embrace the hypothesis that is the more likely given the data, which is also the one which has the lowest surprise associated with.

2.2.2 From Data to Hypothesis and From Hypothesis to Data

As noted, sometimes the probability of an observable event is not clear, and we need to work through the computation of the intersections of all the unobservable events that can happen together with the specific observable. In the following example, we reconsider our two types of six-faced die introduced in the previous section. We observe the outcomes of multiple rolls. For simplicity's sake, we suppose that there exist only two types of dice: *fair* six-faced dice with the distribution $P_f(X)$ of Eq. (2.12) and *biased toward the 5* six-faced dice with the distribution $P_b(X)$ of Eq. (2.13). We would like to know whether the die that generates the observed outcomes is fair or not. Our set of observations consists of this list of rolls:

$$
R = \{1, 5, 5, 5, 5, 5\}
$$

The order does not matter; we consider all the rolls together, as if six dice were rolled at the same time and R was the outcome. Although we may not be sure how to use this data set to make an inference concerning the cause of it (i.e., a fair or a biased die), a historically common reflex has been to compute the probability of observing this set of outcomes assuming a fair or biased die (Clayton, 2022). If we recall, this is the reverse or the sampling probability. *Reverse* because we go from the data to the hypothesis; *sampling* because it tells us what observations we should expect to sample. Clayton (2022) makes this distinction clear via the following pictorial presentation:

$$\begin{array}{lr}\text{Sampling probabilities:} & \text{Hypothesis} \longrightarrow \text{Data} \\ \text{Inferential probabilities:} & \text{Data} \longrightarrow \text{Hypothesis}\end{array}$$

One type is no substitute for the other, although we might feel the pull to content ourselves with the sampling one because it is the easiest to find out. Sampling probabilities are based on a *model*. In the present scenario, our model of a fair die tells us that each outcome has an equal probability of 1/6. Thus, we can compute the probability of R given the fair die hypothesis H_f.

Let us do a bit of counting. Each outcome of R has probability 1/6 according to H_f, and the events (rolls) are independent of each other. (Note: the H just refers to "hypothesis" and has nothing to with the entropy function $H(X)$ from the previous section). Therefore,

$$P_{H_f}(R_{ordered}) = (P_f(X = x))^{|R|} = \left(\frac{1}{6}\right)^6 = \frac{1}{6^6} = \frac{1}{46\,656},$$

where $R_{ordered}$ refers to the specific (ordered) sequence $\{1, 5, 5, 5, 5, 5\}$. But because the order of the outcomes does not matter here, we need to count how many ways there are to arrange R, which corresponds to the number of possible permutations of R, written $\sigma(R)$. In the present case, we only need to account for all the positions that 1 can take (e.g.,

$\{1, 5, 5, 5, 5, 5\}$, $\{5, 1, 5, 5, 5, 5\}$, $\{5, 5, 1, 5, 5, 5\}$, etc.). Hence, there are

$$\sigma(R) = C_k^n = C_1^6 = \frac{6!}{(6-1)!1!} = \frac{6!}{5!} = 6$$

possible rolls of six dice that will result in the outcomes of set R. The function C with the upper script n and lower script k reads "n chooses k", meaning that there are n candidates and k place(s); we want to know all the possible ways to fill the k place(s) with the n candidates. Since each of the six permutations has the same probability of 1/46 656, we multiply this probability by 6:

$$P(R \,|\, H_f) = P_{H_f}(R_{unordered}) = 6\left(\frac{1}{46\,656}\right) \approx 0.000129.$$

This probability is admittedly small, but without a comparison, the number is roughly meaningless. We must compute the probability of the rolls R given the competing hypothesis that the die is biased (H_b).

$$P_{H_b}(R_{ordered}) = P_b(X=1)(P_b(X=5))^5 = \left(\frac{1}{30}\right)\left(\frac{5}{6}\right)^5 = \frac{625}{46\,656}.$$

Again, we account for the permutations:

$$P(R \,|\, H_b) = P_{H_b}(R_{unordered}) = 6\left(\frac{625}{46\,656}\right) \approx 0.0804.$$

Now that we have the two reverse probabilities, let us take their ratio to see how much more likely the data R are given the biased-die hypothesis:

$$\frac{P(R \,|\, H_b)}{P(R \,|\, H_f)} = \frac{\frac{3750}{46\,656}}{\frac{6}{46\,656}} = \frac{3750}{6} = 625.$$

Thus, only on the basis of the reverse probabilities, we would expect to observe the sample R 625 times more often with the biased die than with the fair die. It is understandably tempting to infer that the die that generated R is the biased one. However, by doing so, we would *substitute* the sampling probability for the inferential probability. To finalize our Bayesian inference properly, we need to include the prior probability and the marginal.

The prior plays a cardinal role: it forces us to make our assumptions *explicit*. When computing the reverse probability, we implicitly treated the two hypotheses as equally likely. But do we really believe that fair and biased dice are equally common? Whatever our answer, the prior term forces us to spell it out. Neglecting to do so is often referred to as *base-rate* neglect or fallacy. If we never previously encountered a biased die and the die used to generate R was provided by the most honest person we know, we probably would have high confidence in the fairness of the die. We will assume that this is our situation and set the prior for the fair die to 0.999 and the prior for the biased die to 0.001.

What about the marginal? In our neuroscientist example, the marginal was pretty easy since it was given to us—25 people out of 100 had a brain on their shirt. This die example, by comparison, is trickier. What is the probability of observing R? We just computed something close to answering this question; however, we had to assume the nature of the die—namely, fair or biased. A sensible idea would be to weigh the probability of R given H_i, $P(R \mid H_i)$, by the prior probability of H_i, $P(H_i)$. In other words, this means multiplying (weighing) the probability of observing the rolls R given that the die was fair by the probability that the die was fair. And similarly for the biased die hypothesis. Eq. (2.25) suddenly seems more relevant as it accomplishes exactly this task by summing up the intersections of R (observable event) with its various hypotheses (unobservable events). Plugging our numbers in Eq. (2.25), we get a marginal of

$$\begin{aligned} P(R) &= P(R \mid H_f)P(H_f) + P(R \mid R_b)P(H_b) \\ &= \left(\frac{6}{46\,656}\right)\left(\frac{999}{1000}\right) + \left(\frac{3750}{46\,656}\right)\left(\frac{1}{1000}\right) \\ &\approx 0.000209. \end{aligned}$$

Having all the terms, we can at last work out the inferential probabilities. Iterating over our two hypotheses, we get the following results:

$$P(H_f \mid R) = \frac{P(R \mid H_f)P(H_f)}{P(R)} \approx 0.615$$

and

$$P(H_b \mid R) = \frac{P(R \mid H_b)P(H_b)}{P(R)} \approx 0.385.$$

Despite the fact that under the biased-die hypothesis the observed rolls are much more likely than under the fair-die hypothesis, when we factor in our prior beliefs about the two hypotheses, we are nevertheless led to infer that the die is fair.

In *Bernoulli's Fallacy*, mathematician and philosopher of probability Aubrey Clayton (2022) explores in detail the confusion between sampling and inferential probabilities and its consequences. He recounts a particularly telling real-life example of the type of problematic situations that emerge from mixing the two probabilities—an example closely related to our contrived dice example. In the mid-nineteenth century, parapsychologists Samuel Soal (also a mathematician) and Kathleen Goldney teamed up to scientifically establish the existence of so-called extrasensory perception (ESP), demonstrated by Gloria Stewart. They designed a scientifically sound study with card guessing. In the final analysis, Stewart's performance given the chance hypothesis—that the results are due to chance—was approximately 10^{-139}, while her performance given the ESP hypothesis was approximately 0.005, orders of magnitude higher. The Soal-Goldney study received a lot of traction and citations. Many scholars confirmed that the study was carefully designed to prevent fraud. But these impressive results are based on sampling probability. What do we do about all the knowledge we (scientifically) accumulated over time about the world? In other words, what about the priors? And importantly, what about alternative hypotheses?

One such alternative hypothesis is similar to our biased-die hypothesis earlier. Following Clayton, we will call it *fakery*, denoted by the capital letter F. He writes: "The problem with Soal is that no matter how much scrutiny his experiments were placed under, we'd likely assign a prior probability for F that was several orders of magnitude greater than our prior probability for the ESP hypothesis" (p. 90). Assigning a conservative value of 10^{-5} to the probability that there was fakery given the precautions and a value of 10^{-10} to the probability that Stewart had ESP, we get a very different inferential picture: the ESP hypothesis gets an inferential probability of approximately 0.001% while the fakery hypothesis gets an inferential probability of approximately 99.999%. We are nowhere near having to believe in ESP. Priors really play a crucial role, both in the examples covered here and in active inference. (Denouement: it turned out that Soal had indeed doctored the results, so fakery it was.)

2.2.3 The Generative Model in Active Inference

We can see how the languages of Bayesianism and active inference are similar to each other. This is not by accident since active inference builds on the Bayesian brain hypothesis—and in particular its predictive coding formulation—which originated from this apparent compatibility between Bayesian inference and the brain itself (Ackley et al., 1985; Dayan et al., 1995; Friston, 2005; Rao & Ballard, 1999).

We are in a position to better understand why the former is a suitable candidate to provide the formalism of the latter. In active inference, we want to infer the (unobservable) cause or hidden state of (observable) stimuli. To reiterate a fundamental distinction made earlier, the path from the hypothesis to the data allows for sampling probability; it tells us what data we should expect to observe given the assumption that a specific hypothesis is true. This corresponds to the predicted observations of the *generative model* in active inference. However, we perform an inference when we take the path from the data to the hypothesis. Inferences are fundamentally explanatory; they aim to explain the data by identifying its cause. This corresponds to the so-called *inverting* of the generative model (i.e., the inverting of the inverse probability).

Bayesianism offers inference, but at the cost of introducing *subjectivity* into the equation—literally. Specifically, subjectivity enters along with the prior probability. Often pointed out as a flaw, Bayesians argue that, to the contrary, the prior probability is always present but it is hiding, with the pretension of objectivity. At any rate, within active inference, subjectivity is a desired feature, allowing the unique conditions of a species and even more unique experiences and circumstances of an individual to be encoded in the priors and the marginal. We will circle back to this statement once we derive the free energy equations. Furthermore, in section 3.3 (and Appendix 1), we will explore yet another way in which the personal experience of an adaptive agent can modulate its perceptions/inferences.

In active inference, the generative model of an adaptive system is thus defined as the joint probability of the observations o (which results from the hidden and true state $s*$ of the generative process) and its inferred cause s, which we called "hypothesis" above. The hidden causes s are the hypotheses. This joint probability can be factored into two terms—the likelihood and the prior—as we did in Eq. (2.17). Let us rewrite it with the variables we have used so far when characterizing active inference:

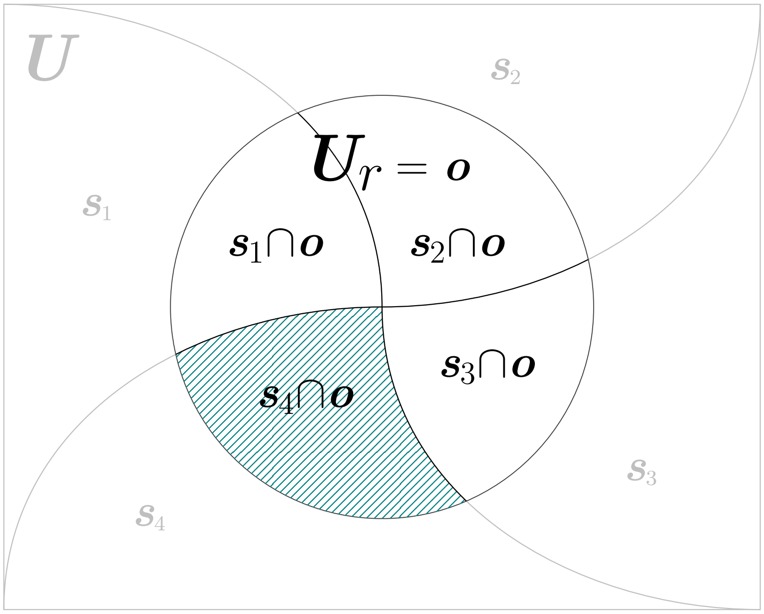
$$P(o,\, s) = P(o \,|\, s)P(s). \tag{2.33}$$

(The comma plays the same role as the symbol $\cap$ used previously, meaning "and".) As before, the joint probability in Eq. (2.33) represents the probability that the data o and the cause s happen together. With only the likelihood and the prior, we can calculate the marginal and then the posterior, as we did in the dice example. These two quantities are therefore obtainable from the generative model.

Unfortunately, a problem arises from the calculation of the marginal. So far, we only explored situations where we had two competing hypotheses or hidden states. But in most situations faced by organisms in the wild, there are many possible hidden states that could have caused an observation. For example, if we see something of a certain size going up, this

Figure 2.7

Reduced Universe of Hidden States s_i Intersecting with Observation o



Note. Once observation y is made, the universe U reduces to o (thus, $U_r = o$). Observation o is partitioned by four hidden states $\{s_1, s_2, s_3, s_4\}$ that are all compatible with it. The coloured area shows the intersection of the particular hidden state s_4 with observation o. (This diagram is inspired by Figure 4.9 of Bolstad and Curran (2016). It is not to scale.)

something could be a bat, a leaf, a ball, or whatever else of a certain size that can go up by any means. In Figure 2.7, we see a Venn diagram depicting an abstract situation where we make an observation o, which is compatible with a few hidden states s_i.

Here, we must generalize Eq. (2.25) to accommodate for an arbitrary number of n hypotheses or hidden states:

$$P(o) = \sum_{i=1}^{n} P(s_i \cap o) = \sum_{i=1}^{n} P(o \,|\, s_i)P(s_i). \tag{2.34}$$

Consequently, the full Bayes' theorem takes the following form:

$$P(s_i \,|\, o) = \frac{P(o \,|\, s_i)P(s_i)}{\sum_{i=1}^{n} P(o \,|\, s_i)P(s_i)}. \tag{2.35}$$

When dealing with continuous probability distributions, the summation becomes an integral. The problem is that this summation or integral is often computationally or analytically intractable. Essentially, this means that the calculation is either too big or too complicated. And without the marginal, we can not compute the posterior probability either. In active inference, the assumption goes that the brain probably does not take this complicated and expensive but exact path, but rather uses approximations. One such approximation is called variational lower bound, one of the variational Bayesian methods for inference, from which the free-energy equations can be derived. By now, we are very close to these derivations.

2.3 Variational Inference and Free Energy

In variational inference, the idea is to approximate the complicated (posterior) conditional probability of latent variables (hidden states s) given observable variables (observations o). To achieve this, we use a family of probability distributions, $\mathcal{Q}_f$, which may contain the exact conditional distribution that we seek to approximate. Crucially, the intra but exact inference is transformed into an optimization problem, where the goal is "[. . .] to find the member of this family, that is, the setting of the parameters, which is closest in KL divergence to the conditional of interest." (Blei, Mucukelbir & McAuliffe, 2017, p. 2) Thus, since we have full control over the distributions $Q(X) \in \mathcal{Q}_f$, we can vary their parameters (e.g., the mean and the standard deviation for a Gaussian distribution) until we obtain a good fit with the complicated and unknown conditional distribution. The quality of the fit is measured by its divergence from the exact conditional—the smaller the divergence, the better the fit.

2.3.1 The Kullback-Leibler Divergence

The distribution with the smallest divergence, referred to as $Q^*(X)$, can then be used as a proxy for the unknown posterior. It is thus the approximate function $Q^*(X)$ that the generative model updates as new observations are collected. Mathematically, this optimization problem takes the following form:

$$Q^*(X) = \underset{Q(X) \in \mathcal{Q}_f}{\arg\min}\, D_{KL}\big[Q(X) \,||\, P(X \,|\, Y)\big] \tag{2.36}$$

where D_{KL} stands for the Kullback-Leibler (KL) divergence, and $P(X \,||\, Y)$ is the *target* conditional probability distribution of hidden states X given observations Y (Blei, Kucukelbir & McAuliffe, 2017).

