universität
wien

MASTERARBEIT / MASTER’S THESIS

Titel der Masterarbeit / Title of the Master‘s Thesis

„Information in the Free Energy Principle of Cognition“

verfasst von / submitted by

Moritz Kriegleder, MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	**UA 066 013**
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	**Masterstudium Joint Degree Programme MEi :CogSci Cognitive Science**
Betreut von / Supervisor:	**Univ.-Prof. Tarja Knuuttila, MSc M.Soc.Sc PhD**

Information in the Free Energy Principle of Cognition

“... a brain is required to write a theory of a brain. From this follows that a theory of the brain, that has any aspirations for completeness, has to account for the writing of this theory. And even more fascinating, the writer of this theory has to account for her or himself.” Foerster, 2003 p.289. “Ethics and Second-Order Cybernetics.” In *Understanding Understanding*, by Heinz von Foerster, 287–304. New York, NY: Springer New York.

Acknowledgments

I would like to thank Professor Knuuttila for supporting my research project and providing me with new perspectives from philosophy of science and computational modeling. The whole ERC group Possible Life shaped my views on modeling in an interdisciplinary context with numerous discussions and coffee breaks, where we were not able to stop talking about our current opinions. Especially, I would like to thank Andrea Loettgers for her encouragement and valuable feedback on drafts of the thesis. Additionally, I want to thank her for the opportunity to visit the California Institute of Technology and offer me a place to stay in Los Angeles for a week.
This thesis would not be possible without the continuous support from my parents, friends and my significant other, who not only listen to my sometimes incomprehensible ramblings about my research but also were there for me through the less motivating times of my studies. Lastly, I am happy to have my two cats around who helped with their random runs over my keyboard and their continuous zoom bombing.

Abstract

In this thesis, I provide an overview of the free energy principle of cognition and the central role that information plays in modelling inference in neuronal systems. Understanding free energy as an informational measure of uncertainty about perceptions helps to clarify the explanatory power of this framework. By relating the free energy principle to the used computational methods and the history of ideas that it is built on, I reinterpret free energy as a relational quantity that we use to describe cognitive processes from the outside. I continue my argument by discussing enactivism and embodied cognition as theories of subjective experience. Linking the ideas of free energy and enactivism, I explore the possibilities and limits of the free energy principle of cognition as a computational model of subjective experience. The goal of this thesis is to clarify the philosophical concepts such as neurophenomenology, the relation between observer and observed, and first-person models to bridge the gap between computational and philosophical approaches to describe experience. I motivate how free energy could potentially be used in computational phenomenology with a relational interpretation of information, contrasting recent arguments about the incompatibility of free energy and enactivism. By understanding information as a relational quantity between cognitive systems and the modeller, we can go beyond the reductionist perspective of information as a description of states of physical systems. In the conclusion, I highlight the importance of how and why we use information as a core concept to model cognitive systems and its central role in the description of subjective experience.

Zusammenfassung

In dieser Arbeit gebe ich einen Überblick über das Free Energy Principle in der Kognitionswissenschaft und die zentrale Rolle, die Information in der Modellierung von Inferenz in neuroyalen Systemen spielt. Ein philosophisches Verständnis von Free Energy als Informationsmaß der Unsicherheit von Wahrnehmung verdeutlicht das erklärende Potential der Theorie. Ich verbinde die mathematischen Methoden mit der Ideengeschichte um Free Energy als relationale Größe zu beschreiben, die wir benutzen, um kognitive Prozesse von außen zu modellieren. Weiter diskutiere ich Enaktivismus und Embodied Cognition als Theorien von subjektiver Erfahrung. Im Spannungsfeld von Free Energy und Enaktvismus analysiere ich die Möglichkeiten und Limitierungen vom Free Energy Principle als mathematisches Modell von subjektiver Erfahrung. Das Ziel der Arbeit besteht darin, die philosophischen Konzepte zu identifizieren, die eine Brücke zwischen quantitativen und philosophischen Aspekten bauen können, um subjektive Erfahrung zu beschreiben. Dabei befasse ich mich im speziellen mit Neurophänomenologie, der Relation zwischen Beobachter und Beobachtetem und Modellen aus der First-Person-Perspektive. Dies könnte den Weg ebnen das Free Energy Principle zu einer Computerphänomenologie zu erweitern, die eine mathematische Beschreibung von phänomenologischen Observablen ermöglicht. In der conclusio erläutere ich die Wichtigkeit des Informationsbegriffs in Modellen von kognitiven Systemen and die zentrale Rolle, die Information in der Beschreibung von subjektiver Erfahrung spielt.

1. Introduction

How do we model cognition? The last century has seen intense discussions between neuroscientific, computational, and philosophical models of the mind that have also spread into everyday life through the increasing availability of computers, artificial intelligence, and better brain imaging techniques. While cognition itself has been the main focus in these discussions, the understanding of the modeling tools is often overlooked in computationally sophisticated theories of the mind. In this thesis, I study the tools and the underlying philosophical commitments related to the mathematical concepts used to explain the brain. Moreover, I discuss the analogies between cognition and logical operations which sparked todays information technology. Specifically, I examine a recent model of the mind that casts cognitive agents as minimizers of uncertainty about their environments, or maximizers of mutual information. This information is quantified as cognitive free energy in analogy to Helmholtz's free energy used in physics and chemistry. This transdisciplinary sharing of modeling tools comes with a variety of interpretations and ontological commitments, which I disentangle using information theory and the philosophy of modeling.

The free energy principle is a current computational theory of the mind and life (Allen & Friston, 2018), which uses Bayesian inference to describe cognitive behavior on different scales from simple neuronal networks to complex human behavior. My thesis focuses on the novelty and possible shortcomings of the framework. How can it be successfully applied and where does it reach its limits? Its proper philosophical foundation is currently a source of heated debate (Colombo & Wright 2018, Andrews 2021, Bruineberg et al. 2022) and my research aims to shine a light on some of the aspects that enable the philosophical and scientific grounding of free energy in the philosophy of mind and the description of subjective experience. The role of information and representation are two crucial concepts for understanding the ontological and epistemological commitments that this computational model of the mind entails. In the free energy principle, different modeling approaches are used together to build an interdisciplinary perspective on what the fundamental quantifiable aspects of cognition and inference are. I will contrast this quantitative perspective with a contemporary understanding of cognition as embodied and enacted dynamical process, because this is where the mathematical intricacies meet with the philosophical bigger picture. Such a perspective is also part of a recent research program aiming to bridge the gap between top-down and bottom up perspectives in consciousness research (Sandved-Smith et al. 2021, Ramstead et al. 2022). The future of the free energy framework depends on how we understand the difference between generative and recognition models, what is internal and external to the agent, and how we think about the borders of these systems. What is informative to and of the agent, is a highly relational question that I want to highlight with this thesis.

The structure of this thesis reflects my analysis of free energy from a mathematical, philosophical, and neuroscientific perspectives. In section 2, I explain the history of free energy in statistical physics and how it migrated over artificial intelligence research into neuroscience. The scientific transfer ties the involved ideas from information theory, neuroscience, and physics together, which at first glance often appear completely unconnected. I suggest that understanding the diverse backgrounds on which Friston and colleagues' work is built is necessary for interpreting information and free energy as they are used in the framework (Friston 2009, 2010). To provide a practice-oriented analysis, it is also vital to discuss the direct applications of free energy in describing the perception and the behavior of cognitive agents. Therefore, I introduce a simple perceptual simulation in section 2.5 to clarify the mathematical backbone of the theory and to visualize one possible application. This is essential in moving forward with a practice-oriented philosophical interpretation of the measures employed in these cognitive models developed from the free energy principle.

In section 3, I will provide an overview of the uses of information in philosophical and mathematical approaches to cognition, with a special focus on Luciano Floridi's classification of information and using free energy modeling as a concrete case study (Floridi 1999, 2002, 2003). This opens up some deeper questions about the nature of information, but to keep the thesis concise, the focus will be on the role of cognitive information and its use in contemporary neuroscience. In my analysis, I limit myself to Shannon, Floridi, and Friston in discussing the role of free energy as an informational measure. My goal is not to find a definitive interpretation of information in all applications but to interpret how information is used in the free energy framework and to suggest a re-interpretation that could open up new ways for a computational and phenomenological models of cognition.

Having defined my scope of discussing the relation of information and cognition in section 3, I move on to present an enactivist reading of information in the free energy principle in section 4. This bridges first- and third-person perspectives in cognitive science and criticizes literalist accounts of free energy as information. Instead of the habitual merging of scientific model-based inference concerning cognition with inference that our assumed mental models supposedly employ, I urge cautionary differentiation between making models of systems and describing systems with models. The scientists' use of models to make sense of targets has to be separated from the assumption that cognitive systems can be described as making models about their environment. Of course, both senses of modeling have significant overlap not only because of the notion of model, but my discussion aims to clarify what ontological commitments these different uses of the notion of a model possess. I argue that an instrumentalist reading of information opens up the possibility to reconnect recent philosophical ideas about cognition, such as embodiment and enactivism, with mathematical progress in neuroscience.

In the conclusion I argue for a relational concept of information that would be necessary to interpret the free energy framework as compatible with embodied and enacted cognition. This includes a pluralist or even instrumentalist reading of cognitive information but allows us to go beyond the classic definitions of information in computer science and neuroscience. The necessity of including first- and third-person perspectives into a computational theory of the mind shows the limits of the current formulation of the free energy principle, since many philosophical underpinnings of the concepts are ambivalent and differ fundamentally in their usage. I defend the perspective that all information, especially cognitive information, has to be understood in a tripartite way as syntactic, semantic, and relational.

In my understanding of relational information - or information as a model of relations between cognitive systems and observers - cognitive aspects go hand-in-hand and cannot be reduced to any of these dimensions. By reinterpreting information as it is used in the free energy principle, I show that enactivism, phenomenology, and free energy are more compatible than recently argued. Applying a concrete philosophical classification to the role of information and representation could help to formulate free energy to become a proper candidate for a theory of computational phenomenology.

2. The Free Energy Principle and Cognition

The free energy principle aims to capture the active process of making sense of our direct environment, by integrating cognitive information in a multi-level system of models about sensory perception in a set of mathematical equations. The level structure fits well with what is known about the structure of the cortex and explains the origin of modularity in the brain. For example, in visual processing, every level from the retina up to the visual cortex minimizes uncertainty about the inputs it gets from the previous level. The free energy principle goes beyond former approaches to formalize the computations of cognitive agents by highlighting the active inference that is necessary to bring forth a perceived world instead of merely representing external states in a passive information processing.

2.1 What Free Energy says About the Brain

Free energy, as it is used in cognitive science, captures the difference between predictions of the environment and the perceived sensations of a cognitive agent. All cognitive agents aim to minimize free energy by changing their perception and action (Friston 2009, 2010). This is a simplified and compressed version of how the Free Energy Principle (FEP) applied to neuronal systems can act as a guide to understand the fundamentals without diving deeper into the mathematical framework. We could, for example, envision coming into a dark room while looking for a black cat. Depending on our confidence in our prediction of where the cat is located, we could either wait until we adjust to the darkness and change our perception, or we could take action and find the light switch. Both approaches minimize the difference between our predictions and sensations by either adjusting our model to the dark circumstances or changing the light conditions of the room.

Another important concept that I will discuss in mathematical detail in section 2.5 becomes clear in this example. To navigate the dark room, we reduce how surprising the environment appears to us. We construct information about it by improving our model or changing it to better fit our predictions. In the FEP, this measure of uncertainty is called surprisal, and minimising surprisal is the underlying goal of why to minimize free energy.

Karl Friston and Klaas Stephan laid the groundwork for the FEP in 2007 (Friston and Stephan 2007) and a number of neuroscientists (Gershman 2019), artificial intelligence researchers (Ramstead 2021), and philosophers (Colombo & Wright 2018, Andrews 2021) have since debated and refined these ideas further. A well-cited Nature article from 2010 even asks the seminal question of whether the FEP can guide as a unified brain theory that forms the basis for all phenomena of perception, action, and experience (Friston 2010). But how could such a recent idea take over fundamental debates of cognitive science in only a bit over a decade?

Starting from the picture of Hermann von Helmholtz's view of the brain as an inference machine, Karl Friston and his collaborators draw on ideas from information theory, statistical physics, and neuroscience to formulate a set of equations that formalize the behaviour of cognitive systems. It is a physics-first approach to modeling the mind (Colombo & Wright 2018), that enables a strict quantification of a diverse set of cognitive phenomena. But such approaches may lack the complexity that different disciplines studying the mind have uncovered already. In this thesis, I will ask what the advantages of such a physics-based model are and what limitations the apparent unification of cognition as free energy minimization entails philosophically. This question is important for the future direction of the cognitive sciences because we have to decide how much we are willing to trade off the diversity of cognitive models for the purpose of unification.

2.2 History of Computational Models of the Mind

Before going into the mathematical and philosophical details, I will take a look at the history of computational models of the mind. Recent attempts to understand cognition through mathematical modeling follow a long line of research that aims to bring computation and mental processing together. The foundations of computer science derive from hypotheses about how computation takes place in our head and abstraction of these supposed processes lead to the invention of the first computing machines. Conversely, advances in artificial intelligence and neuroimaging boosted the efforts to find a computational basis of how the mind works. But what can we learn from treating cognitive systems as computing machines and what do these models leave out?