The KL divergence, also called *relative entropy*, is an information-theoretic measure. It allows us to quantify how much one probability distribution $Q(X)$ diverges from another distribution $P(X)$—it computes the gap between them. This type of function is called a *functional* because it is a function of functions. The arguments it takes are placed in brackets instead of parentheses. Formally, for discrete random variables,

$$\begin{aligned} D_{KL}\big[Q(X) \;||\; P(X)\big] &\triangleq \sum_{x \in X} Q(x) \log \frac{Q(x)}{P(x)} && (2.37) \\ &= \sum_{x \in X} Q(x)\big(\log Q(x) - \log P(x)\big) \qquad \text{by log. quot. rule} && (2.38) \\ &= \mathbb{E}_{Q(X)}\big[\log Q(X) - \log Q(P)\big] \qquad \text{expectation notation} && (2.39) \end{aligned}$$

For continuous random variables (probability density functions), Eq. (2.37) becomes

$$D_{KL}\big[Q(X) \;||\; P(X)\big] \triangleq \int_{-\infty}^{\infty} Q(x) \log \frac{Q(x)}{P(x)} \tag{2.40}$$

Let us note in passing that the KL divergence is not a metric, or true measure of distance, because it does not meet one of the metrics criteria, namely, the symmetry criterion. Thus, in general,

$$D_{KL}\big[Q(X) \mid\mid P(X)\big] \neq D_{KL}\big[P(X) \mid\mid Q(X)\big] \tag{2.41}$$

One way to better see why the KL divergence is also called relative entropy is to rewrite Eq. (2.37) as follows:

$$D_{KL}\big[Q(X) \mid\mid P(X)\big] = \sum_{x \in X} Q(x) \log \frac{1}{P(x)/Q(x)} \tag{2.42}$$

This expression makes salient the inverse probability term. However, the term is a ratio of two probability distributions—one distribution relative to the other. The equation is very close to the entropy equation, as we have the relative surprise of $P(X)$ and $Q(X)$ weighted by $Q(X)$.

To get more familiar with the KL divergence, let us revisit the two examples from the Bayesian inference section. First, we will measure the divergence between our prior probability distribution about the latent/unobservable variable "profession" and our posterior probability distribution on this same hidden state given the observation "has a brain design printed on their shirt." We will write "neuro" to refer to "is a neuroscientist" and "brain" for "has a brain design printed on their shirt." The tilde over a word denotes the negation. Using Eq. (2.38) and plugging in our numbers from the previous section on Bayesian inference, we get:

$$
\begin{aligned}
D_{KL}\big[P(X \,|\, Y) \,||\, P(X)\big] &= \sum_{x \in X} P(x \,|\, y)\big(\log_2 P(x \,|\, y) - \log_2 P(x)\big) \\
&= P(X = neuro \,|\, Y = brain)\big(\log_2 P(X = neuro \,|\, Y = brain) \\
&\quad - \log_2 P(X = neuro)\big) \\
&\quad + P(X = \tilde{neuro} \,|\, Y = brain)\big(\log_2 P(X = \tilde{neuro} \,|\, Y = brain) \\
&\quad - \log_2 P(X = \tilde{neuro})\big) \\
&= (0.64)\big(\log_2(0.64) - \log_2(0.20)\big) \\
&\quad + (0.36)\big(\log_2 P(0.36) - \log_2 P(0.80)\big) \approx 0.737 \text{ Sh.}
\end{aligned}
$$

Therefore, in our scenario, the gap between the prior and the posterior beliefs about the profession of a person is 0.737 shannon. Tangentially, this special type of surprise is called *Bayesian surprise* (Itti & Baldi, 2009). This surprise does not depend on the rarity of an observation, but only on how big a belief update an observation causes. Thus, if we assume that at our 100-person party we know for a fact that German-style pretzels are not sold, not even in the city where the party is taking place, the observation $Y = pretzel$ would certainly produce quite an amount of surprise. However, if we have no reason to believe that neuroscientists have a propensity to eat German-style pretzels, observing a person with one in their hands would not generate Bayesian surprise since $P(X = neuro \,|\, Y = pretzel) = P(X = neuro)$, and similarly for $X = \tilde{neuro}$. Their KL divergence would be zero. Of particular interest, a study by Itti and Baldi (2009) suggests that observations that lead to higher Bayesian surprise attract more human attention than sensory stimuli that lead to a smaller belief update. In principle, we are more likely to look up this paper if we find the result surprising given our prior belief about their finding.

We just saw a simple example of how the KL divergence can be applied to two distributions. To tackle a problem that is a little more complex, we now turn to our two dice distributions from Eqs. (2.12) and (2.13) (shown in Figures 2.3 and 2.8). The fair 6-faced die distribution is uniform—all outcomes have the same 1/6 probability of occurring. If we pick the uniform distribution $Q_1 = Q_u$ from the family of possible distributions $\mathcal{Q}_f$, using Eq. (2.37) we get a divergence

$$\begin{aligned} D_{KL}\big[Q_u(X)\,||\,P_f(X)\big] &= \sum_{x\epsilon X} Q_u(x)\log_2\frac{Q_u(x)}{P_f(x)} \\ &= Q_u(X=1)\log_2\frac{Q_u(X=1)}{P_f(X=1)} + \cdots \\ &\quad + Q_u(X=6)\log_2\frac{Q_u(X=6)}{P_f(X=6)} \\ &= \frac{1}{6}\log_2\frac{1/6}{1/6} + \cdots + \frac{1}{6}\log_2\frac{1/6}{1/6} = 6\left(\frac{1}{6}\log_2(1)\right) = 0 \text{ Sh.} \end{aligned}$$

As expected, the divergence is zero because $Q_u = P_f$; a distribution does not diverge from itself. Let us see what happens if we take the divergence between Q_u and a die that follows the biased distribution P_b from Eq. (2.13), which was $P_b(X) = \{1/30, 1/30, 1/30, 1/30, 5/6, 1/30\}$.

$$\begin{aligned} D_{KL}\big[Q_u(X)||P_b(X)\big] &= \sum_{x\epsilon X} Q_u(x)\log_2\frac{Q_u(x)}{P_b(x)} \\ &= Q_u(X=1)\log_2\frac{Q_u(X=1)}{P_b(X=1)} + \cdots \\ &\quad + Q_u(X=5)\log_2\frac{Q_u(X=5)}{P_b(X=5)} + Q_u(X=6)\log_2\frac{Q_u(X=6)}{P_b(X=6)} \\ &= 5\left(\frac{1}{6}\log_2\frac{1/6}{1/30}\right) + \frac{1}{6}\log_2\frac{1/6}{5/6} \approx 1.073 \text{ Sh.} \end{aligned}$$

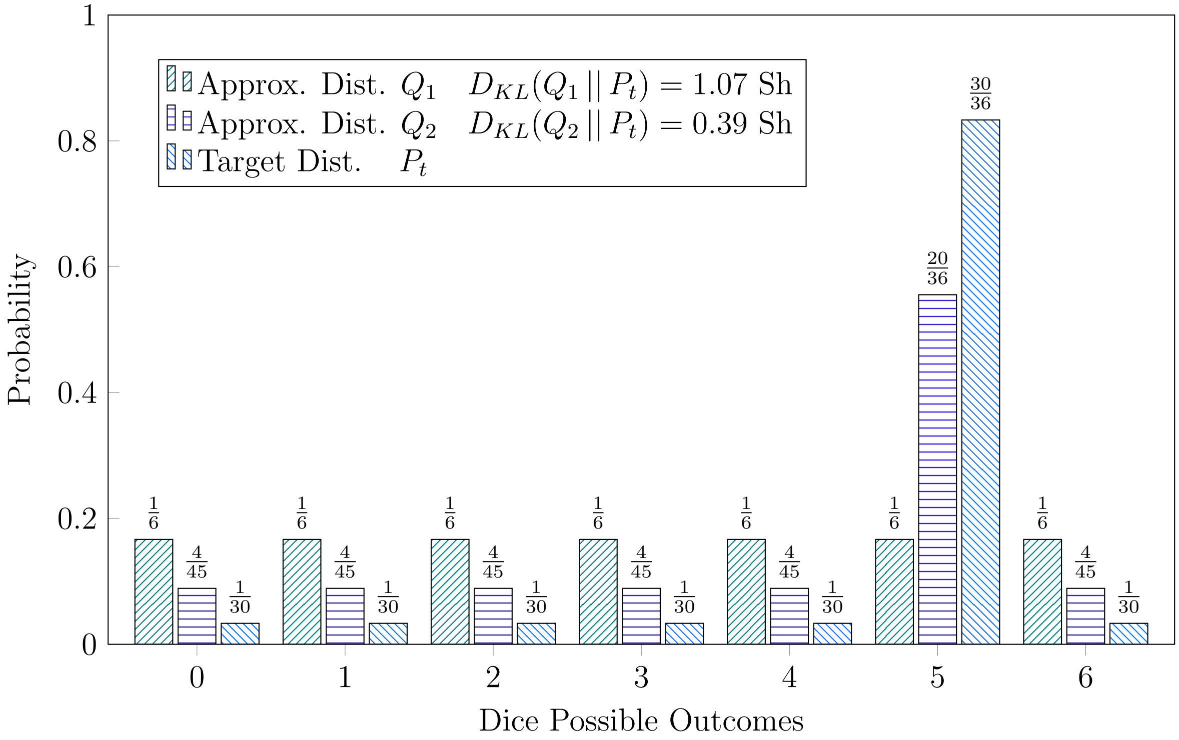
The divergence is not null anymore as we have two different distributions. The ensuing optimization problem would require varying the distribution of Q_1 to generate other distributions from the family $\mathcal{Q}_f$. For instance, the following Q_2 distributions is obtainable from varying the uniform distribution Q_1:

$$Q_2 = \left\{ \frac{4}{45}, \frac{4}{45}, \frac{4}{45}, \frac{4}{45}, \frac{20}{36}, \frac{4}{45} \right\} \tag{2.43}$$

The target distribution $P_t = P_b$ and the approximate distributions Q_1 and Q_2 are shown in Figure 2.8.

Figure 2.8

The Target Probability Distribution $P_t(X)$ and Two $Q(X)$ Distributions from the $\mathcal{Q}_f$ Family.



Note. The target probability distribution P_t is shown in blue. We can see that the violet distribution is closer to the target distribution then is the green (and uniform) distribution. Accordingly, the KL divergence from Q_2 to P_t is the smallest of the two.

We now have $\{Q_1, Q_2\} \in \mathcal{Q}_f$.We find that $D_{KL}[Q_2(X) \,||\, P_t(X)] = 0.390$ Sh, which is a smaller amount of information than the one between Q_1 and P_t. Therefore, if we only consider Q_1 and Q_2, we would use Q_2 to approximate P_f, and Eq. (2.36) would give

$$Q^*(X) = \underset{Q(X)\in\mathcal{Q}_f}{\arg\min}\, D_{KL}\big[Q(X)\,||\,P(X\,|\,Y)\big] = \underset{Q(X)\in\mathcal{Q}_f}{\arg\min}\big\{1.07,\,0.390\big\} = Q_2(X) \tag{2.44}$$

2.3.2 Free Energy as Inherent and Excess Surprise

An insightful way to interpret the KL divergence is in terms of (expected) *excess surprise* that results from approximating the true distribution P by the approximate distribution Q. Indeed Q serves as a model for P and as long as the model makes predictions that differ from the actual conditional P, an amount of surprise is produced. This surprise is in excess since it comes from a suboptimal model.

Let us consider a game of dice that requires a fair six-sided die to work properly. If we play that game with the die P_b biased toward the outcome 5, eventually, an optimal Bayesian inferential process would lead to inferring that the die used in the game is P_b. The distribution P_b becomes our target distribution P_t that we want to approximate. If we keep working with the uniform distribution Q_1 as a model for the die, we will not be able to anticipate the biased nature of the die and its outcomes. Our belief about the die will diverge from P_t. As we see more of the die rolls, we should update our model by varying the parameter of the distribution Q; we might eventually use Q_2 instead of Q_1. This new model would reduce the amount of prediction errors. Hence, as our model improves, our perception improves as well; we anticipate the world better. This improvement of the model can be measured via the KL divergence.

Importantly, the divergence is between our approximate conditional Q and the exact Bayesian conditional P, not directly between our model Q and the generative process, which is the die itself in our example. Still, our updated model about the die is presumably more representative of the die itself that was our previous model with higher divergence from the exact Bayesian conditional. This is because the true Bayesian conditional makes optimal inferences about the die and so we are approximating a very efficient inferential process.

Ultimately, the point is that as the divergence reduces, we fare better in the game; our chances of winning/surviving improve. However, the die is still biased and the game requires a fair one. Having our model in sync with the biased die is certainly useful, but it does not change the fact that we play with a biased die. In the language of active inference, accurate perception is not enough; we also need action.

It is through action that we can change the hidden states of the world that give rise to unacceptable, high-surprise observations. Through action, we can substitute the biased die for a fair one. But if we have not identified correctly what causes the surprising observations, we can not easily change them. In our game of dice, the high-surprise observations are the rolls (the 5 always come up). If we haven't identified the nature of the die as the most likely cause of these observations, but instead believe that the culprit is the table, we might try to change this table. But to no avail, the unacceptable observations will keep coming as our action on the world did not target the correct cause.

It is not by chance that the KL divergence maps onto perception and surprise maps onto action; these two terms constitute in fact variational free energy:

$$\mathcal{F}\big[Q,o\big] \triangleq \underbrace{D_{KL}\big[Q(s\,|\,o)\,||\,P(s\,|\,o)\big]}_{divergence} - \underbrace{\log P(o)}_{log\ evidence} \tag{2.45}$$

$$= \underbrace{\log\frac{1}{P(o)}}_{\substack{inherent\\ surprise}} + \underbrace{D_{KL}\big[Q(s\,|\,o)\,||\,P(s\,|\,o)\big]}_{excess\ surprise} \tag{2.46}$$

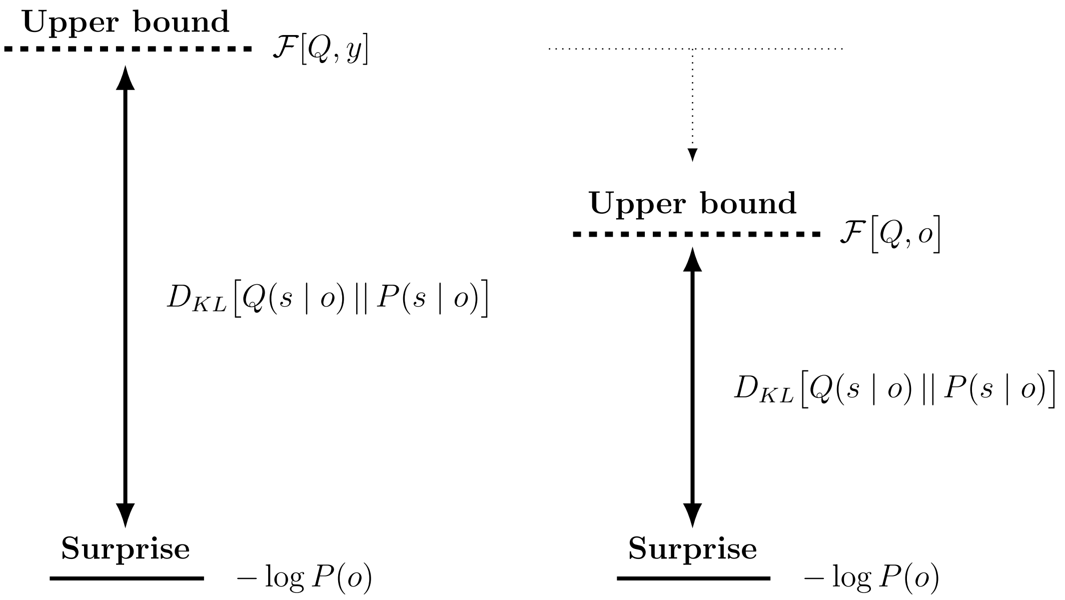
The first equation is the orthodox form of the Divergence-Surprise formulation of variational free energy—surprise is often called negative log evidence because it is the logarithm of the probability of an observation which is the evidence. The second equation is the way I like to write it. It is mostly a labelling difference, but I think the words "inherent" and "excess" naturally fit in the lexicon of active inference and capture rather well the dynamics of the equation. As aforementioned, the divergence term quantifies the amount of surprise resulting from the gap between the exact Bayesian posterior distribution and the approximate distribution. This quantity is an excess in the sense that it is due to a bad model. Once the optimization problem of finding the best approximate distribution Q^* is found, the excess surprise is minimal (see Figure 2.9). And if Q^* happens to be the exact posterior distribution, there is then no excess surprise at all and the variational free energy equals the surprise term. Thus, the inequality

$$\mathcal{F}[Q, o] \geq \log \frac{1}{P(o)} \tag{2.47}$$

holds, meaning that variational free energy is an *upper bound* on surprise—it is always greater or equal to it. When we minimize free energy, we minimize this upper bound and indirectly minimize surprise.