The search for a physical basis of neuronal energy and potential goes back to the ideas of physiologist Hermann von Helmholtz. Interested in the conservation laws of physics and brain physiology, von Helmholtz studied the mind as a set of electrical processes happening in the brain and body (Helmholtz 1882). In his research, he wondered if mental activity could be represented by a type of neuronal energy that follows the typical behavior of physical energy as described by electro- and thermodynamics. To describe the link between perceived sensations and behavior he invented the concept of unconscious inference. According to his idea, when we perceive light from the environment, we do not passively process this information to create the image we see with our eyes. Through previous experience and learning, we are able to construct an image that makes sense to us. The visual process of seeing is, therefore, not only detecting light with the retina but through unconscious inference, ambiguous retinal images turn into active perceptions of objects. We do not need conscious thought to make sense of our environment, we do it already unconsciously by changing the way we look at it. This observation coupled with the idea that our prediction models are encoded probabilistically leads to the Bayesian brain hypothesis, which is fundamental to the development of the FEP (Kersten et al. 2004).

In the 1940s, physiologist Warren McCulloch and logician Walter Pitts investigated the nervous system and derived a calculus for a simplified version of neuronal functioning (McCulloch & Pitts 1943). In the influential 1943 paper “A logical calculus of the ideas immanent in nervous activity” they used early neurophysiological evidence to suggest that neuronal networks can compute complex logical problems because their network structure is analogous to simple logical operators. By assuming that neurons either fire or do not fire (encode binary values), that activity of neighboring neurons add up to excite each other, that synaptic delay is the significant time scale, that inhibitory synapses prevent excitation, and that the network structure does not change with time, they formulated a calculus to solve logical tasks. The connection between computation and thinking was striking, but interestingly the article does not refer to neuroscientific literature. It only contains three references to works from logic, such as the *Principia Mathematica* (Whitehead & Russell 1963). This could be read as foreshadowing the focus of the later-founded field of artificial intelligence, which rather takes ideas from logic and mathematics than from real neuronal architecture. But even if this focus is one-sided, the influence of McCulloch and Pitts's work on neuroscience and the cognitive sciences is still profound.

Taking inspiration from self-organization as defined by cyberneticist William Ross Ashby (1947), the free energy principle builds on a rich and transdisciplinary field that studied control, feedback, modeling, and many other related phenomena in the middle of the 20th century. Cybernetics is usually associated with the works of Norbert Wiener, William Ashby, Ernst von Glasersfeld, and Frank Rosenblatt, who studied various systems of perception and action such as detection devices and artificial models of the nervous system. The perceptron, one of the first computers that imitated the behavior of neuronal networks, was a set of circuits that got input from a camera, adjusted the interaction of separate logical operations, and produced an output that indicated if a specific object has been detected. Even though the perceptron sparked hope that machines can think like humans, the results were limited by computing power and the single layer of artificial neurons (Rosenblatt 1958). Later steps in artificial intelligence such as backpropagation algorithms for multi-layer neuronal networks have shown to be applicable to a wider range of tasks, such as translation, labeling of images, and pattern recognition, but still, the question remains whether these models somehow reflect biological processes or are just effective tools for certain applications that brain can also compute (Rumelhart 1986).

In contrast to neuronal networks, the Bayesian brain hypothesis models higher cognitive function without direct application to the behavior of single neurons. It constitutes a framework that describes how beliefs are encoded and how they change through novel sensory evidence on a meta-level and not on the neuronal level. First of all, beliefs are described as probability distributions over possible states of the environment. These beliefs include everything from low-level beliefs such as perception up to high-level beliefs with emotional content (Demekas et al. 2020) To go back to our previous example, the position of a cat in a dark room could be unclear but our intuition tells us, that some places are more likely than others.

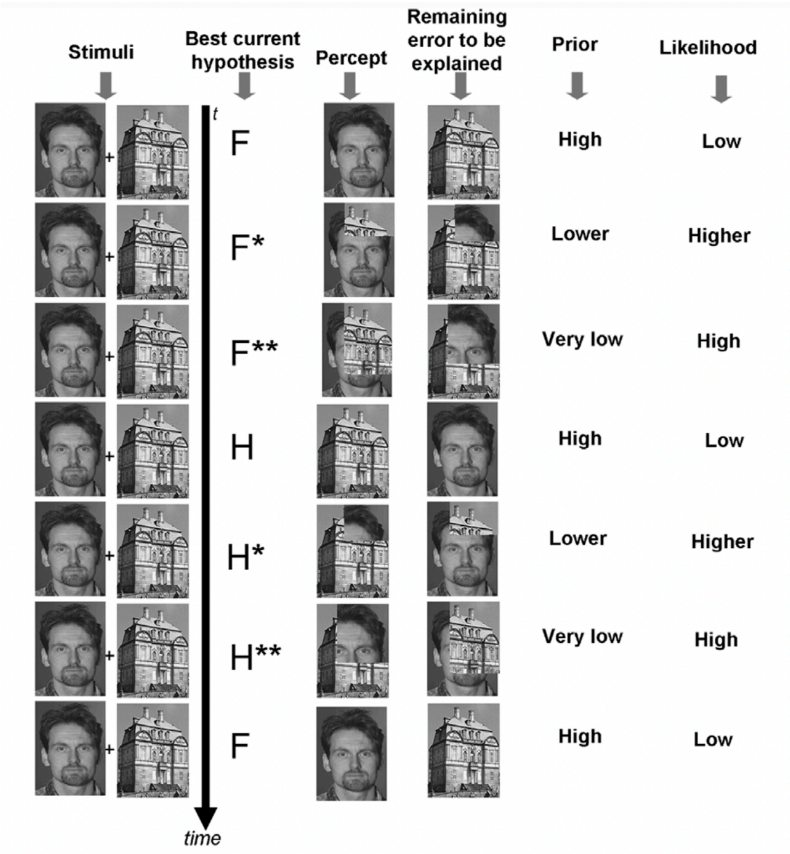

Figure 2.2.1: Binocular Rivalry explained by Bayesian Perception. *When presented with a face on the left and a house on the right eye, we tend to alternate between perceiving one or the other. Since a superposition of both images is unlikely to be seen in the wild, the brain is trying to explain away the conflicting parts of the perception. We form a hypothesis F that what we see is a face and the image of the house is explained away by our mental model quantified by the prior. Since the error on the right eye remains given by the low likelihood of the model explaining the observation, we start to alter our hypothesis to accommodate for the image of the house F*. The face-house hypothesis F** is not plausible for everyday situations so our perception flips to only seeing the house corresponding to hypothesis H. But now the face image remains as error, so the circular dynamics start to flip between F and H without being able to explain the observation in line with our everyday experience and reach a stable state. From: Hohwy et al. 2008*

Maybe we remember where we have seen it last or we are convinced that it is, for example, not in the trash can. All these beliefs about the location based on previous experience are, according to the Bayesian brain hypothesis, encoded in a probability distribution over possibilities. Finding a cat in a box is very likely, but finding it hanging from the ceiling is rather unlikely. This probability distribution is called the prior and represents our beliefs and predictions. The framework tells us now exactly how our beliefs are updated if we switch on the light and construct new information from visual perceptions of the bright room. This is mathematically implemented with Bayes rule, which says that the posterior belief is proportional to the prior belief and the likelihood to perceive each possible state. The Bayesian brain hypothesis does not only describe how we find cats in dark rooms, it has also been applied to perceptual puzzles like binocular rivalry (Hohwy et al. 2008).

When the left and the right eye are presented with different images, we never seem to see two overlapping images. For example, when the left eye looks at a picture of a face and the right eye is presented with a house, experimental data suggests that we never see a combination of both but the perception alters between the face and the house. This phenomenon has been known since the 16th century and is known in visual psychology as binocular rivalry (Wade 1996). Passive computational accounts of perception would imply that we see a direct representation of the stimuli, but Bayesian brain accounts explain the alternation of perceptions with the construction of a mental model and its direct influence on subjective experience. In a nutshell, we don’t see superpositions of faces and houses in the wild so we explain away one part of the conflicting observation. If we see the face, we explain away the house but since the unexplained image on the other side stays there we shift our attention to it and start to explain away the face. This circular process continues to go on without us being able to solve it leading to a bistable perception of both images sequentially.

But to cope with the complexity of daily life one probability distribution is, unfortunately, not enough. To apply to basic tasks such as image recognition, sound identification, and sensory-motor learning the Bayesian updating has to be structured in multiple layers that influence each other continuously. A set of sometimes even contradicting beliefs is constantly being updated while we try to identify an object. This constitutes a multi-layered generative model, that generates a model of causal relationships between different variables that describe the object. Such as how different lighting conditions or our point of view would change our perception of a cat. Additionally, task specification is important in making sense of a visual image.

Computationally, it is not favorable to analyse every single detail of the image, if our goal is just to tell if we see a cat or a dog. On the other hand, if the task involves analyzing the species of cat we are looking at, the identification of two ears, whiskers, and a tail is not sufficient. Other variables, such as color, type of fur, or size, become important, that have been neglected before. The last important point about Bayesian inference is to understand perception as an inverse solution to the generative model (Kersten et al. 2004). The brain generates expectations of how the object looks and contrasts it with the incoming sensations to construct the image we see.

Another important ingredient to understanding the free energy principle and its application to cognitive and other self-organizing biological systems is predictive coding. Often credited to the extensive analysis of the visual cortex by Rao and Ballard (1999), the concept of predictive coding has been discussed in psychology already before (Srinivasan et al. 1982). In a nutshell, the idea states that experiences, such as visual images, are constructed by taking the difference between predictions and sensations. A well-known stimulus is analyzed more coarsely because the prediction is already detailed. Taking in the fine structure would just create more redundant information about something already familiar. By highlighting the importance of both top-down predictions and bottom-up sensations, predictive coding could end the old debate concerning the question of whether perceptions are passively received from the world (an exclusively bottom-up perspective) or actively constructed through the mind (an exclusively top-down perspective). Experiences happen, according to the theory, where predictions and sensations meet.

Empirical evidence from the structure and function of the visual cortex shows that different neuronal structures could be involved in making predictions and passing on sensations (Rao and Ballard 1999). But since the exact neuronal mechanism is still unclear, predictive coding remains a high-level cognitive theory of information processing. The application to the brain and a wider range of cognitive processes has developed rapidly in the last two decades and contemporary theories of predictive brains are nowadays called predictive processing (Clark 2013). As discussed before, the search for a physical basis for the computational models discussed by unconscious inference, the Bayesian brain, and predictive coding led to the first formulation of the Free Energy Principle of cognition. Previously, however, free energy was already borrowed from statistical physics and applied to neuronal networks by computer scientist and “Godfather of AI” Geoffrey Hinton and computer scientist Drew von Camp (1993). Later works also made the link to Helmholtz’s ideas explicit (Hinton and Zemel 1994) and coined a computational model for learning by free energy minimization the “Helmholtz Machine” (Dayan et al. 1995). These methods were used computationally, but it took more than a decade until Karl Friston and Klaus Stephan took them seriously as mechanisms directly explaining cognitive processes (2007). A highly-cited review article by Friston has cemented the FEP as a contender for a unified brain theory (2010).

2.3 Free Energy Definitions for Physics and Cognition

Free energy was first introduced in thermodynamics by Hermann von Helmholtz to explain chemical processes (1882). With this essay and the following works, he made important contributions ~~contributed important work~~ to the interface of chemistry and physics. Helmholtz used free energy to describe how much work a closed thermodynamical system can perform without temperature changes. The symbol F denotes the quantity of ~~and~~ free energy,which is the result of the ~~is made of the whole~~ internal energy of the system U minus the product of temperature T and entropy S. This means that highly entropic systems, with a large number of possible states, have lower free energy than highly ordered systems with low entropy.

$$F = U - TS \tag{1}$$

Helmholtz used this definition to differentiate between free energy, which can be obtained from closed thermodynamic system ~~used to do work~~, and bound energy which cannot be extracted from the system. Therefore, free energy is useful to describe the available energy for chemical and biological processes. He also noted that spontaneous transitions from one thermal state to another can only occur in the direction of minimising free energy. According to the second law of thermodynamics, a closed system in equilibrium is always in the state where the entropy S is maximized. But biological and cognitive systems are thermodynamically open systems because they are in continuous exchange with their environment. An isolated biological system disintegrates quickly. Ideally, free energy can be applied to open systems if temperature and volume are held constant (Chipot and Pohorille 2007). This description is useful for an organism in a stable environment with a heat bath, some kind of reservoir at near-constant temperature such as for a small fish swimming in the ocean for example. In these conditions, the second law of thermodynamics leads to the minimum principle for free energy: the system transitions into the state, where free energy is minimized.

But what does change if we use a description of thermodynamics for cognitive systems that operate in much more complex and dynamic environments and have a diverse internal structure in comparison to usual objects of thermodynamics, such as heat baths, pistons, and engines? Friston and Stephan highlight the unique properties of organisms in their paper “Free-Energy and the brain” (2007):

“ *… biological systems are more than simply dissipative self-organizing systems. They can negotiate a changing or non-stationary environment in a way that allows them to endure over substantial periods of time. This means that they avoid phase transitions that would otherwise change their physical structure. A key aspect of biological systems is that they act upon the environment to change their position within it, or relation to it, in a way that precludes extremes of temperature, pressure or other external fields.*“ Friston and Stephan 2007, p.5

So, the agency of biological systems allows them to not only react to thermodynamic changes but to actively construct and maintain an environment that makes sense for them. Sensing and sense-making both play an important role in explaining cognition with the free energy principle. I will describe the philosophical implications of this model in Chapter 4 in detail. But to understand the scope of the theory, we need to compare the use of Helmholtz free energy and free energy as cognitive information first.

The free energy principle of cognition uses mathematical tools of thermodynamics but does not relate it to directly measurable physical quantities such as heat. It uses the interpretation of entropy as an information-theoretic measure of uncertainty similar to Claude Shannon in his foundational works of information theory (Shannon 1948, Shannon and Weaver 1949). Here, entropy is a measure of uncertainty. For example, a probability distribution that describes the possible location of our black cat in a dark room has low entropy, when the location of the cat is predictable. On the other hand, if we do not have any idea of where the cat is sitting our probability distribution would be highly entropic.

Another important measure that diverges from any use in thermodynamics is surprisal or also called self-information (Friston 2010). Surprisal measures how expected a sensation is. Mathematically it is given by the negative log-probability of a sensation S. It is, therefore, a measure of how good the information is that the cognitive agent has about future sensations.

$$surprisal = -\ln p(S) \tag{2}$$

In this formulation, entropy can then be derived from the average surprisal of an agent. If we reconsider the cat's affinity for laying in boxes, it would be not surprising to find our cat in the box. But if the cat instead regularly could be found in the bathtub our prediction state would be higher in entropy. To be clear, the probability distribution describes our beliefs about the cat's whereabouts. It is relative information between the cat owner that has gathered information about the animal and the cat, that has a shared history of behavior with the owner. Therefore, the owner's belief could be different from what a first-time visitor would find surprising when observing the cat.

Analogous to Helmholtz Free Energy F = U - TS, Friston postulates an equation for cognitive free energy by assuming that the temperature stays constant because of the tendency of biological systems to stay in homeostasis. Due to homeostasis, we are only interested in the internal energy of the cognitive system and its information entropy, the average surprisal over time. The Fristonian version of the free energy equation looks like

$$F = \int dCq(C)E(C,S) + \int dCq(C)lnq(C), \tag{3}$$

where F is the free energy, the first integral represents the average energy analogous to U, and the second part is the negative entropy corresponding to -TS in the Helmholtz equation (Buckley et al. 2017). Before I explain the meaning of the individual variables, I would like to explain the idea behind the recognition density q(C) and generation density p(C,S), which are crucial for the understanding of free energy as a model of cognition. Both densities are dependent on the causes C and the generation density is additionally dependent on the sensations S.

To minimize free energy, cognitive systems construct and update a probabilistic model of their environment. According to the principle, these systems use two separate probability densities that fulfill different purposes in belief updating. First, we have the recognition density q(C) which consists of characteristic variables that are used to represent the causes that shape the environment. This description of the environment includes all likely values that the characteristic variables may possess. To employ our cat example, we could think of these characteristic variables to be the lighting conditions of the room or the angles from which we look at the room. Now, when we receive new sensations such as a change in light, we update our recognition density to reflect what became less or more likely with the new sensory evidence.

Second, we employ the generation density p(C,S) which is our model of how our sensory states change corresponding to environmental states. The assumed link between sensation and environment is crucial to how we update our beliefs. The generation density p(C,S) determines how we change our recognition density q(C) if we receive new sensations S. Philosophically, the variable C plays the role of epistemological causes. These are the causes that the agent assumes to cause the conditions of the environment. They do not play any ontological role or have to possess real causal power. They are only used by the cognizer to explain the situation in which they are in. For example, rain dances can be effective causes for weather changes to some agents, and to others they are not. Even though, surprisal can be quantified as the negative log-likelihood $-lnp(S)$ it is not computationally ~~not~~ tractable in biological situations, since the cognitive system would need to have knowledge of the probabilities of all possible states of the environment.

"…agents cannot evaluate surprise directly; this would entail knowing all the hidden states of the world causing sensory input. However, an agent can avoid surprising exchanges with the world if it minimizes its free-energy because free energy is always bigger than surprise." Friston 2009 p. 294

Free energy, in this formulation, acts as an efficient estimation of what could be a surprising sensation. It is an approximation that is necessary because we do not have access to all possible sensations. Free energy is a useful heuristic for limited beings.

2.4 Cognition as Free Energy Minimisation

As discussed in the previous chapter, the free energy principle gives a mathematical model that defines cognitive systems as systems that aim to minimize uncertainty about sensations caused by the environment. But what cognitive strategies can be employed to minimize the corresponding free energy?

In the generative model that the cognitive agent uses to predict sensations, all states are characterized as either internal or external (Ramstead et al. 2020). Between them, we have sensory states that relate environmental changes to experience and action that makes use of the body to change external states. But both processes are not unidirectional. They are coupled through a nested feedback loop that constructs the internal and the external states simultaneously. This is important to keep in mind because the notion of external states could lure some to think of objective states independent of the observer. But external, in this case, refers only to the assumptions that are not about the mind and the body of the observer. The border between internal and external states is, therefore, dynamic and to model this distinction the free energy principle employs Markov blankets, a concept from statistics to estimate how strongly correlated nodes of a large network are. A Markov blanket between external and internal states indicates that brain states can only indirectly influence external states in contrast to other internal brain states.

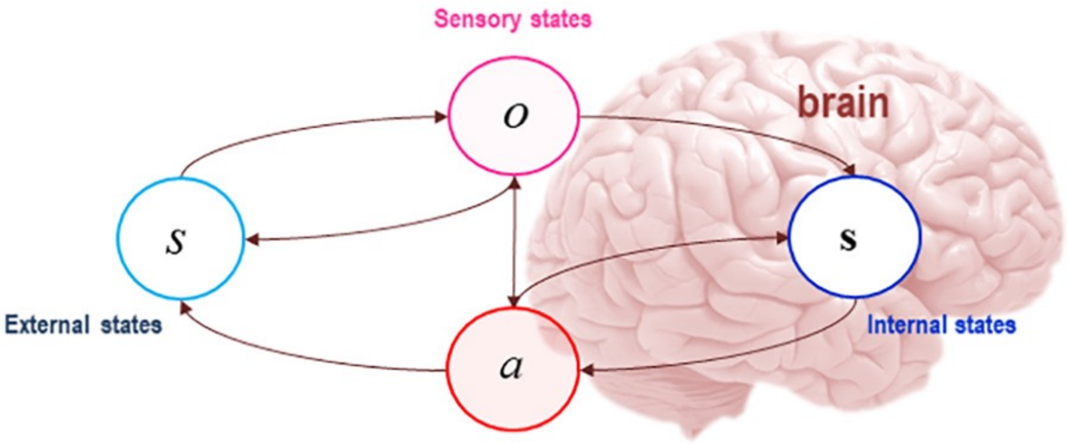

Figure 2.4.1 Interaction Diagram of the Free Energy Principle. *The external states **S** and the internal states **s** are coupled through sensory states **o** and action **a**. The interface between internal and external is described by a Markov blanket, that models how internal states can only indirectly influence external states and vice versa. From: Ramstead et al. 2020*

Finally, we can model our two possibilities in the example of finding a cat in a dark room. The first strategy is to change our internal state and get used to the faint light coming from outside of the room. Maybe the moon is shining through the window. Instead of changing the room to our prediction (most of the time we spend in well-lit rooms), we adapt to new circumstances. We could either wait for our eyes to adjust to the faint light or use our hearing to localize the cat. In this case, we change our internal model to pay attention to weaker sensations. For changes in perception, free energy can be rewritten:

$$\textbf{Perception: F} = \text{Divergence} + \text{Surprisal} \tag{3}$$

Minimising **F** can then be achieved by using perception to reduce the divergence of the probability density or surprisal, which makes sensations caused by the environment more predictable. On the other side of the Markov blanket lies the possibility to reduce the complexity or increase the accuracy of the model by changing the environment.

$$\textbf{Action: F} = \text{Complexity - Accuracy} \tag{4}$$

Standing in the dark room entails limiting the search for the black cat to its favorite spots or making the search more accurate by incorporating sounds such as purring and the direction from where they come from. Friston himself summarises these possibilities for free energy minimization in his characteristic dense style:

"*(i) agents resist a natural tendency to disorder by minimizing a free-energy bound on surprise; (ii) this entails acting on the environment to avoid surprises, which (iii) rests on making Bayesian inferences about the world. In this view, the Bayesian brain ceases to be a hypothesis, it is mandated by the free-energy principle; free energy is not used to finesse perception, perceptual inference is necessary to minimize free-energy." (Friston 2009 p. 295)*

How does this computational model then relate to cognitive behavior? Since we are discussing a physics-based principle only, there are no direct predictions that can be experimentally tested in the brain. Besides stating that cognitive systems minimize free energy, there would need to be a specified mapping to neuronal mechanisms. How is free energy represented in the brain and what mechanisms can we observe through measurement? Recent research focuses on the role of active inference, the consequence of the free energy principle that provides an integrated framework of action and perception. In active inference, action is an integral part of sense-making and cognitive models do not arise from processing external information but by actively sampling an environment and testing cognitive assumptions continuously (Friston et al. 2015). Previous cognitivist theories assumed perception and action to be separated processes that follow computations over external information as perceived by the agent.

So active inference can be experimentally tested through its connection to predictive processing and laboratory studies. One hypothesis assumes that prediction error is transmitted in the brain by dopamine concentrations (FitzGerald et al. 2015). Different cortical areas correspond to the different layers of the generative model proposed by the free energy principle and minimizing prediction error between those areas fulfills the dynamics described by the free energy framework. Simulations of learning under free energy were able to explain how dopamine influences learning efficiency and reproduced experimental results from studies demonstrating dopamine stimulation as a reward (Friston et al. 2012).

The other strong indication of the usefulness of the free energy principle comes from a famous psychological experiment that investigates the concept of body ownership and agency. The Rubber Hand Illusion is an intriguing phenomenon that can be used to test self-representation and its consequences in mental models of the body (Apps & Tsakiris 2014). The experimental design consists of an observer looking at a rubber hand, that gets the same tactile sensations as her real hand. For example, the observer's hand is hidden and stimulated with a brush while simultaneously the experimenter brushes the visible rubber hand in the same way. This leads the participant to feel the touch of the rubber hand, an out-of-body experience that includes the fake hand in the self-model. While many variations of this experiment have been conducted, such as virtual reality studies and full-body illusions, there is no complete theoretical model of it yet (Limanowski & Blankenburg 2013). Through the free energy principle, we can understand the rubber hand illusion as a misleading integration of different sensations into the generative model of our physical self. Feeling the touch and seeing the experimenter touch a rubber hand simultaneously fools our self-model into incorporating the new body part. Surprisal of the correlated sensations is easily minimized by thinking the rubber hand and the real hand are the same thing. This model is way less complex than taking into account the unnatural circumstances of a laboratory setting, the experimenters' intention to fool you, and the mere coincidence of two stimuli happening simultaneously. So we fall for the "simple" explanation.

Still, the experimental evidence is starting to build up and while the discussed phenomena can be described by other theories as well, the unifying power of free energy explanations makes it an interesting candidate for a computational model of the mind (Carhart-Harris et al., 2010, 2019). With its strict mathematical formulation and physics-first approach, it becomes a link between a diverse set of disciplines such as neuroscience, artificial intelligence, psychology, and biology – even though its philosophical implications are not yet clarified systematically (Colombo & Wright 2018). In the next chapter, I analyze the consequences of reading free energy as a measure of information, highlight its relation to philosophical discussions of information theory, and compare the use of the model in an interdisciplinary context.

2.5 Example: Model of a Perceiving Agent

Since the free energy framework is built on a diverse interdisciplinary background and the differences to other approaches from Bayesian cognitive science are hard to draw clearly, I will explain the Bayesian part of perception and the free energy minimization process with a simple model of a perceiving agent. This simulation demonstrates the fundamental mathematical machinery of a cognitive task, but still uses the same equations that would be needed to describe more complex everyday cognitive processes. I aim to demonstrate the computational apparatus without getting lost in the mathematical intricacies. I build on this intuitive understanding to analyze the philosophical implications concerning the actual use of the formalism. It is not only important to analyze where the ideas come from but also how they are actually applied to concrete examples. Only using the framework will make it testable, since the general statement of the free energy principle itself is unfalsifiable (Friston 2010).