Figure 2.9

Variational Free Energy as an Upper Bound on Surprise


Upper bound
$\mathcal{F}[Q, y]$
$D_{KL}\big[Q(s \mid o) \,||\, P(s \mid o)\big]$
Surprise
$-\log P(o)$
Upper bound
$\mathcal{F}\big[Q, o\big]$
$D_{KL}\big[Q(s \mid o) \,||\, P(s \mid o)\big]$
Surprise
$-\log P(o)$

Note. Diagram taken from Figure 2.4 of Parr, Pezzulo and Friston (2022) with the notation adapted to this text.

When the *perception mechanism* of an adaptive system is good, the divergence or excess surprise is minimal. This leaves us with the log evidence term. Why call it *inherent surprise*? Because, as discussed in section 1.5, according to active inference, the conditions of existence of an organism are encoded in this probability. The by-now classic image used by Friton (2010) is the fish's expectation to be surrounded by water. The observation "surrounded by water" is immensely important for the fish to stay alive, and thus the probability of this observation of proportionally set high. In turn, the surprise of observing an environment where it is false that the fish is surrounded by water is extremely high. To stay alive, the fish will act on the world to stay in non-surprising states of being surrounded by water. It is in that sense that the surprise that an observation creates in an organism depends on the organism's specific conditions of existence; the surprise is thus *inherent*. This is also why the surprise can remain high even once the divergence is closed and no excess surprise is left, but the *action mechanism* can minimize this second term. Through action, the fish can do its best to stay in water. But of course, it needs perception to infer the hidden state that causes its observations about the lack of water. (The analysis applied here to the fish is equally valid for the adaptive snowflake, the body temperature, and the glucose level examples from section 1.5.2)

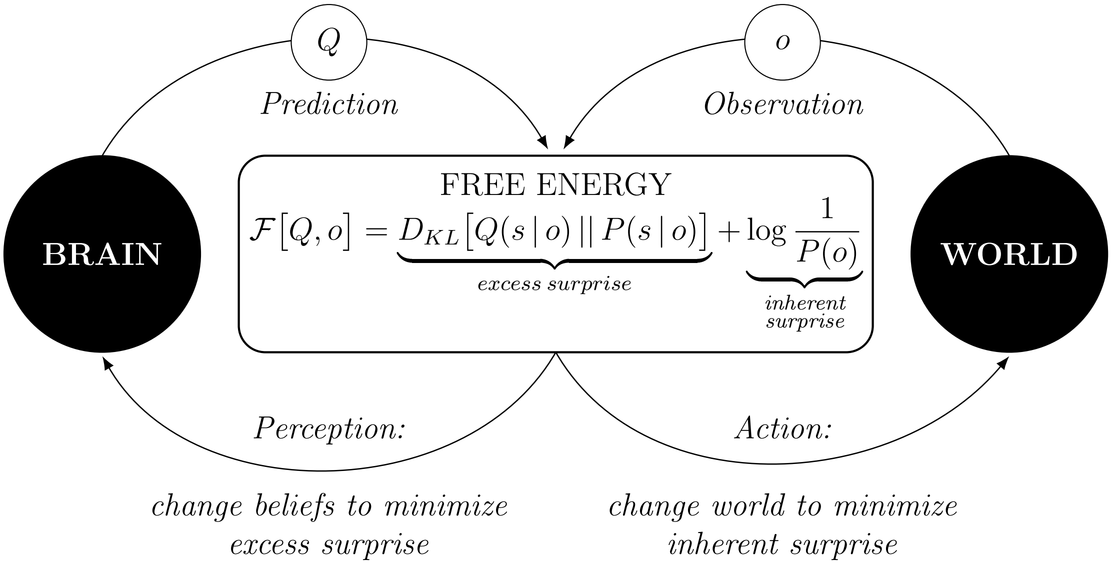
In Eqs. (2.45) and (2.46), we can clearly see how active inference and the free-energy principle unite perception and action, as both minimize the same quantity. By minimizing variational free energy, an adaptive system is led to navigate its environment from low surprise states to low surprise states (e.g., surrounded-by-water states for the fish), which is to say in states conducive to its existence; these low *information* entropy states keep its body in low *physic* entropy states.

2.3.3 Derivations of the Free Energy Equations

Although very elegant on conceptual grounds, the Inherent-Excess Surprise formulation of variational free energy has one obvious problem: it contains the term that we were trying to approximate in the first place, namely, the exact posterior probability. Variational inference works only once we have an expression that only contains terms that we can compute. In what follows, we will derive two other equivalent expressions, which will be computable. But

Figure 2.10

Free Energy Minimization Through the Perception and Action Loops


Q
Prediction
o
Observation
BRAIN
FREE ENERGY
$\mathcal{F}[Q,o] = \underbrace{D_{KL}[Q(s\,|\,o)\,||\,P(s\,|\,o)]}_{\textit{excess surprise}} + \underbrace{\log\frac{1}{P(o)}}_{\textit{inherent surprise}}$
WORLD
Perception:
Action:
change beliefs to minimize excess surprise
change world to minimize inherent surprise

Note. Mathematical version of Figure 1.2. Through perception, the brain can change its model (beliefs) about the world and minimize the excess surprise resulting from the clash between the brain prediction (Q) and the observation from the world (o). Through action, the brain can minimize inherent surprise by changing the hidden state of the world that produces the surprising observation o (This figure is adapted from Figure 2.5 of Parr, Pezzulo and Friston (2022))

beforehand, we still need to derive the Inherent-Excess Surprise formulation from the KL divergence.

The Inherent-Excess Surprise Formulation

We will take the path of the Evidence Lower Bound (ELBO) derivation based on Blei, Mucukelbir and McAuliffe (2017) and Goodfellow, Bengio and Courville (2016). As a *lower bound*, ELBO is the negative variational free energy; by multiplying our final result for ELBO by minus one, we will obtain our desired *upper bound* on surprise. The other and more common way to derive the variational free energy equations in the active inference literature is via Jensen's inequality; these derivations are included in Appendix 1. But I prefer the ELBO approach. The notation is the one used through out the text when characterizing active inference.

We begin by writing the KL divergence between the variational distribution Q over the random variables of hidden states $s \in S$ and the true posterior distribution over $s \in S$ given the random variables of observations $o \in O$:

$$D_{KL}\big[Q(s\,|\,o)\,||\,P(s\,|\,o)\big] = \mathbb{E}_{Q(s\,|\,o)}\left[\log\frac{Q(s\,|\,o)}{P(s\,|\,o)}\right]. \tag{2.48}$$

Next, we rewrite the denominator using Eq. (2.15):

$$D_{KL}\big[Q(s\,|\,o)\,||\,P(s\,|\,o)\big] = \mathbb{E}_{Q(s\,|\,o)}\left[\log\frac{Q(s\,|\,o)}{\frac{P(s,o)}{P(o)}}\right]. \tag{2.49}$$

We now have our intractable marginal or evidence term $P(o)$ explicitly showing up in the expression. Moreover, our generative model $P(s,o)$ is also rendered salient in this equation. Applying the logarithm quotient rule twice, Eq. (2.49) becomes

$$\mathbb{E}_{Q(s\,|\,o)}\left[\log\frac{Q(s\,|\,o)}{\frac{P(s,o)}{P(o)}}\right] = \mathbb{E}_{Q(s\,|\,o)}\left[\log Q(s\,|\,o) - \log\frac{P(s,o)}{P(o)}\right] \tag{2.50}$$

$$= \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o) - \big(\log P(s,o) - \log P(o)\big)\big] \tag{2.51}$$

$$= \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o) - \log P(s,o) + \log P(o)\big]. \tag{2.52}$$

We can distribute the expectation over each term of the expression for the KL divergence, such that

$$\begin{aligned} D_{KL}\big[Q(s\,|\,o)\,||\,P(s\,|\,o)\big] = \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] - \mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big] \\ + \mathbb{E}_{Q(s\,|\,o)}\big[\log P(o)\big]. \end{aligned} \tag{2.53}$$

If we try to compute the value of the last term we realize that the expectation goes away leaving only $\log P(o)$. This is because the expectation is over the variable s, on which $P(o)$

does not depend. Hence,

$$\begin{aligned}
\mathbb{E}_{Q(s\,|\,o)}\big[\log P(o)\big] &= \sum_{i=1}^{n} Q(s_i\,|\,o)\log P(o) && \text{by def. of expectation} \quad (2.54)\\
&= Q(s_1\,|\,o)\log P(o) + Q(s_2\,|\,o)\log P(o) + \\
&\quad \ldots + Q(s_n\,|\,o)\log P(o) && \text{by def. of summation} \quad (2.55)\\
&= \log P(o)\big(Q(s_1\,|\,o) + Q(s_2\,|\,o) + \\
&\quad \ldots + Q(s_n\,|\,o)\big) && \text{by factoring } \log P(o) \quad (2.56)\\
&= \log P(o)(1) = \log P(o) && \text{by Eq. (2.14)} \quad (2.57)
\end{aligned}$$

where n is the size of the set S, that is, $n = |S|$. Based on this result, Eq. (2.53) simplifies to

$$D_{KL}\big[Q(s\,|\,o)\,||\,P(s\,|\,o)\big] = \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] - \mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big] + \log P(o). \tag{2.58}$$

Rearranging the terms, we get

$$\mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big] - \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] = \log P(o) - D_{KL}\big[Q(s\,|\,o)\,||\,P(s\,|\,o)\big], \tag{2.59}$$

where we define the evidence lower bound (ELBO) $\mathcal{L}[Q,o]$ as

$$\underbrace{\mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big] - \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big]}_{\textit{Evidence Lower BOund (ELBO)}} \triangleq \mathcal{L}[s,Q] \tag{2.60}$$

$$\mathcal{L}[Q,o] \triangleq \underbrace{\log P(o)}_{\textit{Log Evidence}} - \underbrace{D_{KL}\big[Q(s\,|\,o)\,||\,P(s\,|\,o)\big]}_{\textit{Divergence}}, \tag{2.61}$$

$\mathcal{L}[Q,o]$ is a lower bound on evidence/surprise because the KL divergence is defined to be greater than or equal to zero. Therefore, $\mathcal{L}[Q,o]$ is either equal or smaller to the log evidence term; it is equal when the divergence is zero and smaller when the divergence is greater than zero. Hence,

$$\mathcal{L}[Q,o] \leq \log P(o) \tag{2.62}$$

But crucially, the two terms labelled ELBO on the left-hand side of Eq. (2.60) are completely composed of *knowable* and *computable* values. This is the *tour de force* that variational inference achieves. We are now in a position to approximate the posterior probability distribution $P(s \,|\, o)$.

The final step in our derivation of variational free energy is to flip this lower bound into an upper bound; we do it by multiplying Eq. (2.59) by minus 1 (i.e., negating the equation) and applying Eqs. (2.2)-(2.4) on the log evidence term (Eq. 2.66):

$$-\mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big] + \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] = -\log P(o) + D_{KL}\big[Q(s\,|\,o) \,||\, P(s\,|\,o)\big]. \tag{2.63}$$

(Therefore, the KL divergence can also be expressed as

$$D_{KL}\big[Q(s\,|\,o) \,||\, P(s\,|\,o)\big] = -\mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big] + \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] + \log P(o). \tag{2.64}$$

This definition for the KL divergence is used in Appendix 1). Eq. (2.63) yields the definition of variational free energy $\mathcal{F}\,[Q,o]$ as

$$\underbrace{\mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] - \mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big]}_{Variational\ Free\ Energy} \triangleq \mathcal{F}[Q,o] \tag{2.65}$$

$$\mathcal{F}[Q,o] \triangleq \underbrace{\log\frac{1}{P(o)}}_{\substack{Inherent\\ Surprise}} + \underbrace{D_{KL}\big[Q(s) \,||\, P(s|o)\big]}_{Excess\ Surprise} \tag{2.66}$$

This last equation is in fact the amply discussed Eq. (2.46), the Inherent-Excess Surprise formulation of free energy.

Let us now turn to the left-hand side of Eq. (2.65).

The Energy-Entropy Formulation

A closer look at the right-hand side of Eq. (2.65) might lead a keen observer to notice the slightly hidden presence of Shannon entropy. Indeed, by applying the logarithm quotient rule and the fact that the logarithm of 1 is 0 (see Eqs. (2.2)-(2.4)), we can extract the negative of the information entropy equation (Eq. (2.10)):

$$\mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] = -\mathbb{E}_{Q(s\,|\,o)}\left[\log \frac{1}{Q(s\,|\,o)}\right] = -H\big(Q(s\,|\,o)\big) \tag{2.67}$$

The other term is called *Energy.* Hence, we can give the variational free energy from Eq. (2.65) its canonical form in the fields of machine learning and statistics:

$$\mathcal{F}[Q,o] \triangleq -\underbrace{\mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big]}_{Energy} - \underbrace{H\big(Q(s\,|\,o)\big)}_{Entropy}. \tag{2.68}$$

This expression requires knowing the variational distribution $Q(s\,|\,o)$ and the generative model $P(s,o)$, to which we have by definition full access.

The energy term is imported from statistical physics (see Parr, Pezzulo and Friston (2022), p. 28). If the logarithm absorbs the negative sign, we clearly see that the Shannon information (Eq. (2.2)) of the generative model is calculated. As Friston (2010) explains, "[...] the energy is the surprise about the joint occurrence of sensations and their perceived causes [...]" (p. 129). By weighing the surprise of the generative model by the approximate distribution Q, we effectively encourage Q to assign a high probability to the states that are most likely under the generative model. States with lower surprise should be lower in energy too. The Entropy term, for its part, indicates that it is best, when lacking data, to hold weak beliefs about the world, which means to assign an equal amount of belief in the hidden states (Friston, 2010; Parr, Pezzulo & Friston, 2022). To have a spread out (high in uncertainty) approximate distribution Q minimizes the risk of overconfidence in some causes. If we refer back to the game of dice example, it would mean embracing the belief in the uniform distribution prior to collecting observations in the form of die rolls.

The Complexity-Accuracy Formulation

To derive our last expression, we start from Eq. (2.65) and express the generative model $P(s, o)$ as $P(o \,|\, s)P(s)$:

$$\begin{aligned} \mathcal{F}\big[Q, o\big] &= \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] - \mathbb{E}_{Q(s\,|\,o)}\big[\log P(s, o)\big] \\ &= \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] - \mathbb{E}_{Q(s\,|\,o)}\big[\log P(o\,|\,s)P(s)\big]. \end{aligned} \tag{2.69}$$

We then factor out the expectation and apply the logarithm multiplication rule:

$$\mathcal{F}\big[Q, o\big] = \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o) - \log P(o\,|\,s)P(s)\big] \tag{2.70}$$

$$= \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o) - \big(\log P(o\,|\,s) + P(s)\big)\big] \tag{2.71}$$

$$= \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o) - P(s) - \log P(o\,|\,s)\big]. \tag{2.72}$$

We can redistribute the expectation in a way that allows for the Complexity term to emerge by applying the definition of the KL divergence (Eq. (2.39)); the remaining term is called "Accuracy."

$$\mathcal{F}\big[Q, o\big] = \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o) - \log P(s)\big] - \mathbb{E}_{Q(s\,|\,o)}\big[\log P(o\,|\,s)\big] \tag{2.73}$$

$$\triangleq \underbrace{D_{KL}\big[Q(s\,|\,o)\,||\,P(s)\big]}_{Complexity} - \underbrace{\mathbb{E}_{Q(s\,|\,o)}\big[\log P(o\,|\,s)\big]}_{Accuracy}. \tag{2.74}$$

Again, we have derived yet a second formulation of variational free energy that is composed uniquely of known values. In this Complexity-Accuracy version, we use the variational distribution Q over which we have full control, and the sampling and prior probability which belong to our generative model.

The Complexity-Accuracy formulation lends itself to a compelling interpretation. It is sometimes described—although a little misleadingly—as Occam's razor stated mathematically. Occam's razor, adapted to the cognitive science context, is about explaining as accurately as possible the sensory data given the simplest (i.e., least complex) explanation for it. The Complexity term takes the divergence between the approximate posterior probability and the generative model prior probability. It thus measures the Bayesian surprise discussed earlier. Parr, Pezzulo and Friston (2022) write that this divergence "[...] scores the degree to which we must depart from our prior beliefs about the world in order to explain data" (p. 268). However, if the prior beliefs about the world are inaccurate and the model needs to make updates, keeping this divergence term small generates an incentive for inaccuracy. And this is exactly where the Accuracy term counterbalances this incentive by making inaccuracy equally costly. In Eq. (2.74), we can identify the sampling probability $P(o \,|\, s)$ that tells us how likely an observation is given the hypothesized explanation s. If we recall, sampling probabilities go from a hypothesis to the data. It is the inferential probabilities that go from the data to a hypothesis. And it is exactly with respect to this inferential probability in the form of the approximate posterior distribution Q that the sampling probabilities of the generative model are weighted. The logarithm of the inverse of the sampling probability constitutes the surprise about the incoming sensations as expected by the generative model. Hence, the Accuracy term prompts the approximate distribution Q to assign a higher probability to causes that tend to explain the sensory input well, that make them least surprising (Friston, 2010; Parr, Pezzulo & Friston, 2022).

Let us finally note that the Complexity-Accuracy formulation of variational free energy is generally the one favoured in predictive coding (Salvatori et al., 2023) as well as by Ororbia and Kelly (2023) in their specification of the neural generative coding units of the CogNGen (Kelly & Ororbia, 2023).

Chapter 3

Active Inference and Affective Phenomena

The relationship between biology and mathematics is a thorny one. As Lindsay (2021) observes, biology is among the slowest branches of science to be mathematized. Mathematics is viewed as too simple to capture all the biological diversity and complexity. So many indispensable details can be lost in a mathematical model, to the point where the model becomes useless if not detrimental. Lindsay nicely captures the tension: “Oversimplification and an obsession with aesthetics are legitimate pitfalls to avoid when applying mathematics to the real world. Yet, at the same time, the richness and complexity of biology is exactly why it needs mathematics” (p. 11).

Given the extent to which modern physics relies heavily on equations, it is hard to believe that in the first half of the 19th century, Ohm (from Ohm’s law) was met with a lot of resistance from some of his contemporary fellow physicists for his attempt to describe mathematically how electricity behaves in different mediums (Lindsay, 2021). Electricity, and most of the phenomena studied in physics, are admittedly much simpler than what we encounter in the biological world, where so many parts interact together. Nevertheless, since the 1950s, mathematics has made many breakthroughs in biology, including in neuroscience. But despite the recent successes of this enterprise, many brain functions pose particularly difficult challenges to mathematically minded scientists, apparently eluding formal characterizations; among them, we find emotions (Seth & Friston, 2016). The expression “mathematization of emotions” itself may strike as oxymoronic. However, recent promising attempts to provide a formal account for affective phenomena have been made, and some come from predictive coding and active inference.

In this section, we will begin with a characterization of affective phenomena that involves notions such as homeostasis and allostasis. Then, we will explore how affects fit within the theory of active inference, by applying concepts from the previous sections. We will continue with a quick exploration of the theorization of researchers who enquired about emotions through the lens of predictive coding and active inference. After discussing how affective phenomena contribute to the minimization of free energy, we will examine how the cortex improves the predictions made by the limbic system, hence further minimizing free energy.

3.1 Characterization of Affective Phenomena

We will base our characterization of affective phenomena on the capstone paper (Schiller et al., 2024) of the Human Affectome Project. With the hope of creating cohesion in the effervescent field of affective sciences, a team of 173 affective researchers joined forces to create a framework that would encompass all aspects of human affective phenomena (hence the suffix “-ome”). The Human Affectome is a framework that brings together a basic set of common assumptions about why (human) affective phenomena exist and what they are. The project aims to act as scaffolding from which existing theories can be reformulated and new theories developed, as well as to promote the emergence of cohesion in affective sciences. As the name suggests, the Human Affectome is about humans. However, the authors believe that the core of the framework also applies to non-human animals. In the present text, we will try to be as inclusive as possible, in the sense that we will consider organisms with a capacity for affective phenomena in general. Moreover, our goal here is not to provide a complete summary of the Human Affectome framework, but rather to borrow only the key components that will be most helpful in our attempt to understand affective phenomena from the perspective of active inference.

3.1.1 Why Are There Affective Phenonomena

According to the Human Affectome, affective phenomena exist to ensure the *viability* of an organism. A living organism is in a constant process of self-recreating its components so that it can maintain its integrity while self-distinguishing from its environment. This process, known as autopoiesis, thus supposes an interaction between an organism and its environment. It also contains the notion of homeostasis, a process (discussed in previous sections) by which an organism persists against the fluctuations of its environment, maintaining its parameters around set points that are viable (i.e., conducive to its autopoiesis). To successfully interact with its environment, an organism must learn which objects are of particular relevance to its viability. But it does not only place its awareness on those objects and passively collect sensory input from them, it also acts on its environment to make it more favourable to its own conditions of existence. Living organisms are said to *enact* their relevance. Affective phenomena are assumed to help achieve this enaction of relevance in order to ensure viability. This is the simplified core of the teleological principle postulated in the Human Affectome, the why of affects.

3.1.2 What Are Affective Phenonomena

Affective phenomena are both experiential (i.e., felt) and mechanistic (i.e., underpinned by biological processes/algorithms). The Human Affectome distinguishes between affective concerns and affective features.

Affective Concerns

Affective concerns are what affects are about, while *affective features* can be regarded as evaluative feedback indicating how well an organism fares with respect to its affective concerns. Affective concerns are defined as processes that reflect the relevance of objects to an organism's viability. Here, we use the term "object" to include physical and mental objects—which include, for instance, other organisms and thoughts—and situations. The relevance manifests itself in what is called the felt implications of an object. Since affective phenomena are geared toward action, the potential for being on the receiving end of an object (to be acted upon), or for acting on an object is central to the process. Moreover, the concerns vary in timescale, ranging from proximal to distal. That is, the relevance of an object can be immediate (proximal) or further in the future (distal). For instance, if we are having a picnic and (1) a violent storm breaks out, the relevance of the storm can be thought of as immediate. But even more immediate would be (2) a bleeding arm, after having been struck by a broken branch. By comparison, if (3) the weather forecast announces a violent storm in three hours, the relevance of the storm is more distant.

Affective concerns can be classified into various categories, each one addressing a type of breach into an organism's viability. The Human Affectome divides them into two main categories: physiological concerns and operational concerns.

Physiological concerns refer to a physiological imbalance—a departure from the organismic comfort zone—that requires immediate action. These are thus directly linked to homeostasis. They arise from a type of sensation called interoceptive. An interoceptive sensation is caused by the body itself. It does not come from the outside world. This is the domain of feeling thirsty or hungry, feeling cold or hot, feeling sleepy or sick, feeling different kinds of pain.

Operational concerns, contrary to physiological concerns, do not involve a departure from homeostatic states, but addressing them might prevent such a departure in the future. They are called "operational" because they require the organism to take a series of actions, to operate, in a manner that will increase the chance of ensuring *future* viability. For instance, scenarios (1) and (3) of the picnic-and-storm example fall within the operational category of affective concerns, while scenario (2) falls within the physiological one.

The operational concerns can be further grouped into various types of concerns. To give a few examples, we have *obstruction* concerns, where an object is in the way of an organism's adaptive capacity; examples include processes such as anger, frustration, or annoyance. There are *threat* concerns caused by sources of danger threatening one's viability; examples include affective processes such as fear, worry, or dread. There are *safety* concerns, encouraging an organism to navigate safe environments that allow, for instance, the safe simulation and practice of actions relevant to the organism in the form of play; safety concerns include joy, happiness and exhilaration. There are *epistemic* concerns, encouraging the acquisition of new knowledge; examples include curiosity, intrigue, and fascination. The Human Affectome also lists the following additional operational concerns among its examples: *loss* concerns (e.g., disappointment, sadness, grief), *cooperation* concerns (e.g., care, love, belonging, trust, empathy), *moral* concerns (e.g., pride, admiration, shame, moral disgust), and *aesthetic* concerns (e.g., awe, appreciation, beauty). These operational concerns, as the authors observe, map rather well onto what is usually called emotions.

Affective Features

The experiential facet of affective phenomena is located in the affective features. The distinction between affective concerns and features made by the authors of the Human Affectome is particularly clear in this passage: "Unlike affective concerns, which highlight the object of interest as actionably relevant during an affective experience, affective features are reflected in the qualitative aspects of the experience itself" (p. 13) These features are valence and arousal. *Valence processes* offer a felt source of information on the homeostasis performance of an organism. Valence is a metric of evaluation (goodness and badness of a state). In common terms, it is about pleasure and displeasure. *Arousal processes*, on the other hand, form a metric of activation (high or low), providing information about the intensity of the affective experience, whether positive or negative; it is the strength of the affective signal. Concretely, arousal is concerned with the level of resources needed by various systems to deal with affective concerns. Together, valence and arousal compose the evaluative feedback on how well an organism deals with affective concerns; they are the affective gauges of the affective concerns. It is important to point out that affective features are commonly defined to be affects while here they are aspects of affective phenomena.

3.1.3 Allostasis

We conclude our simplified tour of the Human Affectome—as adapted to non-human animals—with the notion of *allostasis*. According to Sapolsky (2004), allostasis can be regarded as a modernization of the concept of homeostasis, where the former subsumes the latter. It was first advanced by Sterling and Eyer (1988) to better account for the way biological organisms regulate themselves. Homeostasis is all about a collection of ideal set points for various physiological parameters. But these set points are not the same depending on what an organism is doing. For instance, the amount of blood bringing oxygen to specific muscles is not the same when an organism is at rest or when it is running. The set points constantly fluctuate to accommodate the needs of an organism doing different activities. The term "allostasis" captures this idea in its etymology, denoting something like "stability through change." The second aspect of regulation highlighted by allostasis is that there is a myriad of ways to regulate the set points, not just one local mechanism. Sapolsky (2004) gives the following water shortage example. If the body is running low on water, an animal's kidneys can certainly reduce the production of urine to conserve more water (homeostatic solution). But the brain can also take matters in hand and not only send the kidneys the instruction to conserve more water, but also orchestrate other water-saving measures like transporting water away from body regions with a high evaporation rate, and water-acquiring measures like making the animal thirsty. Sapolsky writes: "Allostasis is about the brain coordinating body-wide changes, often including changes in behavior" (p. 9) More recently, Sterling (2012, 2019) has argued convincingly for the necessity of the allostatic predictive regulation and further explained and characterized the process as orchestrated by the brain—"an organ for predictive regulation."

The characteristic of allostasis most emphasized in the affective literature is the capacity to work in anticipation to a set point deviation, a set point that is predicted to go off the mark. The allostatic process thus allows an organism to prepare, by adapting its body and behaviour, for a change in the environment that is relevant to its organismic comfort zone. The Human Affectome defines allostasis as a "predictive process to maintain stability despite change" (p.18) which produces "measures of predicted homeostatic need" (p.10) But the Human Affectome, as well as the active inference literature, still make use of the concept of homeostasis, even though allostasis was defined to replace homeostasis. Different literatures and authors therefore mean different things when using both concepts (see McEwen and Wingfield (2010) for a discussion). In this text, we will follow the definitions of the Human Affectome, keeping the concept of homeostasis as used throughout this text, and the concept of allostasis as an active process of predictive regulation that works in anticipation of homeostatic needs (McEwen & Wingfield, 2010; Romero, Dickens & Cry, 2009). Therefore, in Sapolsky's water shortage example, we would consider thirst to be homeostatic, as the amount of water is already below the homeostatic set point, and the body is reacting to correct this imbalance. But a fear of a water shortage, which would encourage the acquisition of water for future use, would fall within allostatic regulation.

3.2 An Active Inference Account of Affects

3.2.1 Could Free Energy Be Felt?

Based on our terminology, we could describe physiological affective concerns as being homeostatic affective concerns. Moreover, we could analogously describe operational affective concerns (emotions) as being allostatic affective concerns.

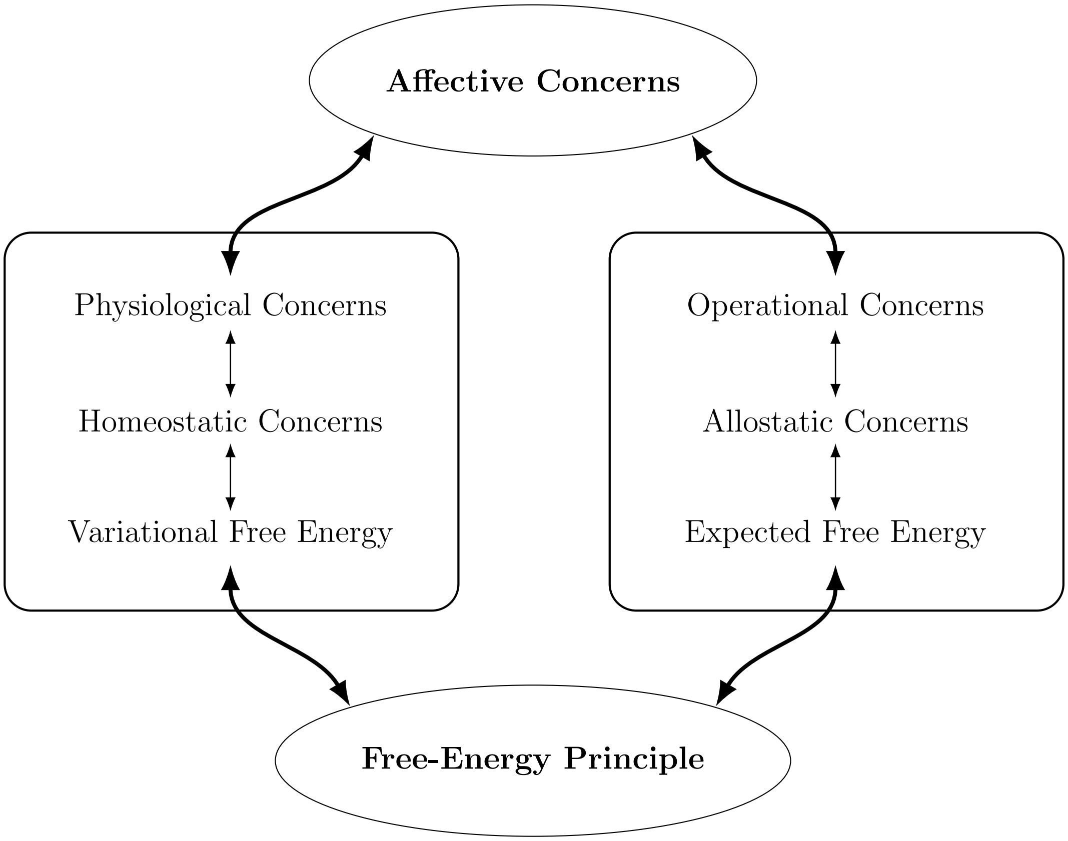
If we recall, according to active inference, there are two kinds of free energies: variational free energy and expected free energy. Variational free energy, as we have seen, is intertwined with homeostasis, as the inherent surprise of an observation is set by an organism's inner model of the world, which is assumed to attribute extremely high probability—and so low surprise—to observations consistent with its own viability. To cite the fish example again, as water is essential to the life of a fish, the observation of being surrounded by water is extremely unsurprising. But the observation of being surrounded by air is inversely extremely surprising and thus generates a high amount of (variational) free energy. Expected free energy, on the other hand, is the free energy we expect with respect to a series of observations and actions to come. Therefore, if our fish sees a predator, the observation should generate a high level of (expected) free energy as the presence of the predator suggests, according to the fish's inner model, a future in which it might lose homeostatic balance, by getting injured or eaten. The fish will thus take a series of actions, a trajectory, that will minimize (expected) free energy, which might involve hiding through algae, for instance. Expected free energy is all about anticipating the future, and under the free energy principle, it is about taking the path that is believed to be best for the enactment of an organism's relevance. The minimization of expected free energy is an allostatic process. Or maybe even the mathematical description of allostasis.