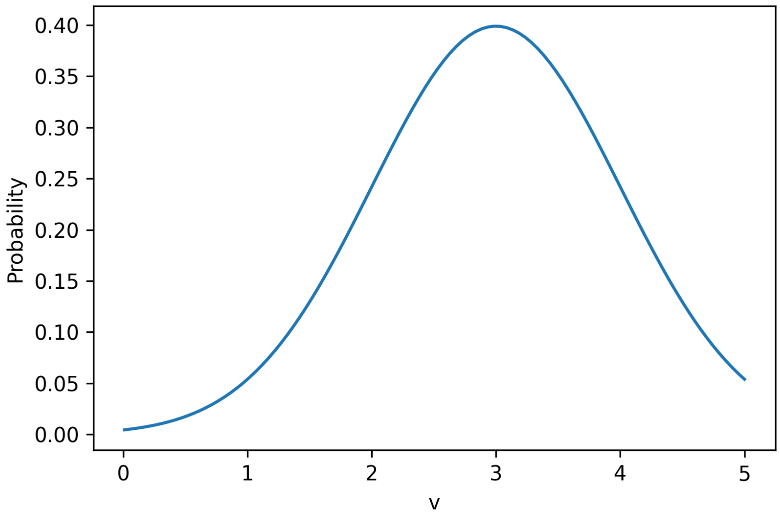

Figure 2.5.1. Prior expectation of perceiving an object. *An agent is perceiving a squared object of diameter v. Let's assume a vegan is looking for a squared falafel with only one receptor sensitive to light. The probability distribution shows the prior belief of the perceiver of possible values for v. The most probable diameter of the object according to the prior expectation is around 3 cm with a variance of the normal distribution that quantifies how sure the agent is about their prior belief. Therefore, we describe a belief as a probability distribution with a most likely value for the size of the falafel as 3 cm and an uncertainty of this maximum.*

This simple example is built on the ideas from Bogacz (2017), where a Bayesian agent is looking for an object with a single photosensitive receptor. I modeled this Bayesian inference about the size of the object as a perceiving agent looking for a square-shaped falafel of about 3 cm diameter, so a typical task of a vegan. My python code used to calculate these values can be found in the appendix of this thesis.

Bayesian belief updating describes how predictions change when new information is acquired. In our example, the vegan agent has a prior expectation of falafel usually being around 3 cm in diameter. But since they know how the sizes of falafel can vary, their expectation is fuzzy. In the free energy framework, such an uncertainty is modeled with a normal distribution that is spread over the most likely expectation of 3 cm. The higher the uncertainty, the higher the variance of the distribution. In **Figure 2.5.1.**, I plotted a normal distribution with a mean of 3 and a variance of 1, which is a model of the prior belief our vegan agent has about the size of a falafel. Bayes rule lets us use this prior and combine it with our noisy observation of v to form an updated belief about the size of the falafel. What we are calculating in this Bayesian inference amounts then to

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}, \tag{5}$$

where the posterior probability $P(H \mid E)$ gives the updated belief, or how likely it is that the falafel has a certain size depending on the observation and the prior belief $P(H)$. The likelihood $P(E \mid H)$ estimates how likely it is to make an observation given the prior belief.

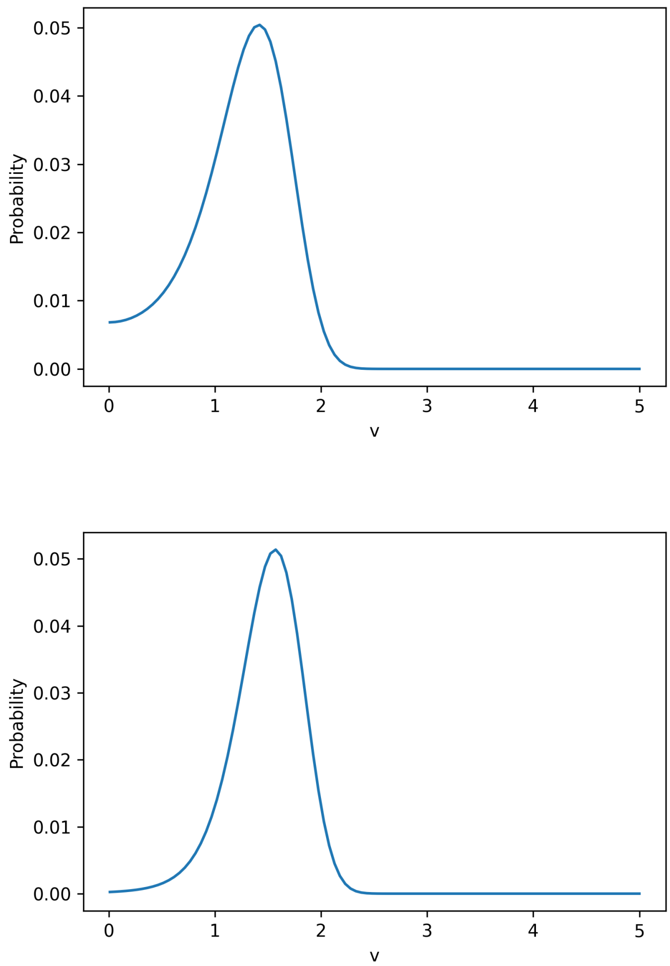

Figure 2.5.2. Updating of a prior belief with an observation. *If we use the prior belief from* ***Figure 2.5.1.*** *and make a noisy observation that is quantified with probability distribution A, we can apply Bayes rule to calculate the posterior belief B. According to A, the agent perceives light that corresponds more to a smaller falafel than anticipated somewhere most likely between 1 and 2 cm in diameter. Since the variance of the observation distribution is small, the posterior belief is closer to what has been observed than what has been expected by the prior.*

Both **Figure 2.5.2** and **Figure 2.5.3** show a belief update of the previous prior according to Bayes rule, but they vary in the precision of observation. The first posterior turns out to be way closer to the observation distribution than to the prior in **Figure 2.5.1.** There, I assumed an observation with little noise which results in a low variance of the distribution. We can be sure that the perception of the falafel is reliable and therefore the influence on the estimated size is big. The posterior distribution shows that our Bayesian agent now believes to see a small falafel between 1-2 cm in diameter. This updating of beliefs is an active process that understands perception as an act of combining predictions and sensations dynamically. There is no passive information processing as it is understood in the cognitivist paradigm, where there is first perception and then information processing which leads to choosing an action. The free energy agent is continuously generating possible perceptions and contrasts them with the states of its sensory apparatus.

As a comparison, in **Figure 2.5.3.** I repeated the Bayes calculation with a much noisier observation. This distribution comes with a higher variance which, in turn, leads to a less drastic updating of the prediction. If we do not trust our senses, we rely on the expectations we had before. The final posterior is closer to the prior predicted falafel size and its maximum ends up between 2 and 3 cm.

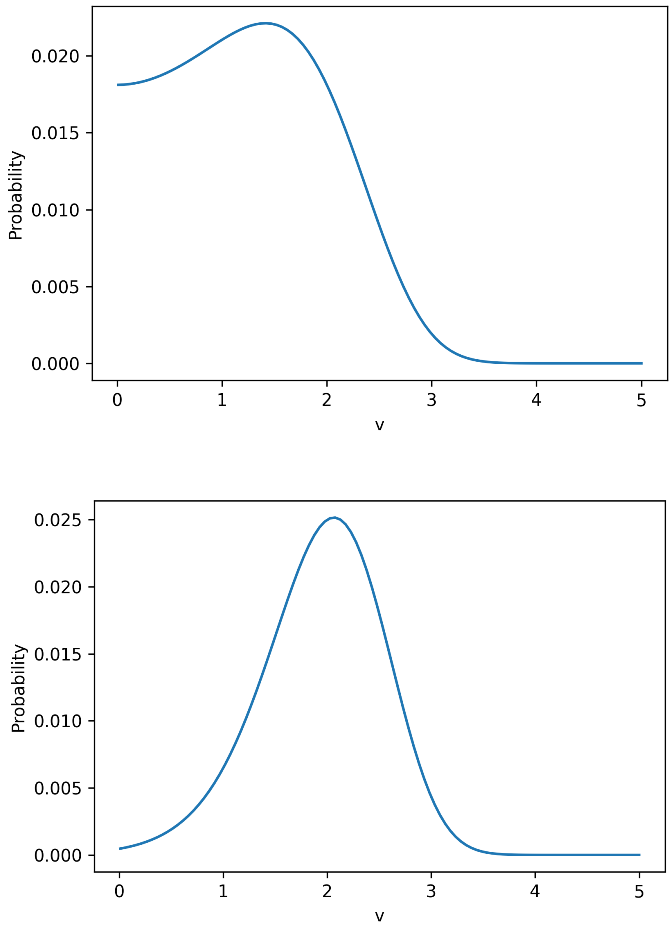

Figure 2.5.3. Updating of the same prior belief with a noisier observation.
Analogous to ***Figure 2.5.2,*** *we use the same prior belief to update the prediction of the falafel size under an observation that is way less clear. For example, if the falafel is further away we are less certain about the precision of our eyesight. This will lead to data with higher variance as shown in the upper panel. Updating again with Bayes rule gives now a posterior belief (lower panel) that is closer to the prior belief, with a mean between 2 and 3 cm. The more noise in our observation, the less effect has the observation on the posterior belief.*

So far, I have only demonstrated what the role of Bayes rule in a simple perception agent could be. But we have not touched the connection to free energy yet. The use of Bayesian inference is common in different approaches to explain the action of the mind computationally, such as predictive coding, predictive processing, and active inference. But free energy as a central measure of surprise in cognitive systems separates the framework from previous predictive approaches (Friston 2009, 2010).

As I exemplified with the simulation in this chapter, the Bayes rule is a powerful statistical tool to update predictions when all information about a perceptual problem is known. But usually, living organisms do not have a precise concept about how likely all possible states of the environment are. Therefore, free energy can be used to estimate the most efficient model of the world even though the evidence or P(E) is unknown.

For this, we can look for a value that maximizes the product of the likelihood P(EIH) and the prior P(H). Since the evidence does not depend on the cognitive agent's predictions about the world, we find a sufficient approximation of the optimal model without being dependent on knowing all the complexities of the state of the environment. Instead of dealing with the whole probability distribution P(HIE), free energy minimization finds the most likely value of the distribution which saves a lot of computation. When putting it into the context of evolutionary systems, it appears much more efficient to have a simpler measure that omits unnecessary details to save energy.

Now, the quantity that we want to minimize is usually called -F and the free energy framework aims to compute the maximum of F (Friston 2010, Bogacz 2017), which depends on the probabilities for the most likely value or mean ϕ, the received noisy stimulus u and the conditional probability $p(u \mid \phi)$:

$$F = ln(p(\phi)) + ln(p(u \mid \phi)) \tag{5}$$

A straightforward way to find the maximum is the gradient descent algorithm, which is often used in statistics and machine learning (Hinton and Zemel 1994). Here, we update ϕ according to the change in F to arrive at the most likely value for the size of our falafel:

$$\frac{dF}{d\phi} = v_p - \frac{\phi}{\Sigma_p} + \frac{g(\phi)}{\Sigma_u} g'(\phi) \tag{6}$$

In equation 6, $g(\phi)$ denotes the function which gives the relation between the observed stimulus and the source. The second term includes the precision of the prior prediction and updates the prediction closer to the mean of the distribution, the third term integrates the observation to better fit the sensation - again weighted by the precision of the observation.

How is this related to the prediction error? Without doing a full derivation, I would like to motivate the relation by extending the previous example of the perception of a falafel. When we update the prediction about the most likely size, we implicitly get the prediction error ϵ_p between expectation and observation from calculating the following quantity:

$$\epsilon_p = \phi - v_p / \Sigma_p \tag{7}$$

So, the prediction error of perception is just the difference between the estimated most likely value and the inferred size at this time step divided by the width of the distribution Σ_p. What this means is drawn in figure 2.5.4., where I plotted the change of inferred size in response to the prediction error ϵ_p. The updating of perception follows the strength of the prediction error. The lower the error, the smaller the change in belief that the perceiver goes through. This leads in my simple example to convergence at around 1.6 cm after 6 time steps. Comparing the two measures directly, we can see that the prediction error always changes earlier indicating the dependence of free energy minimization on it.

So why this lengthy discussion of the mathematical apparatus when my goal is to understand the philosophical roots of the framework? First, I will come back to most of the concepts such as negative free energy, prediction error, and precision in the rest of the thesis. Second, I believe it is important to clearly state the mathematical applications when we are interested in how to interpret the mathematical outcomes of the model. An interpretation of computational tools is always dependent on the context in which it is used (Humphreys 2002) and the goals of applying the same equations can be completely different (Herfeld & Lisciandra 2019). The philosophical question remains if Bayes rule is really what is computed in the brain (Colombo et al. 2021), or only acts as a useful approximation. A recent analysis of the use of the free energy framework revealed that many researchers have different modeling assumptions, ranging from instrumentalism to ontological commitments (Bruineberg et al. 2022). Now that I have clarified the use of the equations, we can finally dive into the claims that free energy research is making about its application to biological (Friston 2013), psychological (Limanowski & Blankenburg 2013), and artificial systems (Sajid 2021).