What we are doing here is to explicitly connect homeostatic affects to variational free energy and allostatic affects to expected free energy. Our analysis so far is illustrated in Figure 3.1.

The probabilities of making various observations flow from the organism's inner model, which contains affective concerns. The surprise of an observation, and consequently its free energy or prediction error, is set by homeostatic affective concerns, in the case of variational free energy, and by allostatic affective concerns, in the case of expected free energy.

Figure 3.1

Affective Concerns and The Free-Energy Principle

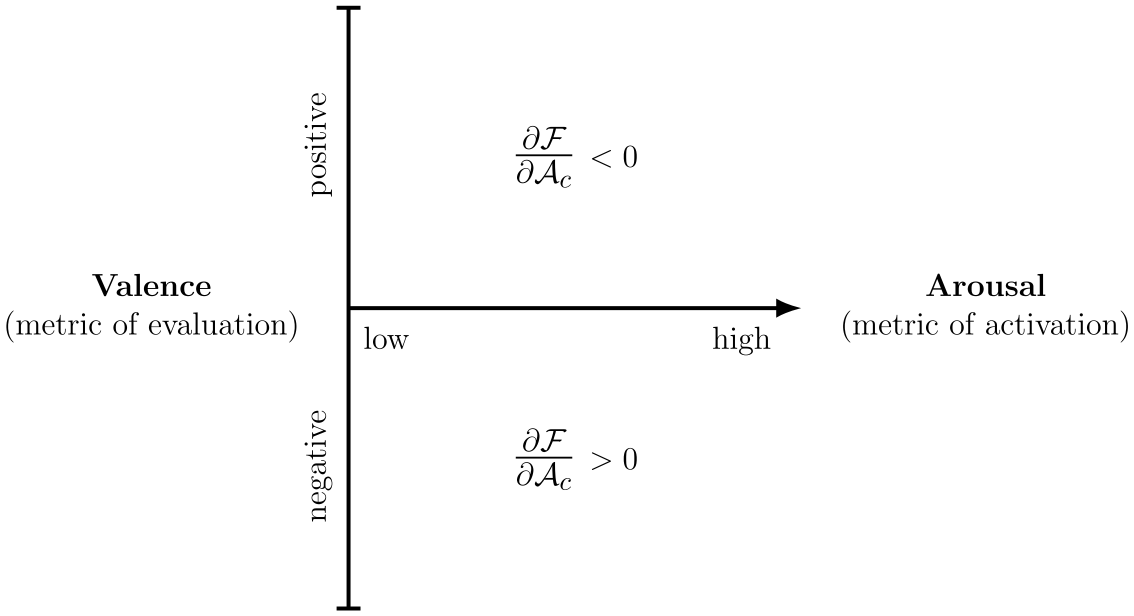


Note. An attempt to connect the free-energy principle and affective concerns through the homeostatic and allostatic processes.

What about affective features? Arousal could be related to the amount of free energy; the higher the free energy, the higher the arousal. However, free energy is a positive quantity, which seems to be an issue for valence. But what if valence were associated with a variation in free energy? After all, although always positive, free energy can vary. This idea, it turns out, is at the core of the first formal account of affects under active inference; Joffily and Coricelli (2013) define pleasure as a positive rate of change in free energy and displeasure as a negative rate of change in free energy. Hence, an affective experience would be produced by passing from one free energy level to another, rather than with respect to a static level of free energy. Figure 3.2 seeks to capture our active inference account of affective features.

Figure 3.2

Affective Features and Variation in Free Energy (First Derivative)


positive
$\frac{\partial \mathcal{F}}{\partial \mathcal{A}_c} < 0$
Valence
(metric of evaluation)
low
high
Arousal
(metric of activation)
negative
$\frac{\partial \mathcal{F}}{\partial \mathcal{A}_c} > 0$

Note. On the y-axis, we have the valence metric, which takes values on negative-positive dimensions. On the x-axis, we have the arousal metric, which takes values on the low-high dimensions. The inequality in the upper quadrant represents the partial derivative (variation or rate of change) of free energy $\mathcal{F}$ with respect to a particular affective concern, denote $\mathcal{A}_c$. However, an affective experience can be global, integrating the valence and arousal with respect to multiple affective concerns (see Schiller et al. (2024), the Human Affectome paper). The magnitude of the arousal depends on the magnitude of the rate of change.

According to this view, the experience of pleasure is a felt reduction in free energy that encourages an organism to take actions that minimize free energy, and the experience of displeasure is a felt augmentation of free energy that discourages an organism from taking actions that increase free energy. This account fits nicely with Marcus' (2008) comment on the happiness treadmill, the idea that we have to constantly do things to stay in a state of happiness. He describes the opportunity to pursue happiness as "[...] little more than a motor that moves us. The happiness treadmill keeps us going: alive, reproducing, taking care of children, surviving for another day. Evolution didn't evolve us to be happy, it evolved us to pursue happiness" (p. 139) If an organism encounters an object (including living organisms and situations) that, according to its model, would be good for its viability with respect to an affective concern, this inference would have the effect of raising free energy, which would translate into an incentive to close the free energy gap by taking actions that help the organism interact with the object in the way believed to further ensure its viability. And this reduction in free energy would feel pleasurable. Similarly, hunger would be a felt manifestation of the increase of free energy with respect to the energy level of the body.

Of course, this definition of affects under active inference raises many issues that may not seem to be easily reconcilable. For instance, if one's body is below the organismic comfort zone, the warming process will reduce free energy with respect to body temperature (homeostatic concern), and possibly feel good. However, a body temperature that deviates from the homeostatic set point seems to produce displeasure even when it becomes static.

In the end, according to Clark, Watson and Friston (2018), arguments based mostly on theoretical considerations have converged to an understanding of affective phenomena as variations in uncertainty about the physiological consequences of actions (Barrett, Quigley & Hamilton, 2016; Joffily & Coricelli, 2013; Gu et al., 2013; Seth, 2013; Seth & Friston, 2016). This understanding is closely related to a particular view of emotional experience where emotions are thought to be built on basic affective features and inferred from the brain's model which integrates cues from both the body (interoceptive) and the environment (exteroceptive), such as the immediate context (Barrett 2017, 2019; Seth & Friston, 2016). The resulting definition is articulated around the idea of interoceptive inference, where felt experiences would be generated by "[...] active interpretation of changes in the physiological conditions of the body" (Gu et al., 2013, p. 3372). Moreover, this active interpretation is related to an organism's prior beliefs about the anticipated somatic consequences of a given action (Seth & Friston, 2016). Affective phenomena would thus be directly related to descending interoceptive predictions that encounter ascending interoceptive signals (Seth & Friston, 2016). From this process emerges a sense of control when uncertainty is resolved, and a sense of loss of control when uncertainty increases. Clark, Watson and Friston (2018) compactly summarize what valence is according to active inference:

> [...] positively valenced brain states are necessarily associated with increases in the precision of predictions about the (controllable) future—or, more simply, predictable consequences of motor or autonomic behaviour. Conversely, negative emotions correspond to a loss of prior precision and a sense of helplessness and uncertainty about the consequences of action.

This understanding of emotions has led to the elaboration of computational models of interoceptive inference and computational architectures of the brain (Barrett, 2017; Barrett & Satpute, 2013; Seth & Friston, 2016; Smith et al., 2019; Smith, Parr & Friston, 2019; Tschantz et al., 2022). This impressive and growing literature could constitute a primary source of inspiration for the affectivization of Ororbia and Kelly's CogNGen (Ororbia & Kelly, 2022a, 2022b, 2023).

3.2.2 Affects as Free Energy Minimization Mechanism

If everything the brain does minimizes free energy, and the brain does affective phenomena, then it follows that affective phenomena minimize free energy. This conclusion holds irrespectively of our ability to characterize affective concerns and features through the free energy principle.

In short, affective phenomena can be considered to contribute significantly to the minimization of free energy by being felt incentives for an organism to enact its relevance. These incentives take the form of felt evaluations of how well an organism fares with respect to affective concerns that are directly tuned to allostatic and homeostatic needs. An affective experience pushes, in a sense, the organism toward an action that is believed by the model to minimize free energy. Affective phenomena also attract the organism's attention to the information that is most critical to its integrity; this information is harnessed from both the inside and the outside of its body. By doing so, they can sometimes act as task organizers, automatically ranking affective concerns in order of priority based on the inner model of the organism (Bolles & Fanselow, 1980). For instance, at least in some mammals, a painful stimulus is often ignored if a serious threat is present, from rats ignoring a burning tail when exposed to a cat to injured soldiers on the battlefield (Bolles & Fanselow, 1980; LeDoux, 1998). In such situations, dealing with the allostatic concern of threat is prioritized over the homeostatic concern of pain, which suggests that, depending on their relative severity, the threat stimulus can raise more free energy than the painful stimulus.

That said, we must not lose sight of the many situations where affective phenomena seem to be a nuisance. An anger that incentivizes the takedown of an object causing obstruction might lead to a subsequent state of affairs where the organism fares worse than if it did not enact its anger that way. In a scenario like this, some planning and strategizing might lead to a better future. But before entering the territory of higher cognition, let us observe that a wave of immediate anger can be thwarted by a more distal fear of the consequences of enacting the anger. As highlighted above, an affective concern integrates object relevance that varies on the proximal-distal time scale. Each affective process acts like a heuristic, a simple rule that is overall helpful on a given timescale but that is expected to fail in many situations. We now turn to the interaction between affect and cognition.

3.3 The Cortex as a False Positives and Free Energy Reductor

According to a simplified model of mammalian brains (LeDoux, 1998, Panksepp 2012; Sapolsky, 2017), the core brain structures involved in affective experience are built on top of the older brain structures concerned with basic homeostatic regulation. We will refer to this collection of brain regions as the limbic system. The circuitry that allows for allostatic processes came after the fundamental circuitry that maintains basic physiological regulation—although all structures never stopped evolving. As such, the evolution of brains could be framed in terms of free energy minimization, where the abstract selective criterion for the success of novelty in the brain is its contribution to the overall task of minimizing free energy. Hence, from the standpoint of active inference, we would expect the cortex, the most recent brain layers which envelope the inner brain, to be yet another free-energy minimization innovation of nature. There are many ways for the brain to further reduce free energy through cognition, but one that does not require too much creativity and that follows rather straightforwardly from the allostatic limbic system consists of simply improving the affective mechanism and allostasis more generally. That is, improve the predictions the limbic system makes. In that light, not only does active inference unite perception and action, the theory also unites emotion and cognition—both work toward the minimization of a unique quantity (free energy).

It is important to note that in this last section, we embrace the view of basic emotions (e.g., LeDoux, 1998; Panksepp & Biven, 2012) that has dominated the field of affective science. However, since the early 2000s, this view has been seriously challenged (See Barrett (2019) for a discussion of the weaknesses of the theory of basic emotions and for her suggestion of an alternative approach that goes by the name of the theory of constructed emotions; see LeDoux (2012) for his updated view of the emotional brain).

In what follows, we will sketch an abstract, high-level portrait of how cognitive modules could interact with affective modules. Moreover, the focus will be on how these interactions follow the logic of active inference and further reduce free energy. We will cover only a few cases of how cognition helps regulate affect. But a similar analysis could be done in the other way around since affect also regulates cognition. To be concretely helpful in building cognitive architectures, going beyond this preliminary analysis and performing a more complete characterization of how these influences work in both directions would be necessary. (This is indeed what many of the projects referenced at the end of section 3.2.1 are doing).

3.3.1 True/False Positives Asymmetry and Free Energy

In many contexts, a false positive is less costly than a false negative; they are asymmetric in their consequence. For instance, let us consider the classic snake example used by LeDoux (1998) in his discussion on fear conditioning in *The Emotional Brain.* For an organism vulnerable to snakes (like primates), mistaking a stick for a snake (false positive with respect to the snake) is less problematic than the opposite, namely mistaking a snake for a stick (false negative with respect to the snake). An organism that consistently mistakes a stick for a snake has a brighter future—more chance to reproduce before dying—than one that consistently mistakes a snake for a stick—a mistake that can't be made many times. Based on this simple evolutionary reasoning, we would expect an elementary affective mechanism to be more sensitive to objects that an organism believes (consciously or not) to be particularly dangerous.

Table 3.1

Confusion Matrix for the Detection of a Snake

	Cause = Snake	Cause = Stick
Prediction = Snake	true positive	false positive
Prediction = Stick	false negative	true negative

Note. In this confusion matrix, we treat the cause and prediction "stick" as "not a snake."

In the language of active inference, raising the sensitivity of an object gets translated into raising the prior probability of that hidden state. For instance, in an environment rich in sticks and poor in snakes, the allostatic relevance of snakes could make the affective system artificially raise the probability of encountering a snake, hence diminishing the probability of a false negative. The expression *allostatic relevance* maps to the more general notion of *utility*, that is, the worth of an object. Although not the focus of this section, we might as well note in passing that this is an example of how affects modulate perception through utility as encoded in the prior probability of a cause. For a mathematical example, please refer to Appendix 2. This differs from the way observations consistent with homeostatic set points are encoded in the marginal (see section 2.3.2). It may even be contradictory. We are nonetheless going to try and see where this approach leads us.

To recap, some false negatives due to defective perception ("this is a stick") can be very costly. However, the lower cost of a false positive, as compared to a false negative, does not make it free of charge. And a high rate of false positives is not cheap, energetically speaking. For instance, in a fear response, the arousal triggered is energy-consuming—elevated heart rate and blood pressure, etc. These false positives can be cast as events that artificially raise free energy. Hence, a reduction of the frequency at which an organism is needlessly primed to face danger—or other affective concerns—is a reduction in prediction error and in free energy. Interestingly, cortical areas seem to be doing just that. LeDoux (1998) writes: "The information received from the thalamus is unfiltered and biased toward evoking responses. The cortex's job is to prevent the inappropriate response rather than to produce the appropriate one" (p. 165). Let us unpack this.

3.3.2 The Fast Track and the Slow Tack to the Amygdala

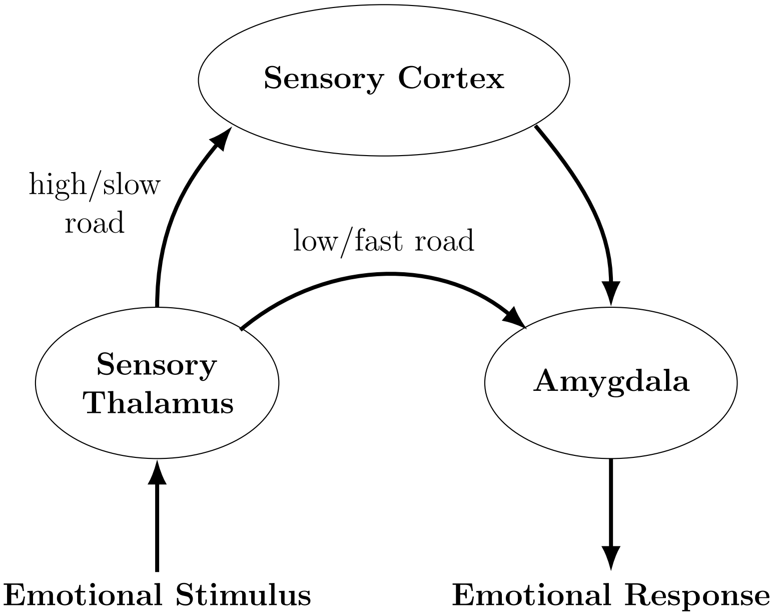
From the late nineteenth century throughout the twentieth century, many researchers interested in aggressive behaviours have investigated the effects of removing the cortex of animals—like rats, cats, dogs and monkeys (Kaada, 1967). As many studies confirmed, after ablation, the animals displayed "abnormal emotional hyperexcitability" (Spiegel, Miller & Oppenheimer, 1940). But what was most surprising, given the state of neuroscience and physiology at the time, was that these poor animals were still capable of emotional behaviours at all (LeDoux, 1996). For instance, cats with their cerebral cortex removed reacted to provocation as regular cats do—they arced their back accompanied by an autonomic piloerection, hissed at the provocateur and so on. However, although the emotional responses were apparently intact, their triggers were not: these animals were now reacting to a much wider set of events as if their provocation threshold was lowered and widened—the phenomenon became referred to as “sham rage” (Kaada, 1967).

To better understand the interplay between cortical areas and emotional responses, precise studies of how the brain processes sensory stimuli were needed. LeDoux, Sakaguchi and Reis (1984) carried out such a study, focusing on acoustic stimuli. The flow of information coming from the ear travels from the cochlear nucleus to the auditory brainstem, then to the auditory midbrain before reaching the auditory thalamus. This was known. What they wanted to know was if the auditory thalamus projects to brain regions other than the auditory cortex once it receives signals from the ears. To be able to observe neuronal projections, they used a chemical tracer. These chemicals are injected into the desired area of investigation; they enter the bodies of cells and then travel along the axon with the other chemicals used to send signals. Moreover, the substance reacts with other chemicals and stains the neurons as it moves through, making it possible to see its trace. Using this technique, the three neuroscientists were able to identify four sub-cortical regions to which neurons of the auditory thalamus project. Before that study, it was generally believed that the thalamus signalled only to the cortex. In a follow-up study, LeDoux and his team (1986) investigated the effect of preventing the flow of information from the auditory thalamus to each of the four pathways. The only lesions that brought the rats’ susceptibility to fear conditioning to a halt was the one interrupting the efferent connections between the thalamus and the amygdala; when this pathway was unavailable, so was fear conditioning—or at least its magnitude was greatly reduced.

Enlightened with that knowledge, Jarrell and colleagues (1987) designed an auditory fear conditioning study on rabbits. The rabbits learned to discriminate between two tones. A mild electric shock followed the first tone but not the second. Then, the researchers lesioned the neuronal pathway from the thalamus to the auditory cortex. Post-surgery, the first tone still elicited conditioned fear, but so did the second tone. The cells that project from the thalamus to the auditory cortex are said to be narrowly tuned so that the auditory cortex can distinguish similar acoustic stimuli and then send its fine-grained analysis to the amygdala. In contrast, the analysis performed by neurons projecting from the thalamus to the amygdala is coarser—the cells are broadly tuned. As LeDoux (1998) evocatively sums up: "The Beatles and Rolling Stones [...] will sound the same to the amygdala by way of the thalamic projections but quite different by way of the cortical projection" (p. 162) In the rabbit study, the two tones fell into the same thalamus-to-amygdala bucket. Therefore, the thalamus relayed the same message to the amygdala for both tones, whereas the auditory cortex of pre-lesioned rabbits was receiving from the thalamus and sending to the amygdala different messages—hence the absence of a fear response when the second tone played. Furthermore, the study suggests that the role of the pre-lesioned auditory cortex was to inhibit the amygdala which was already activated by the signal from the thalamus—the thalamus-to-amygdala track is faster than the thalamus-to-auditory-cortex-to-amygdala track (see Figure 3.3).

Figure 3.3

The Fast and the Slow Road to the Amygdala



Note. This figure closely replicates Figure 6-13 from LeDoux (1998), with minor adaptations to the present text. A signal from the sensory thalamus that passes by the sensory cortex before reaching the amygdala follows a longer and slower path than a signal that flows directly from the sensory thalamus to the amygdala.

Generalizing the inhibitory role of the auditory cortex to other sensory cortices, LeDoux (1998) concludes the snake-and-stick example as follows: “The curvature and slenderness reach the amygdala from the thalamus, whereas only the cortex distinguishes a coiled up snake from a curved stick. If it is a snake, the amygdala is ahead of the game” (p. 165). Canceling the reaction of the fast-but-mistake-prone track lends a higher chance of survival than waiting for the slow-but-more-accurate track to make their call. This setting reduces free energy overall by accepting to pay small energetic costs due to prediction error but avoiding high costs due to slow accurate predictions. It is no coincidence that, as a general rule of animal neurobiology, among the sensory modalities present in a given species, it is the predominant one—like olfaction in rodents or vision in primates—that has the most privileged access to the limbic system (Sapolsky, 2017). If affect modulates the utility of perceptions, this pattern of direct access by the prevailing sense is something to expect.

3.3.3 Frontal Cortex Regulation of the Limbic System—Two Cases

This modulation of the emotional response by the cortex has been documented many times for various functions and is now a widely accepted notion (Eysenck & Keane, 2020; Munakata, 2011; Postle, 2015). But while the previous example of acoustic stimuli illustrated the general idea of sensory cortex inhibition, the region most famously known for its inhibitory activity is the frontal cortex. In his magistral synthesis of human behaviour, *Behave*, Sapolsky (2017) gives many examples of this frontal regulation pattern. In what follows, we will briefly explore two of them and see how they represent false positive rate reduction through cortical regulation of the limbic system.

Interracial Fear Regulation

The setting for the first example involves exposure to subliminal stimuli. A subliminal sensory input is one that goes unnoticed by our conscious brain but is nonetheless processed. This type of perception is called subliminal perception—or perception without awareness

(Eysenck & Keane, 2020, p. 81). The threshold for conscious perception is around 500 milliseconds (Elgendi et al., 2018). A subliminal stimulus could be a picture that flashes so briefly before our eyes that we are not sure whether we have seen anything at all. But other parts of our brain notice aspects of the stimulus. For instance, when people are shown a picture at subliminal speed (100 hundred milliseconds), they nevertheless do better than chance at guessing the colour of the face when asked to—after having been informed that they did see something and that the something was a face (Kubota, Banaji & Phelps, 2012). When the face shown is one from a different race, people's amygdala activates despite having no conscious experience of the image—not that surprisingly, the level of activation correlates with the level of implicit racial bias (Ito & Urland, 2003).

Of particular interest to us, in a study where white participants were shown pictures of a black face long enough to fall out of the subliminal range and thus be consciously experienced, their frontal cortex (specifically the anterior cingulate cortex and the dorsolateral prefrontal cortex) was able to recruit its knowledge and general beliefs about the world and inhibit the activity of the amygdala (Kubota, Banaji & Phelps, 2012; Richeson et al., 2003; Richeson & Trawalter, 2005). The detour to the frontal cortex allowed the avoidance of a false positive. As Sapolsky (2017) writes: "It's the frontal cortex exerting executive control over the deeper, darker amygdaloid response" (p. 85).

Social Rejection in Adolescents versus Adults

For the second example, the setting consists of a paradigm organized around an online game of catch developed to study social exclusion (Williams, Cheung & Choi, 2000). Based on her work on pain and social exclusion (Eisenberg, 2003), Eisenberger and her colleagues used the virtual game of catch to compare the neural correlates of rejection in adolescents and adults (Masten et al., 2009). Study participants played the game from within a brain scanner. Importantly, they did not know that the two other players were computer programs; they believed that they were controlled by other people. The real experiment began after a while when the two computer players stopped throwing the virtual ball to the participant. With adults, here is what happened: When they began to feel excluded by the two other players, brain areas associated with pain, anger and disgust light up (periaqueductal gray, anterior cingulate, amygdala, insular cortex). Then, this neuronal activity got tuned down when the frontal cortex intervened (specifically, the ventromedial prefrontal cortex), and the person reminded themselves that they were just playing a silly game that did not matter. But when the participants were teenagers, with their teenage frontal cortex not fully matured, the cortical inhibition barely happened, and so the feeling of social rejection stayed strong—only in the few teenagers that were the least susceptible to rejection and were often surrounded by their friends did the inhibitory signal do its job.

Commenting on the study, Sapolsky (2017) observes: "Rejection hurts adolescents more, producing that stronger need to fit in" (p. 166, emphasis in the original text). Again, in adults, cortical inhibition helps discriminate between situations where subcortical social rejection calls are true or false positives. But in adolescents, this inhibition does not take place as much, and the more broadly tuned limbic system makes sure due attention is paid to a wide range of possible situations where rejection is possible, causing a higher rate of false positives. Of course, we should note that the delayed maturation of the frontal cortex and the effects it engenders—like a more powerful incentive (via the pain of not belonging) to be part of a group in adolescence—is probably adaptative. Missing out on signals that suggest we are in the process of being excluded from a group is probably more costly at an age when we haven't reached maximum autonomy.

In summary, one way the cortex contributes to minimizing free energy is through regulating the limbic system and reducing its prediction errors, especially the rate of false positives. The regulation is inhibitory in part because cortical pathways take longer to travel and a certain amount of computation is needed for the cortex to be able to take advantage of the knowledge about the world it has built over time. In cognitive modelling, this type of inhibitory control could be built in an architecture that is endowed with a form of affective hub. Moreover, each aspect of cognition would also have to be modulated by the affective module. For instance, we saw how affect-based utility can modulate perception. Of course, many more aspects of cognition are modulated by affect and vice-versa. To model brains, the influence between affect and cognition in a cognitive architecture should be mutual and ubiquitous.

Discussion

In this text, we tried to understand the basic conceptual and mathematical formulation of active inference. This objective was motivated by a desire to explore how affective phenomena could be integrated into cognitive architectures, which would perhaps be more appropriately called *mind* architectures (LeDoux, 1998). Attempting to formalize affective phenomena via active inference made sense on at least two counts: First, the CogNGen was the cognitive architecture chosen for the initial goal of thinking about the implementation of affective modules that would interact with the cognitive modules; the learning algorithms in the CogNGen, based on neural generative units, are already built around the free-energy principle and predictive coding. Second, active inference provides a powerful mathematical formalism to develop algorithms and models of biological processes, and many of the promising emerging computational models of emotions are within this framework. The Human Affectome provided our characterization of affective phenomena. This characterization acted as a cornerstone on which we tried to sketch an active inference account of affects. We then explored a few ways in which the most recent and primarily cognitive brain structures further reduce free energy by better controlling the rate of false positives.

Counterfactuals

As we saw with emotions, active inference offers a fertile ground to revisit familiar psychological and biological phenomena from a new perspective, but also to ask new or different questions altogether. To illustrate, let us consider counterfactual reasoning. Counterfactuals are about trajectories of events that could happen; they usually take the form of what-if questions. In active inference, references to counterfactuals are sometimes made when expected free energy is discussed. As we saw in section 1.3.2, adaptive systems with enough brain complexity can select the series of actions that lead to the most favourable consequences among many such alternative trajectories. In that sense, an organism can "engage its generative model vicariously to run 'what if' or counterfactual simulation of the consequences of its possible actions [....] (Parr, Pezzulo & Friston, 2022, p. 32). For instance, an organism can consider what would happen if it were to go through a fire or around a fire. Although not yet fully explored, this line of investigation has been noted a few times (Millidge, 2023; Parr, Pezzulo & Friston, 2022; Seth & Friston, 2016) and leads to interesting, if not perplexing questions. Could counterfactuals be statically based? Could other animals be capable of counterfactual reasoning? It could be that active inference answers in the affirmative to both questions, which are usually answered in the negative.

For instance, in a recent paper, a team of AI scientists analyzed different approaches to counterfactuals; they categorized statical models as unable to answer counterfactual questions (Schölkopf et al., 2021). Moreover, counterfactual reasoning has long been thought to be one of the distinguishing features of the human mind, a belief we find in Judea Pearl's ladder of causation, where counterfactual reasoning thrones on the highest rung only reached by humans (Pearl, 2018). The hidden presence of counterfactuals in the non-human animal world would be consistent with the trend that we observed during the last few decades where many of the traits of mind that were once declared to set humans apart turned out to be observed in other species in various degrees. Such traits are now better characterized by a continuum than a binary have or have not a dichotomy, and so, to reuse the words of primatologist Frans de Waal (2010), through Darwinian evolution, "[...] we are continuous with all other life forms, not only in body but also in mind" (p. 207). Maybe the causal relationships that other non-human animals are capable of, due to their perhaps less powerful hierarchical networks, are in the order of seconds or milliseconds, such that the capacity went unnoticed.

Or maybe there is a misunderstanding. Could it be that active inference researchers use the term "counterfactuals" in an unconventional way? Perhaps.

Counterfactuals deal with events that did or did not happen in the past. To use classical illustrative examples, the two common forms of counterfactuals are: "If Oswald didn't kill Kennedy, someone else did" (indicative form) and "If Oswald didn't kill Kennedy, someone else would've" (subjunctive form) (Starr, 2022) This is indeed how Pearl uses counterfactual ("What if X had not occurred?" (Pearl, 2018, p. 28)) as well as the team of AI scientists just cited ("Counterfactual problems involve reasoning about why things happened, imagining the consequences of different actions in hindsight, and determining which actions would have achieved the desired outcome."). Based on a quick analysis, it seems that what active inference sometimes call "counterfactual questions" are elsewhere referred to as "interventional questions." For instance, "how does the probability of heart failure change if we convince a patient to exercise regularly?" is an interventional question; conversely, "would a given patient have suffered heart failure if they had started exercising a year earlier?" is a counterfactual one (Schölkopf et al., 2021, p. 615). The question about going through or around the fire seems to be an interventional question, not a counterfactual one.

The Brain's Philosophy of Science

Declaring a question "new" is a bold thing to do. And while new theories often do it, we will not take the risk here. Instead, let us ask a question that is probably not new, but on which someone exploring the ideas of active inference might land: What is the *brain*'s philosophy of science? After all, throughout its evolution, the brain had to deal with many of the challenges faced by scientists and philosophers of science. How does the brain answer to the problem of induction (Hume, 2008)? How does the brain answer to the problem of underdetermination of theory by data (Stanford, 2023)? How does the brain answer to the problem of unconceived hypotheses (Stanford, 2006)? What does the brain do when it makes observations recalcitrant to its model of the world? Does it reject the model as a falsificationist in the style of Popper would? Does it take the stand of scientific realism or antirealism (Chakravartty, 2017)? Does the brain embrace a form of pluralism, entertaining many competing theories at once (Chang, 2012)?

Although we did not comment on it, in section 3.3.1 and Appendix 2, we did provide a tentative answer to the question of the underdetermination of theory by data. Put simply, this problem emerges when two competing theories are consistent with the same set of data; both theories are able to explain the observations. Perception is confronted with this challenge all the time when input stimuli fall in the intersection of many hidden states. One way we suggested that the brain handles such cases is by giving more weight to causes higher in relevance with respect to the organism that is having the sensation.