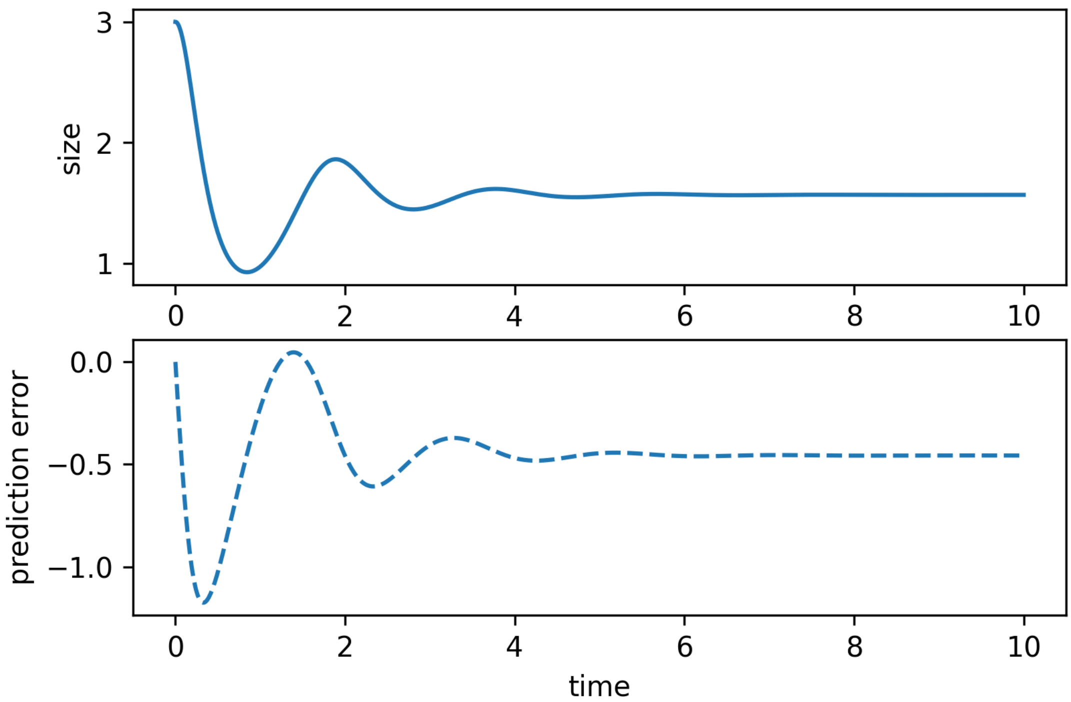

Figure 2.5.4. Prediction error calculation from free energy minimization. *The free energy formalism directly provides a measure of prediction error, which demonstrates the connection between the mathematical approaches to explain cognition. In the upper graph, I have simulated the estimated size of the square falafel updated according to free energy minimization after 10 time steps. The function converges around step 6 at a value of around 1.6 cm. The lower graph shows the change in the prediction error over time, which is part of the calculation and we see that it also converges when the free energy is stable. When the prediction error grows, the estimated size changes more dramatically. This simple example is intended to highlight the crucial role the difference between predicted and observed size plays when we update beliefs with free energy.*

3. **Philosophy of Information**

3.1 Free Energy as Information

Information has played a significant role in philosophy way before our current age of information technologies and data accumulation. While related to abstract concepts such as Eidos in Platonic and Aristotelian philosophy (Capurro 1978), which was often used to differentiate concepts from matter, information in a philosophical and scientific context, was widely discussed in the English Empiricist works of John Locke and David Hume (Locke 1823, Hume 1777/1902). There, we first see the real implications of considering information in a more scientific sense, beginning to understand its epistemic consequences. According to Locke, information constitutes the bridge between a thing and knowledge through our perception of it (Locke 1823). Hume (1777/1902) identified the relation between the number of possible choices and informational content, saying that more information reduces the degree of uncertainty. Picking from a single option is not informative, but deciding between 10 different options informs others about my preferences to a certain degree. Capurro identifies Ernst Mach's and Hermann von Helmholtz's work on statistical physics and entropy, which I will discuss in detail when discussing physical information, as influential for the work of W. Whewell, who argued for the objective character of information. In Whewell's perspective, the objectivity of information is grounded in its construction between direct sensations and a priori ideas (Capurro 1978). This is in contrast to the empiricists, who considered information to be transformed experiences. Traces of Whewell's clashing of ideas and sensations can be rediscovered in the bottom-up versus top-down discussion of neural processes in the cognitive sciences.

In this thesis, I focus on contemporary theories of information formed by cybernetics, computer science, and current debates in the philosophy of mind and the rather recent field of philosophy of information (Floridi 2002, Adriaans 2020). This chapter contrasts the influences of philosophy and physics on information theory to highlight their synergies and differences. With Luciano Floridi's scheme differentiating data and syntactic and semantic information, I tackle the free energy principle as it is used to construct a computational model of cognition.

3.2 What is Information?

Instead of looking for a single, well-defined answer, I discuss the different roles that information has played in science and philosophy. From the difference between matter and ideas to the use of entropy in statistical mechanics, and the quantification of communication channels, information plays diverse roles in philosophy, physics, and technology. It is one of those deceiving notions that appear to have a clear and intuitive meaning, but a closer analysis reveals the different concepts we employ when using the word information. The historical and etymological roots of the word have been analyzed by Capurro (1978), and a review of the philosophical development of the notion of information can be found in the works of Adriaans (2020). Here, I want to focus on the relations of information as defined in the mathematical field of information theory in the 20th century, the use of information in describing cognitive systems in cybernetics, artificial intelligence, and computational neuroscience, and the recent attempts to unify the philosophy of information.

The first conceptualization of information that students of mathematics, the natural sciences, or technology likely encounter is Shannon entropy (1948). Shortly after the end of the second world war, Claude Shannon published his foundational work in information theory in an article titled “A Mathematical Theory of Communication” (1948) while working at Bell Labs. Interestingly, the formalization of communication was still at the forefront of the mathematical endeavor that would later be the first milestone in information theory. Shannon started his paper and his book based on this work (1949) with clearly defined roles of sender, transmitter, and receiver. The Shannon entropy H, which can be interpreted as the inverse of information, is only meaningful between two parties communicating over a common channel. Shannon entropy without interpreters is not meaningfully defined. It appears that information is therefore validated between subjects and not as its own objective quantity in the world (**Figure 3.2.1**).

Weaver’s work with Shannon (1949) already splits the notion of information into three connected but different issues. They talk about

1. technical problems concerning the quantification of information and dealt with by Shannon’s theory (syntactic information)

2. semantic problems relating to meaning and truth

3. ’influential’ problems concerning the impact and effectiveness of information on human behavior

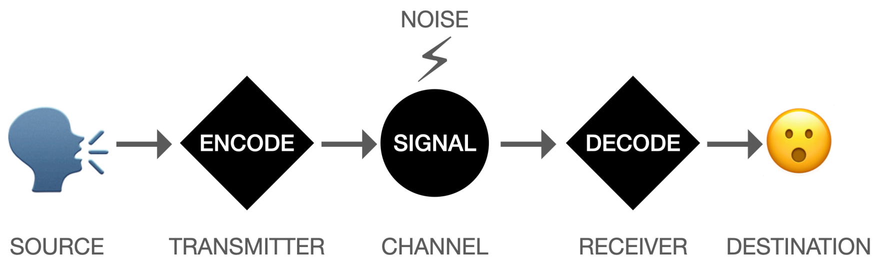

Figure 3.2.1. Shannon-Weaver Communication Model. *In the original formulation of information theory, Shannon and Weaver assumed a one-way communication between an information source and a destination (1948, 1949). Encoding a message by the transmitter and decoding by the receiver enables the transfer of a signal through an information channel. Noise can interfere with the signal even if encoding and decoding are uniquely specified. This reduces the capacity of the channel when the errors of the signal are corrected. Interestingly, this simplified model already establishes information as a relational quantity, that acquires its meaning through the communicating parties and not through objective features of the signal itself. The Shannon-Weaver model does not take into account context and feedback in the transmission of information, two properties that are essential when we try to explain cognitive information processes.*

This nested interpretation of information already shows the intricate relations of each one to the other. The splitting of syntactic, semantic, and cognitive problems is not as definite as it may seem. They act more as complementary dimensions of the same process. I return to this interpretation of understanding again when discussing models of the mind and enactivism, because the intersubjective co-construction of meaning is also found in recent theoretical developments in the cognitive sciences. In the next section, I discuss the mathematical approaches to quantifying information from physics and computer science and their limitations to resolve philosophical questions around information in detail.

3.3 Mathematical and Physical Concepts of Information

While Claude Shannon is usually regarded as the starting point for a formalized information theory, earlier research in physics and philosophy already included many of the points discussed in his mathematical theory of communication and its philosophical interpretation. In the second half of the 19th century statistical physics started to model thermodynamic quantities such as heat and pressure as average behavior of particles. James Clerk Maxwell published the first important statistical law of particle dynamics with the intention to describe the motion of an ensemble of particles with different energies instead of giving a deterministic description of what every single component of the system is doing. Further investigations into Maxwell's work by Ludwig Boltzmann lead to the formulation of the fundamental postulate of statistical mechanics, which puts the entropy S of a system - originally interpreted as the logarithm of the number of possible states Ω of the system - in the center of the theory:

$$S = k_B log \Omega \tag{8}$$

Here, the entropy S depends on the Boltzmann constant kB and the logarithm of the number of microstates that match the energy of the whole system. This was shown by Ludwig Boltzmann for an ensemble of particles in a fixed volume with constant internal energy U (1882). How is this equation related to information? When we consider the statement of Hume that every option chosen is increasing information, we can interpret negative entropy as proportional to information content.

$$-S = -k_B log \Omega = k_B log(1/\Omega) \propto I \tag{9}$$

But the formalizations in the 19th century did not make the many possible uses apparent. The field of cybernetics starting with Norbert Wiener in the 1940s set out to apply most of the mathematical ideas of information theory to develop a research discipline of control systems. Influential cyberneticist Will Ashby described the discipline as the study of systems that are open to energy but closed to information and control systems that are information tight (1956). There is a bit to unpack here. First, open systems are used to describe processes where the exchange with the environment is necessary to describe the dynamics. In contrast, a closed system is constituted just by internal structure and external function (Zwick, 2015). So having an open, information-tight model concerns a system that exchanges energy with its environment but not information. The communication theory of Shannon deals with the possibilities of information exchange between two such systems and analyses the capacity of communication channels with probabilities of the appearance of symbols and information transmission with entropy. This entropy is inspired by thermodynamics but quantifies an abstract, non-physical process.

Shannon on his own (1948) and later together with Weaver (1949) derived mathematical measures that manifested his influential “theory of communication”, which lead to what we now usually refer to as “information theory”, an abstract, mathematical framework that is decoupled from the physical substrates on which information is stored, processed, and transmitted (Landauer 1961). The important background to understanding the notion of cognitive information is the concept of entropy that Shannon used to formalize the amount of information a source can produce. This quantity H is given by the probabilities p of possible states with

$$H = -\sum p_i \cdot log p_i \tag{10}$$

Imagine a two-sided coin, the Shannon entropy of it amounts to:

$$H = -(p \cdot log(p) + (1-p) \cdot log(1-p)) \tag{11}$$

Figure 3.3.1. shows the number of bits that we are able to communicate when we use a coin with different weights. A regular, fair coin should have equal probabilities for landing on one of two sides, so according to Shannon we can transmit one bit of information with it. If the coin is heavily prepared to always land up on one side, we are not able to transmit any information at all since the coin toss will always tell us the same answer.

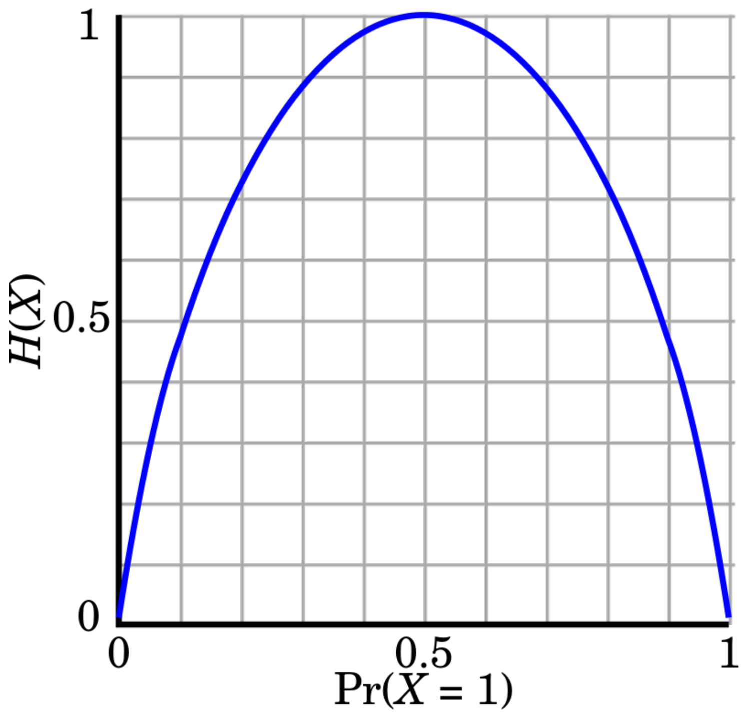

Figure 3.3.1 Shannon Entropy of a two-sided Coin. *The amount of information transmittable with a system depends on the number of possible states it can be in. The simplest example is a regular, fair coin that lands on either of two sides with equal probability. According to Shannon's theorem, such a two-state system can transmit exactly one bit, but if the probabilities are not 50:50 the amount of transmittable information decreases. For example, a coin that has heads on both sides will not transmit any information, since every coin toss leads to the same outcome.*

The strict formalization of information, entropy, and communication is why Shannon's work is one of the most-cited papers in computer science today. But what is often omitted is the full complexity of the Shannon-Weaver model as pictured in **Figure 3.2.1**. Between source and destination are many steps of encoding, transmitting, and decoding and the context of message passing is usually only implicit as well. To understand the use of information, we need to go beyond the analysis of structural properties alone and include the context and semantics of a message. For this, I will use Luciano Floridi's recent account of information to put the use of information in cognitive systems into perspective.