Let us turn to the question of disconfirming evidence. One of the most counterintuitive aspects of active inference, namely, how actions are propelled by beliefs contradicted by observations/evidence (see section 1.4), might make more sense when treated as a consequence of the brain's philosophy of science. To borrow the terminology of French physicist and philosopher of science Étienne Klein (2013), when an observation seems to contradict a theory, scientists are offered at least two solutions: the legislative solution and the ontological solution. The *legislative* solution involves modifying the theory and the hypothesized laws that are believed to regulate the world. The *ontological* solution involves modifying our ontology, that is, the set of objects we believe the world contains. To use Klein's examples, when the predictions about Uranus' orbit made by Newtonian physics did not agree with observations, astronomers could modify the Newtonian model or modify their ontology. They took the second solution, postulated the existence of another planet, hence modifying their ontology, and then went on to discover Neptune. When a similar situation arose, this time the issue was with Mercury's orbit, the same ontological solution led to the postulation of planet Vulcan, which was never discovered. The accumulation of these disconfirming observations contributed to motivating a legislative solution that took the form of General Relativity decades later. The parallelism with active inference is not perfect, but the similarities are still striking. With a slightly misleading way of using language, *perception* corresponds to the legislative solution (updating the model through observations), while *action* corresponds to a kind of ontological solution (modifying the world so that the observations agree with the model once more).

This is only a sketch of the project of identifying the brain's philosophy of science, a project for which active inference offers many insights. Of course, uncovering how the brain answers puzzling philosophical and practical questions that scientists and philosophers of science wrestle with should by no means be taken as normative. It is not because the brain answers a question in a particular way that we want or should take the same path. In doing so, we would commit the naturalistic fallacy, where one seeks to illegitimately derive an *ought* from an *is* (Hume, 2000). However, in the case of artificial intelligence, such an enterprise could constitute a source of inspiration.

References

Abelson, R. P. (1963). Computer simulation of “hot cognition”. *Computer simulation of personality*, 277-298.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1), 147-169. https://doi.org/10.1016/S0364-0213(85)80012-4

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1), 1-23. https://doi.org/10.1093/scan/nsw154

Barrett, L. F. (2019). *How Emotions Are Made.* Providence Book Festival, May, 25, 2019.

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1708), 20160011. https://doi.org/10.1098/rstb.2016.0011

Barrett, L. F., & Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Current opinion in neurobiology*, 23(3), 361-372. https://doi.org/10.1016/j.conb.2012.12.012

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Bolles, R. C., & Fanselow, M. S. (1980). A perceptual-defensive-recuperative model of fear and pain. *Behavioral and Brain Sciences*, 3(2), 291-301. https://doi.org/10.1017/S0140525X0000491X

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to Bayesian statistics.* John Wiley & Sons.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, 10(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Chakravartty, Anjan, "Scientific Realism", The Stanford Encyclopedia of Philosophy (Summer 2017 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/

Chang, H. (2012). *Is water H2O?: Evidence, realism and pluralism* (Vol. 293). Springer Science & Business Media.

Clark, J. E., Watson, S., & Friston, K. J. (2018). What is mood? A computational perspective. *Psychological medicine*, 48(14), 2277-2284. https://doi.org/10.1017/S0033291718000430

Clayton, A. (2021). *Bernoulli's fallacy: Statistical illogic and the crisis of modern science.* Columbia University Press.

Carrie L. Masten, Naomi I. Eisenberger, Larissa A. Borofsky, Jennifer H. Pfeifer, Kristin McNealy, John C. Mazziotta, Mirella Dapretto, Neural correlates of social exclusion during adolescence: understanding the distress of peer rejection, *Social Cognitive and Affective Neuroscience*, Volume 4, Issue 2, June 2009, Pages 143–157, https://doi.org/10.1093/scan/nsp007

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural Computation*, 7(5), 889-904. https://doi.org/10.1162/neco.1995.7.5.889

De Waal, F. (2010). *The age of empathy: Nature's lessons for a kinder society.* Crown.

Dukes, D., Abrams, K., Adolphs, R., Ahmed, M. E., Beatty, A., Berridge, K. C., Broomhall, S., Brosch, T., Campos, J. J., Clay, Z., Clément, F., Cunningham, W. A., Damasio, A., Damasio, H., D'Arms, J., Davidson, J. W., De Gelder, B., Deonna, J., De Sousa, R., ... Sander, D. (2021). The rise of affectivism. *Nature Human Behaviour*, 5(7), 816–820. https://doi.org/10.1038/s41562-021-01130-8

Eisenberger, N. I. (2012). The pain of social disconnection: examining the shared neural underpinnings of physical and social pain. *Nature reviews neuroscience*, 13(6), 421-434. https://doi.org/10.1038/nrn3231

Eysenck, M. W., & Keane, M. T. (2020). *Cognitive Psychology: A Student's Handbook* (8th ed.). Psychology Press. https://doi.org/10.4324/9781351058513

Frankish, K., & Ramsey, W. (Eds.). (2012). *The Cambridge handbook of cognitive science.* Cambridge University Press. https://doi.org/10.1017/CBO9781139033916

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2), 127-138. https://doi.org/10.1038/nrn2787

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 815-836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159(3), 417–458. https://doi.org/10.1007/s11229-007-9237-y

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press. https://doi.org/10.1038/nature14539

Gould, H., & Tobochnik, J. (2021). *Statistical and Thermal Physics: with Computer Applications.* Princeton University Press.

Gu, X., Hof, P. R., Friston, K. J., & Fan, J. (2013). Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology*, 521(15), 3371–3388. https://doi.org/10.1002/cne.23368

Hume, D. (2000). *A treatise of human nature.* (Oxford Philosophical Texts), David Fate

Norton and Mary J. Norton (eds.), Oxford, Clarendon Press.

Hume, D. (2008). *An enquiry concerning human understanding* (P. Millican, Ed.). Oxford University Press. (Original work published 1748)

Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology*, 85(4), 616–626. https://doi.org/10.1037/0022-3514.85.4.616

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10), 1295-1306. https://doi.org/10.1016/j.visres.2008.09.007

Jarrell, T. W., Gentile, C. G., Romanski, L. M., McCabe, P. M., & Schneiderman, N. (1987). Involvement of cortical and thalamic auditory regions in retention of differential bradycardiac conditioning to acoustic conditioned stimuli in rabbits. *Brain Research*, 412(2), 285–294. https://doi.org/10.1016/0006-8993(87)91135-8

Kaada, B. (1967, January). Brain mechanisms related to aggressive behavior. In *UCLA Forum Med Sci* (Vol. 7, pp. 95-133).

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, 10(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Klein, É., on CentraleSupélec. (2013, March 25). Étienne Klein - Cours introductif de Philosophie des Sciences 6/9. [Video]. *YouTube*. https://youtu.be/Gnunx9V5EP8?si=7-z3vn91ye52ZkcP

Kubota, J. T., Banaji, M. R., & Phelps, E.A. (2012). The neuroscience of race. *Nature Neuroscience*, 15(7), 940–948. https://doi.org/10.1038/nn.3136

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine*, 38(4), 13–26. https://doi.org/10.1609/aimag.v38i4.2744

LeDoux, J. E. (1998). *The emotional brain: The mysterious underpinnings of emotional life.* Simon and Schuster.

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron*, 73(4), 653-676. https://doi.org/10.1016/j.neuron.2012.02.004

LeDoux, J. E., Sakaguchi, A., & Reis, D. J. (1984). Subcortical efferent projections of the medial geniculate nucleus mediate emotional responses conditioned to acoustic stimuli. *Journal of Neuroscience*, 4(3), 683–698. https://doi.org/10.1523/JNEUROSCI.04-03-00683.1984

LeDoux, J. E., Sakaguchi, A., Iwata, J., & Reis, D. J. (1986). Interruption of projections from the medial geniculate body to an archi-neostriatal field disrupts the classical conditioning of emotional responses to acoustic stimuli. *Neuroscience*, 17(3), 615–627. https://doi.org/10.1016/0306-4522(86)90034-5

Lindsay, G. (2021). *Models of the mind: how physics, engineering and mathematics have shaped our understanding of the brain.* Bloomsbury Publishing.

Lemons, D. S. (2013). *A Student's Guide to Entropy.* Cambridge University Press. https://doi.org/10.1017/CBO9780511984556

McEwen, B. S., & Wingfield, J. C. (2010). What's in a name? Integrating homeostasis, allostasis and stress. *Hormones and Behavior*, 57(2), 105. https://doi.org/10.1016/j.yhbeh.2009.09.011

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746-748. https://doi.org/10.1038/264746a0

Miller G. A., and Johnson-Laird, P. (1976).*Language and perception* (Cambridge: Cambridge University Press).

Millidge, B. (2023, April 14). Predictive coding networks can perform causal and counterfactual inference. *Beren's Blog.* https://www.beren.io/2023-04-14-Predictive-coding-networks

Moerland, T. M., Broekens, J., & Jonker, C. M. (2018). Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, 107, 443-480. https://doi.org/

10.1007/s10994-017-5666-0

Munakata, Y., Herd, S. A., Chatham, C. H., Depue, B. E., Banich, M. T., & O'Reilly, R. C. (2011). A unified framework for inhibitory control. *Trends in Cognitive Sciences*, 15(10), 453–459. https://doi.org/10.1016/j.tics.2011.07.011

Newwell, A., Rosenbloom, P. S., and Laird, J. E. (1989). Symbolic architecture for cognition. In *Foundations of cognitive science*, M. Posner, ed. (Cambridge: MIT Press).

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Ororbia, A. G., & Kelly, M. A. (2022a). Cogngen: Building the kernel for a hyperdimensional predictive processing cognitive architecture. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44). Retrieved from https://escholarship.org/uc/item/35j3v2kh

Ororbia, A. G., & Kelly, M. A. (2022b, August). Maze learning using a hyperdimensional predictive processing cognitive architecture. In *International Conference on Artificial General Intelligence* (pp. 321-331). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-19907-3_31

Ororbia, A. G., & Kelly, M. A. (2023). A neuro-mimetic realization of the common model of cognition via hebbian learning and free energy minimization. In *Proceedings of the AAAI Symposium Series* (Vol. 2, No. 1, pp. 369-378). https://doi.org/10.1609/aaaiss.v2i1.27702

Panksepp, J. (2010). Affective neuroscience of the emotional BrainMind: Evolutionary perspectives and implications for understanding depression. *Dialogues in Clinical Neuroscience*, 12(4), 533–545. https://doi.org/10.31887/DCNS.2010.12.4

Panksepp, J., & Biven, L. (2012). *The archaeology of mind: Neuroevolutionary origins of human emotion.* W. W. Norton & Company.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle* *in Mind, Brain, and Behavior*. MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001

Postle, B. R. (2015). *Essentials of cognitive neuroscience* (1st ed.). John Wiley & Sons.

Richeson, J. A., & Trawalter, S. (2005). Why do interracial interactions impair executive function? A resource depletion account. *Journal of Personality and Social Psychology*, 88(6), 934–947. https://doi.org/10.1037/0022-3514.88.6.934

Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., & Shelton, J. N. (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature Neuroscience*, 6(12), 1323–1328. https://doi.org/10.1038/nn1156

Rosenblum, L., on BBC. (2010, November 10). Try This Bizarre Audio Illusion! [Video]. *YouTube*. https://www.youtube.com/watch?v=G-lN8vWm3m0

Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: Demystified and compared. *Neural Computation*, 33(3), 674–712. https://doi.org/10.1162/neco_a_01357

Salvatori, T., Mali, A., Buckley, C. L., Lukasiewicz, T., Rao, R. P. N., Friston, K., & Ororbia, A. (2023). Brain-Inspired Computational Intelligence via Predictive Coding (arXiv:2308.07870). arXiv. https://doi.org/10.48550/arXiv.2308.07870

Sapolsky, R. M. (2004). *Why zebras don't get ulcers: The acclaimed guide to stress, stress-related diseases, and coping*. Holt paperbacks.

Sapolsky, R. M. (2017). *Behave: the biology of humans at our best and worst*. New York, New York, Penguin Press.

Schiller, D., Yu, A. N. C., Alia-Klein, N., Becker, S., Cromwell, H. C., Dolcos, F., Eslinger, P. J., Frewen, P., Kemp, A. H., Pace-Schott, E. F., Raber, J., Silton, R. L., Stefanova, E., Williams, J. H. G., Abe, N., Aghajani, M., Albrecht, F., Alexander, R., Anders, S., . . . Lowe, L. (2024). The Human Affectome. *Neuroscience & Biobehavioral Reviews*, 158, 105450. https://doi.org/10.1016/j.neubiorev.2023.105450

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612-634. https://doi.org/10.1109/JPROC.2021.3058954

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11), 565-573. https://doi.org/10.1016/j.tics.2013.09.007

Simon, H. A. (1967). *Motivational and emotional controls of cognition.* Psychological review, 74(1), 29. https://doi.org/10.1037/h0024127

Smith, R., Lane, R. D., Parr, T., & Friston, K. J. (2019). Neurocomputational mechanisms underlying emotional awareness: insights afforded by deep active inference and their potential clinical relevance. *Neuroscience & Biobehavioral Reviews*, 107, 473-491. https://doi.org/10.1016/j.neubiorev.2019.09.002

Smith, R., Parr, T., & Friston, K. J. (2019). Simulating emotions: An active inference model of emotional state inference and emotion concept learning. *Frontiers in Psychology*, 10, 2844. https://doi.org/10.3389/fpsyg.2019.02844

Spiegel, E. A., Miller, H. R., & Oppenheimer, M. J. (1940). Forebrain and rage reactions. *Journal of Neurophysiology*, 3(6), 538–548. https://doi.org/10.1152/jn.1940.3.6.538

Stanford, P. K. (2006). *Exceeding our grasp: Science, history, and the problem of unconceived alternatives* (Vol. 1). Oxford University Press.

Stanford, Kyle, ”Underdetermination of Scientific Theory”, *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/sum2023/entries/scientific-underdetermination

Starr, W., Counterfactuals, *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/win2022/entries/counterfactuals

Stone, J. V. (2022). *Information Theory: A Tutorial Introduction* (2nd ed.). Sebtel Press.

Tiippana, K. (2014). What is the McGurk effect?. *Frontiers in psychology*, 5, 91962. https://doi.org/10.3389/fpsyg.2014.00725

Tschantz, A., Barca, L., Maisto, D., Buckley, C. L., Seth, A. K., & Pezzulo, G. (2022). Simulating homeostatic, allostatic and goal-directed forms of interoceptive control using active inference. *Biological Psychology*, 169, 108266. https://doi.org/10.1016/j.biopsycho.2022.108266

Von Helmholtz, H., 1866. Concerning the perceptions in general, 3rd ed. *Treatise on Physiological Optics*, Vol. III (translated by J. P. C. Southall 1925 Opt. Soc. Am. Section 26, reprinted New York: Dover, 1962).

William, M. (2018), *Introduction to Statistical Physics.* Summer 2018. Massachusetts Institute of Technology: MIT OpenCouseWare, https://ocw.mit.edu/courses/res-8-010-introduction-to-statistical-physics-summer-2018/

Williams, K. D., Cheung, C. K. T., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the Internet. *Journal of Personality and Social Psychology*, 79(5), 748–762. https://doi.org/10.1037/0022-3514.79.5.748

Appendices

Appendix 1 | Variational Free Energy from Jensen's Inequality

The orthodox way of deriving variational free energy in the active inference literature is by using Jensen's inequality (e.g., see Definition 3 of Sajid et al. (2021)). In short, Jensen's inequality states that for any concave function "the logarithm of an average is always greater than or equal to the average of a logarithm" (Parr, Pezzulo & Frison, 2022, p. 64-65). Translating this definition from natural language to symbolic language gives the following formula:

$$\log \mathbb{E}[x] \geq \mathbb{E}[\log(x)]. \tag{3.1}$$

The surprise function, shown in Figure 2.1 from the Information Theory section, is indeed concave, and therefore a valid function for Jensen's inequality.