Interestingly, Shannon himself later commented on the use of his theorems in a more pragmatic way than most of his followers have interpreted him:

"The word 'information' has been given different meanings by various writers in the general field of information theory. At least a number of these will likely prove sufficiently useful in certain applications to deserve further study and permanent recognition. It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field."
Shannon 1993, p. 180

This quote shows that Shannon himself did not give information the unificatory potential that neuroscientists and especially free energy theorists sometimes ascribe to it. A more pragmatic approach can encompass the different uses of information as a scientific tool much better.

3.4 The General Definition of Information

Since there exist many different accounts of information, both from philosophical and mathematical perspectives, I present in detail the recent account by Luciano Floridi that takes both perspectives seriously and builds a consistent framework to construct bridges between philosophy and mathematics. Most comparable philosophies of information focus either on the background in logic or historical roots (Adriaans & van Benthem 2008). Floridi's account is applicable to current developments in the cognitive sciences because it takes the physical, philosophical, and ethical aspects of information into account from the very start. By paying equal attention to theory and applications, it is possible to approach the mind as an information processor without undue focus on foundational debates. Contemporary philosophy of the mind needs to be empirically informed to contribute to the debates in neuroscience, psychology, psychiatry, and artificial intelligence.

Many philosophical conceptualizations of information have been put forward by for example Tarski (1944), Kolmogorov (1965), and Wheeler (1990). Here, I will employ a recent classification that is inspired by empirical uses, and current developments in information theory and ethics. As it is often claimed, that we are living in the age of information, I want to use a contemporary classification that reflects how we use the term in relation to computation, data science, and neuroscience, especially as these are the areas that influenced the development of the free energy principle of cognition. My goal is to capture the theoretical environment in which these ideas have been developed, culminating in the claim of a unified brain theory.

Luciano Floridi is a practice-oriented philosopher who is most well-known for his contributions to the debates on big data (2003), technological innovation (2010), and ethics of artificial intelligence (1999). His account has significant overlap with the claims of researchers in the free energy framework without being directly informed by it. Building bridges between the mathematical and the philosophical accounts allows me to formulate a philosophical analysis that critically reflects on the quantitative aspects and helps with the interpretation of these concepts in the empirical realm. Floridi's work takes off from an influential paper on the "philosophy of information" (2002) that categorizes uses of the term in different fields, contributing to computational applications in philosophy, such as modeling of epistemic agents (De Langhe 2014) or data-driven analysis of philosophical developments (Noichl 2019).

The philosophy of information as a separate field deals with both the influence of information technology on philosophy itself and the philosophical grounding of information theory. These interactions are found in many pressing questions right now, and Floridi's 2011 overview has possibly even significantly expanded over the last decade:

"Concepts or processes like algorithm, automatic control, complexity, computation, distributed network, dynamic system, implementation, information, feedback, or symbolic representation; phenomena like HCI (human–computer interaction), CMC (computer-mediated communication), computer crimes, electronic communities, or digital art; disciplines like AI or Information Theory; questions concerning the nature of artificial agents, the definition of personal identity in a disembodied environment, and the nature of virtual realities; models like those provided by Turing Machines, artificial neural networks, and artificial life systems . . . these are just a few examples of a growing number of topics increasingly perceived as new, useful, of pressing interest, and academically respectable. Informational and computational concepts, methods, techniques, and theories had become powerful metaphors acting as "hermeneutic devices" through which to interpret the world. They had established a unified language that had become common currency in all academic subjects, including philosophy."
Floridi 2011 p.6

Obviously, such broad applications make a unique and general definition of information unlikely. From the beginning of information theory, it was clear that intuitive and mathematical definitions do not map onto each other neatly. Instead of providing a new all-encompassing definition of information, I will give a working definition that I will tailor specifically to my case study from cognitive science.

A proper working definition is provided by Luciano Floridi's analysis of the use of information in philosophy (2011). Floridi's information classification schema involves something he identifies as the General Definition of Information (**GDI**), which is spread over many philosophical disciplines making use of information in one way or the other (Chalmers 1995, Floridi 1999). The General Definition of Information consists of three conditions:

GDI: x is an instance of semantic information if and only if:

> **GDI.1** x consists of n data, for n >= 1
> **GDI.2** the data are well-formed (wfd)
> **GDI.3** the well-formed data are meaningful (mwfd)

The condition **GDI.1** deals with the definition of data, which can consist of measurements, observations, relations, pointers, or whatever can be quantified in this way. **GDI.2** on the other hand makes an assumption about syntax, the structure in which we can make sense of data. Without making a commitment about which structures are informative and which are not, philosophers that share this definition assume that there must be a structure that separates information from random data points. But **GDI.3** makes clear that there is a semantic component to this structure and if we follow the Shannon-Weaver model discussed in **figure 3.2.1** then this is the part, where the sender and receiver, the coding and decoding agent comes into play. Any message needs to be structured and meaningful to the agents. There is no information to transmit if they don't know how to understand the data. The meaning is a relation between the sender and the observer. This is another argument for a relational interpretation of information that it is context- and observer-dependent, a point that is not always clear in the literature about the free energy principle. In the next section, I give an overview of the use of information in this literature and argue that open questions require the understanding of cognitive information as a relational bridge between first- and third-person perspectives.

3.5 Information in the Free Energy Principle

Since the free energy principle was mostly developed and applied by neuroscientists, there is not much discussion available on what the information in the model represents. An important interpretation comes therefore from philosopher Andy Clark, who writes in his monograph on predictive processing how he views Friston's use of the term:

"It is also worth stressing that the term 'information' is here used simply as a description of 'energy transfer'… Talk of information, that is to say, must ultimately be cashed simply in terms of the energies impinging upon the sensory receptors. This is essential if we are to avoid, yet again, the illicit importation of an observer's perspective into our account of how informed observers are naturally possible in the first place."
Clark 2015, p.15

Clark's observation directly addresses the first-person problem of computational models of the mind. How can we talk about information in a cognitive system without talking about what is informative from our perspective, from the researchers' point of view? In line with Shannon's view of information transfer as communication between two agents, presented in chapter 2, we have to keep in mind that the semantic content of information arises between the observer and the observed. Observing an observer creates a second-order coupling, where what we find informative from the outside can easily be confused with what is informative to the observer. But the mapping between the two is more complicated than that. Biological systems capable of observing could therefore find sensations informative in the first person that from the third person are trivial or appear random. And the question is whether there is a transformation between what the outsider believes to be informative and what the system itself regards as information. The question of whether the free energy principle can provide such transformation will be discussed in chapter 4. Clark proceeds to more pragmatic observations, arguing for an information view that springs from the embodied mind.

"Information talk, thus used, makes no assumptions concerning what the information is about. This is essential, since sorting that out is precisely what the animal's brain needs to do if it is to serve as an empowering resource for the control of an environmentally apt response. It is thus the task of the (embodied, situated) brain to turn those energetic stimulations into action-guiding information."
Clark 2015, p.16

While being in line with cognitive science research in the embodied tradition (Varela et al. 1991, Thompson 2007), this does not provide an explanation of what free energy theorists mean when they use free energy as an informational measure. Ramstead and colleagues understand the research direction of free energy and embodied cognitive science as compatible with each other, but the main philosophers of the tradition point out the differences between the approaches (Thompson 2007, Di Paolo et al 2022). They argue that autopoiesis, or the self-production of the organism, is a different process than free energy minimization to maintain structural coherence. Self-distinction in the operationally closed systems of enactivism is different from the Markov blanket formalism that free energy researchers apply to separate internal from external states. Last and most importantly, the non-representational nature of enactive cognition is something that is according to the authors incompatible with the free energy framework in general.

A recent analysis of the ontological commitments of researchers sparked a heated debate not only about the correct interpretation of informational free energy but further revealed the different stances of authors working on the same topic. Bruineberg and colleagues analyze how many of the most influential papers take their claims to be literal statements concerning the modeled systems and how many adopt a more instrumentalist take on modeling itself. According to their analysis, the views vary widely revealing not only a diversity of interpretations about the meaning of free energy but also hinting at an unclear concept of information. Since free energy is directly dependent on entropy, which is a measure of information, the range of interpretations of one quantity implies that there is no single answer to what information means in the free energy principle. While Bruineberg and colleagues give some directions on how to move forward, the question of how to interpret free energy is part of a current debate in the field and I want to clarify that this interpretation is contingent on our understanding of the concept of information.

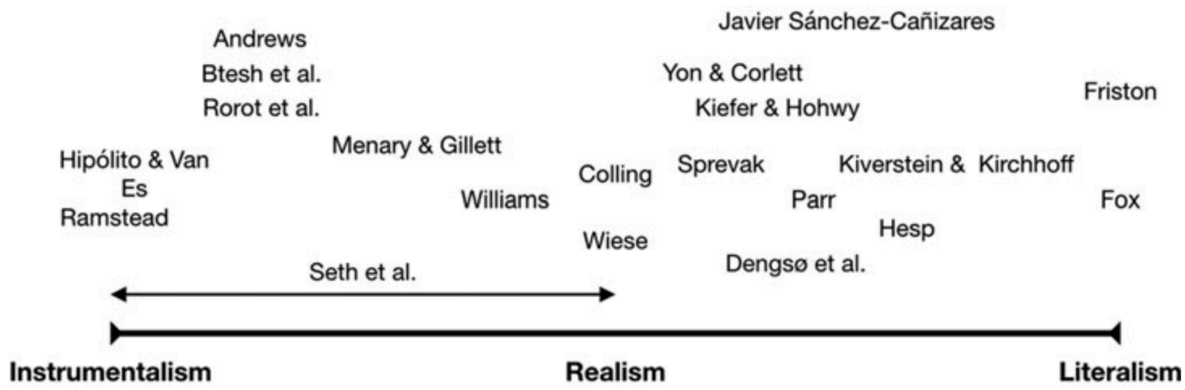

Figure 3.5.1 Ontological Commitments of Free Energy Interpretations.
Bruineberg and colleagues (2022) analyze the belief in free energy uttered by authors of publications in the field on a spectrum from instrumentalism, over realism, to literalism. On the far left, we find the pragmatic modelers and to the right, we find the interpretations that understand the free energy principle as a mechanistic process implemented in the organism itself.

These perspectives range from taking free energy and Markov blankets, the distinction between inside and outside of the system, as a formal tool of analysis, a physical feature of the system, or all there is in biological systems. Bruineberg and colleagues comment on a collection of open peer commentaries to their original paper and defend their classification as the starting point of a much needed discussion.

“The fact that a problem solved by a target system can be described using a Markovian formalism does not require a commitment to Markov blankets being explicitly represented (or instantiated) by the system itself. We can talk about realism only once theorists commit to the target system explicitly possessing or instantiating Markov blankets, either as representations or boundaries.”
Bruineberg et al. 2022 p. 72

Here it seems like two perspectives are colliding: First, the idea that the biological system models the environment in the way we describe it, and second, the view that it is empirically useful to model the biological systems ‘as if’ they model their environment. This controversy goes back to cyberneticist Ashby’s (1947) good regulator theorem, which states that every good regulator is a model of the system. Which can be taken literally or instrumentally as “every good regulator behaves as if it is a model of the system”. The fundamental tension arises in conflating scientific inference with inference through free energy minimization. Researchers use the free energy model to make inferences about cognitive systems, but the literalism account posits that free energy minimizing systems are models of the environment in the sense that we describe them. Some philosophers even go so far to call this perspective Kantian (Khachouf et al. 2013). If we accept the reality of our modelling, the Kantian transcendental is the generative process that we try to model and the modelling itself is our epistemic scope.

Relating the discussion of instrumentalism versus realism to the role of information appears straightforward. If we assume informational content to be physical only, we take a literalist stance on free energy as a measure of information. The neurobiological states of the system are only contingent on how physical information is processed. The transformation laws of information processing should in principle reveal behavior that is directly observable in brain imaging and related measures. But we are far from this mapping in the empirical tests of free energy. Is this because the fundamental formulation is not mathematically sophisticated enough to make concrete predictions? Or does informational free energy only provide a useful statistical tool like Markov blankets? In the first case, new quantitative predictions and measurements may make a difference. If we succeed to find unique signatures of information processing in the brain that agree with our models, these philosophical debates will surely move into the background. But if there is no clear mapping between free energy and brain function, which I assume to be highly likely given the difficulty of defining non-ambivalent tests, a clarification of what the framework does and does not tell us about cognition is necessary.

In the next section, I will develop the outlines for a conceptualization of information that aims to bridge the gap between the first-person and third-person perspectives necessary to develop the free energy principle into a computational, phenomenological model of cognition. This enactive information interpretation connects neuroscientific developments with recent philosophical accounts in cognitive science such as the embodied and enactive mind (Thompson 2007).

4. **Cognitive Information**

4.1 Cognition as Information Processing

Previous approaches in cognitive science have been highly successful in formulating the cognitive as some kind of information processing. The computer metaphor of the brain has been used to tackle perception, decision-making, and action as input-output relation with cognition as a mapping process (Pinker 1997). In this view, the senses act as measurement devices that passively relay information about the outside world to the computational unit, the brain. In this picture, deterministic neuronal processes convert information into concrete decisions and an accompanying set of actions, or policies (Di Paolo et al. 2022). This linear nature of perception and action is useful because of its simple causal structure, which fuels the hope to cast behavior in simple mathematical relations.

Cognitive neuroscientist Steven Pinker summarises the computational view of the mind this way:

"The brain's special status comes from a special thing the brain does, which makes us see, think, feel, choose, and act. That special thing is information processing or computation."
Steven Pinker 1997, p.24

While arguing that a computational perspective does not boil down to the computer metaphor itself, Pinker's argument omits how much the computer metaphor has enabled the reductionist take on cognition as information processing. A point that is often brought up by enactivists is to clarify how the computational paradigm in the mind sciences has to be extended to incorporate embodied and dynamic cognition (Varela et al. 1991, Thompson 2007, Di Paolo et al. 2022).

In contrast to contemporary paradigms in cognitive science, the computationalist paradigm can be understood as a third-person approach. Which is analogous to the perspective in computer science that computation can be implemented and verified from the outside. The computationalist paradigm is a scientific model that uses data external to the system, such as visual stimuli and observed behavior, to identify its internal workings. It does not capture the perspective of what it is like to be the system (Nagel 1974). The computational paradigm can therefore be understood as the next step in the behaviorist tradition. In it, we don't treat the brain as a black box anymore, but the black box becomes a Turing machine that maps inputs to outputs. As it turns out, this linear processing is only applicable to very simple cognitive processes and does not even scratch the surface of the complexity that cognitive systems are capable of (Di Paolo 2005).

4.2 First and Third-Person Approaches

In this chapter, I will draw on two separate but complementary ideas by philosopher Thomas Nagel. First, the now classic reading in philosophy of mind "What is it like to be a bat?" (1974), in which Nagel provocatively asked if all our outside understanding of the experience of a bat is sufficient to describe its subjective dimensions. From understanding the biological and neurological mechanisms of the bat, can we understand how it is like to be one? Of course, this is a non-trivial problem, since we try to use knowledge from one reference frame to explain what is happening in a different frame, in this case, the first-person perspective of being a bat. To approach the intersection of both perspectives, we need to let go of the strict dichotomy between objective and subjective descriptions of systems.

Here, the second important line of thought from Nagel comes into play. In his book "The View from Nowhere” (1989) he analyses the philosophical problems we encounter when assuming an objective, fact-based perspective of the mind. The way out of this epistemological dilemma is to see subjective and objective perceptions as idealizations on different ends of the same spectrum. Subjective knowledge reduces to intrinsic information on one’s own sensations and thoughts. Objective knowledge usually refers to observer-independent facts that are completely detached from the act of observation. But where do these facts come from if they are shared between different observers? In the end, both of these views run into paradoxes when applied to the mind-body problem. Neither an idealistic nor a materialist perspective can resolve these paradoxes. So, understanding knowledge as a set of relations between observers opens up a kind of phenomenological middle way in the sense of Husserl (1976). I call this the second-person or relational perspective.

Neurobiologist Francisco Varela describes neuroscientific methodology as exclusively interested in studying the brain and behavior from the outside to derive objective facts about cognition (1996). This ignores all the subjective dimensions that Nagel discussed in his “what it is like to be a bat” paper. Describing the structure of the brain and the dynamics of neuronal activation will not give us insight into how the subjective experience of the animal. Consequently, we could either ignore it as Daniel Dennett does (1988) and regard it as an uninteresting epiphenomenon, or pick up again the methods of introspectionist psychology (Danziger 1980). The latter has fallen out of favor due to problems of replication and not being in line with the goal of quantifying psychology. But Varela’s program of neurophenomenology aims to bridge the gap between qualitative first-person reports and third-person neuroscientific data. Maybe there are measures of subjective experience that can be tested empirically.

To further strengthen the connection between first- and third-person accounts, I propose to have another look at what information means to each observer and to a collective of people. Chalmers' (1995) dual view of information gives a good point of departure for my analysis of cognitive information.

"… information (or at least some information) has two basic aspects, a physical aspect, and a phenomenal aspect. This has the status of a basic principle that might underlie and explain the emergence of experience from the physical. Experience arises by its status as one aspect of information when the other aspect is found embodied in physical processing."
Chalmers 1995, p.23

In opposition to Dennett (1988), Chalmers does not think that understanding all the physical aspects will give rise to a fundamental understanding of phenomenal qualities of experience. Since he sees a dualism of physics and phenomenology, this perspective has been classified as dual-property monism or proto-phenomenology. But interestingly, Chalmers makes a direct connection to the aspects of information itself, instead of only staying within the philosophy of mind. According to him, cognitive information reveals not only the world around us but also how it feels to perceive. Thomas Nagel already discussed this in relation to the subjective experience of a bat (1974), but the phenomenological aspect of perception is best exemplified by the gedankenexperiment of Mary the color scientist by Frank Jackson (1986). In the experiment, hypothetical scientist Mary is a color-blind expert in optics and knows everything about the physical properties of color; which wavelength corresponds to which excitation of the retina, how the visual cortex decodes the input, and so on. But what happens if she suddenly can perceive a red apple? Is her physiological knowledge sufficient to not learn anything new or does she have a novel experience that explains something that goes beyond what she learned from the third-person perspective of studying color? The physical and the phenomenological quality of how it feels to perceive seem to be distinct aspects of information. But in his article, Chalmers leaves open where and why this split happens or what the relation of the physical and phenomenological aspects are.

He goes on to justify this separation of information as having two distinct properties by adopting a computational perspective and the mind as an information processor:

"… analysis of the cognitive explanation of our judgments and claims about conscious experience—judgments that are functionally explainable but nevertheless deeply tied to experience itself—suggests that explanation centrally involves the information states embedded in cognitive processing. It follows that a theory based on information allows a deep coherence between the explanation of experience and the explanation of our judgments and claims about it."
Chalmers 1995, p.23

This quote makes clear that there is no biological priority for subjective experience. Every processor of Chalmers information (in contrast to Shannon information from before) has a specific experience of being. For example, the phenomenon of color perception is not sufficiently explained by giving the wavelength of light and the physiological processing on the retina, but the phenomenological aspects of how it feels to perceive is embedded in the cognitive processing from the system's first-person perspective. How to include the phenomenological aspects into neuroscience is what Chalmers called the hard problem of consciousness and this way of thinking inspired one of the most prominent computational theories of consciousness: integrated information theory (Tononi 1998), where phenomenal properties arise in a quantifiable way by analysis of information passing in the network.

Shannon made clear that his type of information is to be found in communication systems, telephones, morse code, and letters, but Chalmers directly applies his dual-aspect view of information to cognitive systems and not only to their interactions. Do these notions of information have different domains? I argue that neither an externalist nor a strictly internalist sense of information will be useful to understand cognitive information. We need a relational interpretation of information as relating external and internal states from an agent's point-of-view. But to make the full argument, I first discuss where we find cognitive information and what free energy as an informational measure tell us about this relation.

4.3 Where to Find Free Energy

The practical applications of the free energy framework are continuously growing since the mathematical formulation was published in 2005. Proposed applications include explanations of perceptional illusions such as binocular rivalry (Hohwy et al. 2008), a better understanding of mental diseases such as schizophrenia (Friston et al. 2016), and self-organization of biological systems (Friston 2013, Ramstead et al. 2018). But the question remains if the proposed mechanisms are explicitly found in these systems or if the Free Energy Principle presents a useful heuristic of more complex underlying dynamics. Moreover, if we were able to group all these different phenomena under the free energy umbrella, would it demonstrate its unificatory potential or only highlight the generality of the description? The free energy principle has to yield useful predictions to be a candidate to explain the origin of cognition and its unfolding in biological, neuronal and even social systems (Veissiere et al 2019).

Free energy in cognitive science does not seem to be a directly measurable physical quantity as its counterpart in thermodynamics, it acts more as an informational measure in the modeling practice of self-organization (Andrews 2021). As described above, information in the classic sense of Shannon needs encoding and decoding, a transmitter, and a receiver in between the source and destination. The semantic content of a message relies therefore on the conventions set up between sender and receiver. If we don't set up a shared mapping between morse code and the regular alphabet, the message of the telegraph would be meaningless. But besides the syntactical relation, we also need a semantic component. A sending agent transmits a first-person perspective, and a receiving agent reconstructs meaning out of the message. While we can describe the mathematical properties of transmission and the encoding and decoding process, Shannon information is not sufficient to describe how the experience changes over communication. How does one subjective experience map to the perceived experience by the receiver? A mathematical approach to capture this experience necessitates some phenomenological qualities that information theory in general does not cover. For inspiration, I turn to neurophenomenology, the mutual study of first and third-person perspectives of cognitive agents.

An important philosophical and methodological tool is the separation of inside and outside, perception and action, and of the subjective and the objective. All these distinctions are made based on the difference between first and third-person qualities. In the free energy principle, the difference maker is the Markov blanket, a statistical border or separation of dependent and independent variables. Everything inside the blanket is considered to be intrinsic information of the organism. Everything on the outside are the environmental states that the organism probes. This is a mathematically elegant formalism that was introduced by mathematician Judea Pearl (1989). It neatly separates inside from outside and we can think of biological structures like the membrane of a cell as real analogies of Markov blankets. If we want to find free energy in the brain, we have to find the Markov blankets. While empirical evidence is still ambiguous concerning the physical structures that could fulfil this role, philosophical problems have begun to accumulate around the definition itself.

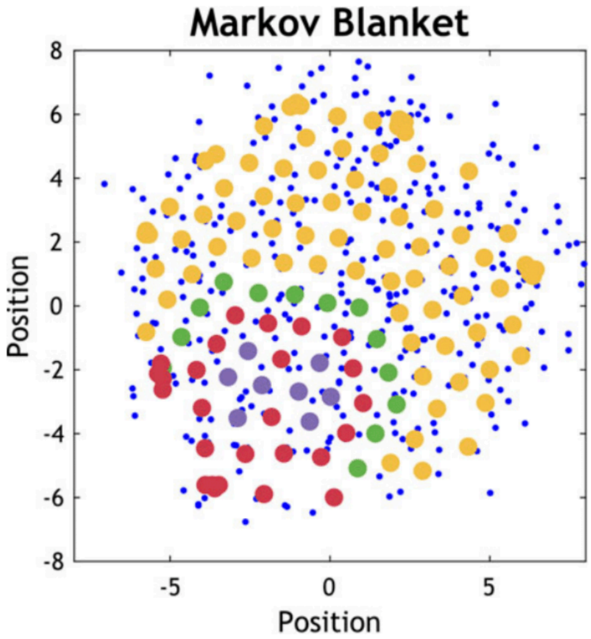

Figure 4.3.1. A Markov Blanket in a Spatial Model. *This simulation of simple cells shows the building up of a boundary, the Markov blanket, between internal and external states of a perceiving and acting collective. The violet dots represent the operationally closed internal states, and red marks the acting cells that together with the green sensory cells make up the Markov Blanket. External states, which are marked in yellow, build up around the boundary. The blue dots govern the interactions between cells and represent the electrochemical states of the model. From Bruineberg et al. 2022.*

In a recent article, Bruineberg and colleagues analyze the definitions given for Markov blankets and critically examine the specific uses of the mathematical tool in the context of the free energy principle (2022). After a much-needed discussion with commentators on the original paper, their meta-commentary concludes that the

Markov blanket is not a physical construct but a modeling tool, that changes shape in different applications (Bruineberg et al. 2022). While the mathematical background is clear, it is not clear how to apply the tool to brain images. The distinction between inside and outside is done in the model by the scientist and not in the brain itself.

4.4 Active Inference and Experience

The question remains: How can we capture subjective experience in a mathematical model? The answer is not straightforward since a reduction of first-person perspective to simple physical principles seems completely off the table. In line with Chalmers (1995) and Thompson (2007) I use subjective experience to describe the hard problem of consciousness, also referred to as qualia. In my argumentation, I follow their main principles:

1. Conscious experience is the first-person perspective of an organism.
2. Conscious experience is embodied and enacted between agent, body, and environment.
3. We can approach these perspectives in a mathematical framework using dynamic systems theory and the free energy framework.