We can now derive variational free energy by applying Jensen's inequality. We start with the logarithm of our marginal, which is the surprise term. I will use here the standard negative log evidence form for surprise (i.e., $-\log P(o)$ and not the equivalent but uncommon $\log \frac{1}{P(o)}$ used in the main text). As in the main text, $o \in O$ are the observations, $s \in S$ the hypothesized hidden state, $P(o, s)$ is the generative model, and $Q(s \mid o)$ is the variational distribution that we adjust to approximate the exact conditional posterior distribution $P(s \,|\, o)$:

$$\begin{aligned}
\log P(o) &= -\log \sum_{s \in S} P(o, s) && \text{by def. of marginal (Eq. (2.34))} && (3.2)\\
&= \log \sum_{s \in S} P(o, s) \frac{Q(s \,|\, o)}{Q(s \,|\, o)} && \text{multiplying by 1} && (3.3)\\
&= \log \mathbb{E}_{Q(s \,|\, o)} \left[\frac{P(o, s)}{Q(s \,|\, o)} \right] && \text{expectation notation} && (3.4)\\
&\geq \mathbb{E}_{Q(s \,|\, o)} \left[\log \frac{P(o, s)}{Q(s \,|\, o)} \right] && \text{by Jensen's inequality} && (3.5)\\
&\triangleq -\mathcal{F}\big[Q, o\big]. && && (3.6)
\end{aligned}$$

Positive variational free energy is thus defined as

$$\mathcal{F}\big[Q, o\big] \triangleq -\mathbb{E}_{Q(s \,|\, o)} \left[\log \frac{P(o, s)}{Q(s \,|\, o)} \right] = \mathbb{E}_{Q(s \,|\, o)} \left[\log \frac{Q(s \,|\, o)}{P(o, s)} \right], \tag{3.7}$$

where we flipped the fraction, which led to the absorption of the negative sign via the logarithm quotient rule. The manipulations from Eq. (3.2) to (3.6) also revealed that variational free energy is an upper bound on surprise:

$$\mathcal{F}\big[Q, o\big] \geq -\log P(o). \tag{3.8}$$

To make the meaningful formulations of variational free energy emerge, we need to further manipulate Eq. (3.7):

$$\begin{aligned}
\mathcal{F}\big[Q,o\big] &= \mathbb{E}_{Q(s\,|\,o)}\left[\log\frac{Q(s\,|\,o)}{P(s\,|\,o)P(o)}\right] && \text{by Eq. (2.18)} && (3.9)\\
&= \mathbb{E}_{Q(s\,|\,o)}\left[\log\frac{Q(s\,|\,o)}{P(s\,|\,o)} - \log P(o)\right] && \text{by log. quotient rule} && (3.10)\\
&= \mathbb{E}_{Q(s\,|\,o)}\left[\log\frac{Q(s\,|\,o)}{P(s\,|\,o)}\right] - \mathbb{E}_{Q(s\,|\,o)}\left[\log P(o)\right] && \text{by distr. expectation} && (3.11)\\
&= D_{KL}\big[(Q(s\,|\,o)\,||\,P(s\,|\,o)\big] - \mathbb{E}_{Q(s\,|\,o)}\left[\log P(o)\right] && \text{by def. of KL div., Eq. (2.38)} && (3.12)\\
&= \underbrace{D_{KL}\big[(Q(s\,|\,o)\,||\,P(s\,|\,o)\big]}_{Divergence} - \underbrace{\log P(o)}_{Evidence} && \text{by Eqs. (2.54)-(2.57)} && (3.13)
\end{aligned}$$

We can recognize the Divergence-Evidence formulation in Eq. (3.13). We can extract the Energy-Entropy formulation from Eq. (3.13) as follows:

$$\begin{aligned}
\mathcal{F}\big[Q,o\big] &= D_{KL}\big[(Q(s\,|\,o)\,||\,P(s\,|\,o)\big] - \log P(o) && && (3.14)\\
&= -\mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big] + \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] && && \\
&\qquad + \log P(o) - \log P(o) && \text{by Eq. (2.64)} && (3.15)\\
&= -\mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big] + \mathbb{E}_{Q(s\,|\,o)}\big[\log Q(s\,|\,o)\big] && x - x = 0 && (3.16)\\
&= -\underbrace{\mathbb{E}_{Q(s\,|\,o)}\big[\log P(s,o)\big]}_{Energy} - \underbrace{H\big(Q(s\,|\,o)\big)}_{Entropy} && \text{by Eq. (2.67)} && (3.17)
\end{aligned}$$

As for the Complexity-Accuracy formulation, we can derive it from Eq. (3.7), by substituting

$P(s, o)$ by $P(o \,|\, s)P(s)$.

$$
\begin{aligned}
\mathcal{F}\big[Q, o\big] &= \mathbb{E}_{Q(s\,|\,o)}\left[\log \frac{Q(s\,|\,o)}{P(o, s)}\right] && \text{Eq. (3.7)} \\
&= \mathbb{E}_{Q(s\,|\,o)}\left[\log \frac{Q(s\,|\,o)}{P(o\,|\,s)P(s)}\right] && \text{by Eq. (2.17)} \qquad (3.18) \\
&= \mathbb{E}_{Q(s\,|\,o)}\left[\log \frac{Q(s\,|\,o)}{P(s)} - \log P(o\,|\,s)\right] && \text{by log. quotient rule} \qquad (3.19) \\
&= \mathbb{E}_{Q(s\,|\,o)}\left[\log \frac{Q(s\,|\,o)}{P(s)}\right] - \mathbb{E}_{Q(s\,|\,o)}\left[\log P(o\,|\,s)\right] && \text{by dist. of expect.} \qquad (3.20) \\
&= \underbrace{D_{KL}\big[Q(s\,|\,o)\,||\,P(o\,|\,s)\big]}_{Complexity} - \underbrace{\mathbb{E}_{Q(s\,|\,o)}\big[\log P(o\,|\,s)\big]}_{Accuracy} && \text{by Eq. (2.37)} \qquad (3.21)
\end{aligned}
$$

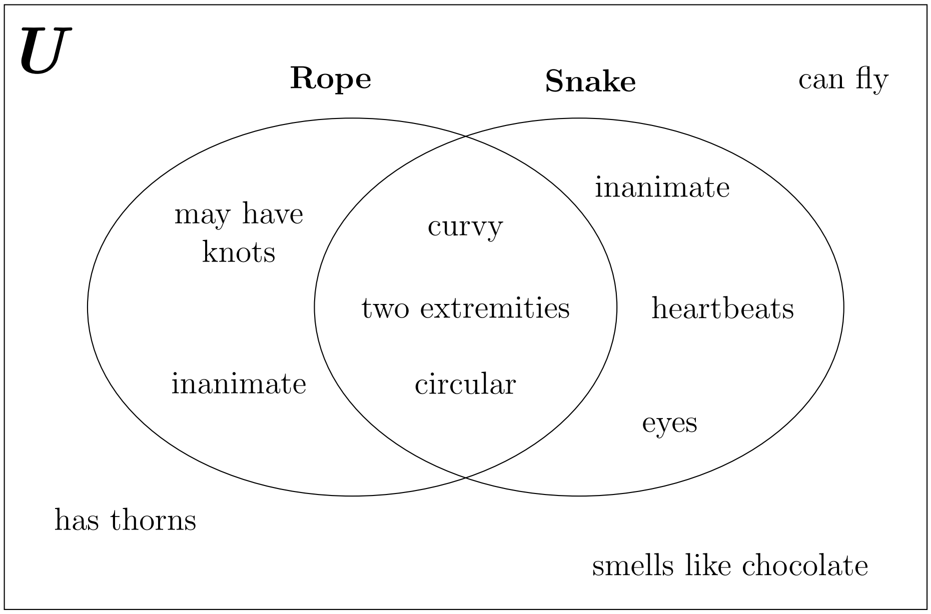
We have thus derived, via Jensen's inequality, all three canonical equations for variational free energy.

Appendix 2 | Modulation of Perception Through Utility-Based Prior Probabilities

In this appendix, we will mathematically illustrate the stick and snake example discussed in the main text (see section 3.3). However, we have substituted sticks with ropes to help explain the distinction between frequency and probability that will intervene later on. We begin with the following Venn diagram of the characteristics of ropes and snakes shown in Figure 3.4.

Figure 3.4

Characteristics (Data) of a Rope and a Snake (Hidden Causes) in the Universe U



Let's define A as the intersection set that has all the characteristics that ropes and snakes have in common, such that

$$A = \{curviness,\ circularity,\ two\, extremitiveness,\ etc.\} \tag{3.22}$$

Any subset $B \subseteq A$, is itself a set of characteristics that ropes and snakes share. (There are

$2^{|A|} - 1$ such sets, where $|A|$ stands for the size of set A. We note that the minus one comes from excluding the empty set.)

The task that a brain faces when it gets a subset B—an intersection set—is to make an inference with insufficient information to discriminate between ropes and snakes. In reality, the set of possible causes of subset B is much bigger than ropes or snakes, but for simplicity's sake, we will work with only these two.

In Bayesian terms, we get the two following expressions:

$$P(\mathit{Rope} \mid B) = \frac{P(B \mid \mathit{Rope}) P(\mathit{Rope})}{P(B)} \tag{3.23}$$

and

$$P(\mathit{Snake} \mid B) = \frac{P(B \mid \mathit{Snake}) P(\mathit{Snake})}{P(B)} \tag{3.24}$$

For now, let us assume that $P(B \mid \mathit{Rope}) = P(B \mid \mathit{Snake}) = k$ is true even though it is extremely likely to be false. We will let go of this assumption later. Additionally, we observe that both $P(B \mid \mathit{Rope})$ and $P(B \mid \mathit{Snake})$ share the same denominator. Thus, the marginal $P(B)$ will not impact the outcome of the inference—it just acts as a normalizing value that ensures the probabilities are between zero and one. Now, with the marginals gone and the two likelihoods fixed at the same value k, on the left-hand side of Bayes' equation, we are left with the priors multiplied by k:

$$P(\mathit{Rope} \mid B) = k \cdot P(\mathit{Rope}) \tag{3.25}$$

$$P(\mathit{Snake} \mid B) = k \cdot P(\mathit{Snake}) \tag{3.26}$$

In this contrived setting, the outcome of the inference rests entirely on the priors. The notion of utility is absorbed into the prior (Parr, Pezzulo & Friston, 2022). As such, we can distinguish between two types of priors: the prior based on probability and the prior based on utility. If we walk into a forest and see ten ropes and one snake, based on frequency, the prior for ropes will be higher than the prior for snakes. However, probability is not only frequency (as nicely argued by Clayton (2022) in *Bernoulli's Fallacy*), but it also includes our more general beliefs about the world. For instance, let us imagine that we are about to enter a forest, and someone tells us that in the past year, they saw one rope and ten snakes. This is our frequency estimate. But should the probability of seeing a rope be so low? Ropes are human-made and we might believe that they are often used in forests. The probability of seeing a rope should probably increase. Then, upon looking online, we learn from a very trustworthy source that the forest is located in a region of the world where there are no snakes. We need to take this newly acquired belief about the world into account. What if the next person we come across tells us that the first person is a pathological liar? The probability of running into a snake just dropped. But what if the third person asked us if we had visited the zoo on the other side of the forest? The probability of running into a snake increases. Etc.

The second type of prior is the one based on utility. Since a snake can throw us out of homeostatic balance, knowing about the nearby presence of a snake has high utility. If the prior absorbs utility into probability, we can end up with a prior for the snake that is higher than the rope's prior, even though the probability of a rope is much higher than the probability of a snake. This situation could be altered if we were desperate to find a rope for a good homeostatic reason, such that according to our beliefs, the utility of a rope would increase drastically. As Parr, Pezzulo and Friston (2022) point out, standard Bayesian decision theory usually keeps the probability of an event and its utility separate. But "[...] this distinction is somewhat superficial, as a utility function can always be rewritten as encoding a prior belief [...]" (p. 207). That said, in our example, we want to illustrate the difference between the two priors and, thus, will keep them apart. It is nonetheless important to remember that they can easily be merged.

Let us denote P_{prob} the prior based on *probability* and P_{util} the prior based on *utility*. In our scenario, we will assume that $P_{prob}\,(Rope) > P_{prob}(Snake)$ and $P_{util}\,(Rope) < P_{util}(Snake)$.

The two first rows of Table 3.2 offer an example of made-up numbers.

Table 3.2

Summary of the Values for Key Bayesian Terms of the Snake and Rope Example

	Snake	Rope	Inference
Probability Prior P_{prob}	10%	90%	Rope
Utility Prior P_{util}	90%	10%	Snake
Likelihood $P(B \mid Cause)$	70%	80%	Rope
Probability-Based Posterior $P_{prob}(Cause \mid B)$	8%	63%	Rope
Utility-Based Posterior $P_{util}(Cause \mid B)$	72%	7%	Snake

Note. The last two rows are not normalized posterior probabilities as we left out the division by the marginal $P(B)$. This explains why the posterior probabilities do not sum up to one.

Using these numbers, we conclude that, given the set B, we would infer that the cause of B is a rope if we use probability priors, but the inference outcome would switch to a snake if we used utility priors.

Now that we have dealt with the simplest case, we can include distinct likelihoods and calculate the Bayesian inference based on probability and based on utility as follows:

$$P_{prob}(Rope \mid B) = P(B \mid Rope)P_{prob}(Rope) = (0.7)(0.90) = 0.63 \tag{3.27}$$

$$P_{prob}(Snake \mid B) = P(B \mid Snake)P_{prob}(Snake) = (0.8)(0.1) = 0.08 \tag{3.28}$$

$$P_{util}(Rope \mid B) = P(B \mid Rope)P_{util}(Rope) = (0.7)(0.1) = 0.07 \tag{3.29}$$

$$P_{util}(Snake \mid B) = P(B \mid Snake)P_{util}(Snake) = (0.8)(0.9) = 0.72 \tag{3.30}$$

The results of the four posterior probabilities are summed up in rows four and five of Table 3.2.

As we see, the inference based on probability leads to the rope as the hidden state (63% > 8%), while the inference based on utility leads to the snake as the hidden state (72% > 7%)

In this example, we assumed that the inference was based on B, but what if the set of stimuli is not B, but B with other characteristics? If one of the characteristics in the set of stimuli from which we make an inference is, say, "has thorns," should that categorically exclude the cause "snake"? I don't know; this would probability be best answered empirically. But from a theoretical point of view, I would speculate that the answer is *no*. Each characteristic in the set from which a final inference is made about the cause of the stimuli received has at least two parameters: a weight and a confidence score. Every characteristic should not be weighted equally. For instance, the colour green might be weighted more heavily for snakes than for ropes. That is, the likelihood of observing the colour green given that the hypothesized cause is a snake might be higher than for a rope. Therefore, the presence of "thorns" in the set of characteristics could be weighted at zero—although this might not be desirable since "thorns" is a rather solid reason to infer that we are not in the presence of a snake.

Then there is the confidence score. One important aspect of the set and subsets of characteristics that we have used in our rope and snake example is that each element of these sets is itself inferred. Backing up in the construction of a percept, like a visual percept, we see that each piece (e.g., curviness) was itself inferred from yet smaller pieces of perception until we get to the bottom of this hierarchical network where we have the raw input of light (visual perception in the brain seems to work a lot like convolutional neural networks (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Lindsay, 2021)). Consequently, they also come with a probability. For instance, "thorns" must be inferred from a set of stimuli. Maybe they are not thorns but only thorn-like patterns. Or they might be thorns, but they belong to a plant located behind the rope-or-snake object. Considering these two attributes, especially the second one—the characteristics' confidence scores—, I believe that the cause "snake" is not excluded from the possible causes in the inference race just because the set from which the inference is made contains elements incompatible with snakes.