By building on information theory, phenomenology, and the free energy framework we can come up with a methodology that acts as a candidate for computational phenomenology. This fits under the umbrella of the recently founded field of Mathematical Consciousness Science (Kleiner) and complements contemporary philosophy of mind discussions by constructing computational models that make the assumptions and approaches of different theories explicit and comparable.

Such an approach enables the empirical testing of competing theories and a progressive development of the models in line with state-of-the-art neuroscientific data. This would amount to another big step into testable models helping to go beyond mere speculation about the character of subjective experience, facing a common critique of the research into first-person characteristics from introspectionist psychologists, for example, the research of Wundt (Danziger 1980).

What does the concrete application of free energy or the whole active inference framework to subjective experience look like? The extension to consciousness itself is rather recent (Sandved-Smith et al. 2021, Ramstead et al. 2022), but has already been applied to phenomenological concepts such as Husserl's description of time perception (Albarracin et al. 2022). In these attempts, Husserlian phenomenology acts as the point of departure for a rigorous description of subjective experience. By combining Friston's work with a formalized framework of phenomenology, the authors explain the specific experiences of time in the free energy framework. In their mapping of perception to uncertainty, time arises as a parameter that separates recollection from prediction and helps us to construct causal models in our brain. Therefore, time is used to sort perceptions into sequences according to the most efficient model.

4.5 Free Energy and Enactive Information

The main differences between enactivism and the free energy principle of cognition lie in their philosophical assumptions about what mental action looks like. Enactivism regards action and perception both as parts of dynamic interactions with the environment, which cannot be reduced to internal mental models. The enactivist paradigm posits that first and third-person perspectives are not separable and consciousness cannot be explained by data about the brain alone. Its aim is to bridge the gap between neuroscientific and phenomenological methods by formalizing a common framework (Varela 1996). On the other hand, a literalist interpretation of the free energy principle separates these perspectives neatly via the Markov blanket. If we take free energy as exclusively dependent on physical information, we should be able to understand consciousness via a physical description alone. This reduces the phenomenological perspective to just another frontier of physics, even if a new theory might be necessary. Enactivism urges that such a reduction is impossible and phenomenal data is as valid as neuroscientific data in the cognitive sciences. Di Paolo and colleagues (2022) go as far as stating that the ontological assumptions are completely incompatible. They propose a forking path between the two approaches because of two main reasons:

Forking 1: Historicity

In the derivation of the free energy principle, researchers make use of the assumption that all organisms stay intact by reaching a non-equilibrium steady state. They self-organise around a stable set of values also known from dynamical systems theory as random global attractors. This process assumes that at a certain point the rate of change over time is not important for the dynamics anymore. In contrast, many systems biologists and also enactivism think the historicity of an organism is crucial to understand its dynamics. The time dependence of possibility spaces of a biological system and symmetry-breaking effects, like long-term memory, are impossible to explain in a non-equilibrium steady state. (Colombo and Wright 2018, Di Paolo et al. 2022)

Forking 2: Co-constitution

The mutual construction of the internal and the external is a main tenant of enactive cognition, because it is built on a dynamic coupling between the organism and its environment. The organism is as much part of the environment as the environment is part of the organism, they co-constitute each other. This is a quality that free energy in a strict representationalist reading, where internal models represent external information, does not consider. The friction appears because the direct representation of the external by the internal already gives a directionality, a hierarchy as the inside as a subset of the outside. This hierarchical picture is in tension with co-constitution in enactive approaches, but only if the boundary dynamics are understood as a representational process.

By taking an enactivist reading of information, I argue that the position of Di Paolo and colleagues is built on a misreading of the free energy framework. Understanding free energy as a mechanistic framework (as for example Badcock et al. 2019 do), within which mathematical tools would have to be interpreted literally as being part of the biological workings leads to the conclusion of forking paths. But if we adopt the relational view of information that I have outlined in chapter 3, we can build a bridge that not only benefits the free energy framework but also forces enactivism to formalize its claims and thereby makes them testable in experiment. How this formalization could look like is still open, but recent work in category theory points in the direction of a useful mathematical formalism (Kauffmann 2017).

As put forth by Di Paolo and colleagues (2022), most of the philosophical friction between enactivism and free energy arises from the concept of representation used. There are two main representing factors here. First, does a cognitive system internally represent external features of the world? And second, how do our models represent features of the system studied? A clear separation of these two aspects already disentangles a significant chunk of the debate between instrumentalism and realism. While philosophy of mind is most interested if cognitive systems have internal representations of the external world, philosophy of science extends this question by asking what the variables in our cognitive models represent in the actual systems. So we cannot be instrumentalists about our models if we assume that agents have a model of their environment in their head. On the other hand, it is consistent to be a realist about mental representation and the ontological status of our models or completely instrumentalist about both notions of representation. In the cognitive sciences, one central debate concerns whether cognitive systems have to represent something external or not. This debate is now known as the "representation wars" (Sims & Pezzulo 2021). A starting point to connect representation in modeling and cognitive representation is a systematization of the different roles representation plays in the free energy principle.

1. Organization: representing external by internal states leads to a separation between the inside and outside. The Markov blanket acts as a boundary that dictates the dichotomy of organization. If internal states represent outside variables, we need to assume an organizational hierarchy that would not be there without it. In my cat example, the external state of its position in the room is separated from my internal distribution of possible locations given by my knowledge about the dark room. The view of Hohwy (2013) uses this internalist representation argument to ground the realism of free energy in biological systems. The enactivist view on the other hand interprets this organization as a mere side-effect of coupling between states (Bruineberg et al. 2022).

2. Structure: since internal models are used to navigate the world, there is the argument that models have to share some structure with the external processes they are describing (Sims and Pezzulo 2021). But as with all structuralist realist positions, the proof of an isomorphism between representation and phenomena is impossible without some given transformation. Instead, the empirical adequacy of the model could be sufficient. For energy conservation, it also makes more sense that organisms create simple models that are enough to survive most external disturbances, instead of going for complex dynamics of the environment, which are very energy costly to maintain and act on.
3. Content-related: if we take a more biologically-minded action-oriented perspective, we can evade the pitfalls that come with structural representation. An adaptive organism may look for the best action to survive in a dynamic environment, instead of looking for the most accurate representation of external states. With this perspective, we change the content of representation in the free energy framework. Instead of representing features of the environment, we could read cognitive representations as adaptive actions in synchrony with the environment. This embodied view of representation (representation as coupling and not as a map from internal to external states) is a perspective that for example, Clark (2016) defends passionately. Here, an often-used example is ball-catching. Instead of internally representing the trajectory of a ball throw as accurately as possible, embodied representations enable the cognitive agent to catch the ball by just running at a similar speed to stabilize it in its field of view. What is represented here is not a complex external state, but a relation between the catcher and the ball. The free energy principle allows for the minimization of uncertainty in this simple relation, which is much more efficient than modeling the whole trajectory of the ball and the catcher. Running at the same speed as the ball and keeping the angle constant is an action that guides perception, and this predictive perception is exactly what

allows us to solve a complex problem with a simple model that does not track external objects in all their complexity.

4. Function: The functional role of representation is already discussed by Piaget (1954). In the free energy framework, we encounter the open question if the function of representation is to describe external states as accurately as possible, or if representations are just marks of the coupling between agent and environment. The latter prioritises guiding action and regulating sensory information over accurately mapping the world. But according to more radical views, all these cognitive functions do not necessitate representations to function properly (Hutto and Myin 2017).

This overview of the role that representation plays or could play in future formulations is just the tip of the iceberg of the many different philosophical interpretations that researchers have put forth in connection to the free energy principle of cognition. Here, I want to demonstrate how debated the concept of mental representation is and how far we can go without it. A relational perspective on information, which before has been understood as a representational quantity already covers the adaptive and action-oriented aspects of cognition without tapping into the paradoxes of direct representation (Weisberg 2007). These relational aspects are also in line with the enactivist paradigm of cognition and their arguments for a science of the mind that takes phenomenology and lived experience seriously. If we understand our free energy model of a modeling agent as an instrumentalist and contextual practice, we circumvent the core of the arguments about the incompatibility raised by Di Paola and colleagues (2022). A relational interpretation of the information used in the free energy principle also must include an updated understanding of what representation in scientific modeling entails, something that is so far missing in the free energy literature.

5. Conclusion

This thesis does not present a conclusion to the research question, it acts more as a process statement for future investigations into the relations of the free energy principle and enactive cognition. I highlighted the different modeling assumptions of the two approaches, with the central question concerning the meaning of cognitive representation. Current debate points into the direction that free energy is used in a representational and non-representational way by different users. So it could either be something we have to identify in the brain or a modeling tool that we can employ in different scientific cases, from biological organisms to cognitive and cultural systems. Here, I emphasize that one should judge the success of the free energy principle as an explanation in different epistemological contexts, and not only whether it provides a unified brain theory or not. A pluralist reading of free energy could be useful in the philosophy of modeling without necessitating strong ontological statements.

By understanding the history of ideas that lead to the free energy framework, it becomes obvious that free energy and Bayesian cognitive science already have many applications from statistical physics to perception and artificial intelligence. So even if the big claims of a unification of all these levels by the free energy framework do not prove to be true, a better understanding of how and why to use free energy as a computational tool elucidates its modeling assumptions and epistemic affordances. Specifically, I analyzed the role that information plays in the framework because it is often introduced as a simple mathematical description going back to Shannon. Yet, when we look deeper into the different texts using the concept of information, it turns out that information comes with a syntactic, semantic, and relational component and this tripartite structure is often overlooked in discussions about the ontological commitments of free energy research.

Especially the semantic, and relational component is necessary to tackle such a methodologically challenging topic like subjective experience, where usual third-person methods are not sufficient. A model of phenomenological properties needs to include dimensions of how it feels to be the specific system modeled, and what perception from the first-person perspective is like. A mere syntactic description does not suffice here. Embodied and enactive cognition research starts from these dimensions, and therefore already maps the territory to which, I argue, the free energy framework should be extended. Beyond recent arguments about the incompatibility of the two, I am motivating to take the tripartite structure of information seriously. Only by considering syntactic, semantic, and cognitive aspects, we regain the relational nature of perception and action in free energy models. Something that has been criticized as missing from the current framework (Di Paolo et al. 2022).

But including all three dimensions of information is only possible if we give up the mechanistic, realistic interpretation of the free energy principle of cognition. Free energy is not a direct observable in the brain, it is rather a computational tool that allows us to build concrete models which we can test in experiments, but this does not verify or falsify free energy itself. It only lets us judge the specific model we derived from the general principle. Consequently, the concept of information we use in modeling cognition carries a significant chunk of the ontological commitments we make in our models.

5.1 Outlook

Is the map the same as the territory it represents? Of course, a map gets more informative with added information and increased scale but at a 1:1 mapping the map becomes useless. This map-territory discussion is a central part of philosophy of science and especially debates regarding scientific representation since Alfred Korzybski raised the point nearly a century ago (1931). The recent debate about the philosophical grounding of the free energy principle of cognition has focused on the infamous “map-territory fallacy” in philosophy of science, where a map is mistaken for the environment to be mapped. From simple beginnings, where researchers assumed that free energy is a physical quantity in the brain and Markov blankets are material structures to be found in biological systems, the focus has shifted to the modeling strategies used in explaining cognitive systems and organisms by the same framework. Do we need to hold on to the idea that the assumed biological processes create an internal model of the environment?

Critics such as Mel Andrews (2021) have put forth points of why this is a reification of modeling assumptions - creating a map of cognitive functions does not imply that these structures are really there. But quite recently, the Fristonians made this critique their own by putting forward the idea in a preprint that it is a useful modeling strategy to model biological systems “as if” they modeled the environment (Ramstead et al. 2022). After these many rounds of recursion, the arguments become loaded with self-referentiality and the jury is still out on what the authors exactly intend to say with their “map-territory fallacy fallacy”. Their proposed notion of nested representations is definitely a concept that could use some further unpacking from a philosophy of modeling perspective. While their philosophical position is not fully developed yet, a recent preprint explains nested representations in the free energy principle as a representation of the representational capacities of cognitive agents. In their own words: "…it is ultimately a model that is apt to model the modeller." (Ramstead et al. p.17.). While the authors argue that the map-territory fallacy does not apply to their computational model, they leave open how this perspective does not collapse back to a purely instrumentalist use of the free energy principle.

An enactive interpretation of information as a relational quantity between actions might be necessary to come up with a proper first-person description necessary for computational phenomenology. Since enactivism is currently the main driver behind including phenomenological perspectives into neuroscience and the free energy principle is a currently influential computational model of cognition, it is inevitable to construct bridges and see if the ideas behind both approaches could be made compatible or are fundamentally incommensurable.

References

Apps, M. A. J. & Tsakiris, M. 2014. “The free-energy self: A predictive coding account of self-recognition.” *Neuroscience & Biobehavioral Reviews* 41: 85–97.

Ashby, W.R. 1956. *An Introduction to Cybernetics*. London: Chapman & Hall.

Ashby, W.R. 1947. “Principles of the Self-Organizing Dynamic System.” *The Journal of General Psychology* 37 (2): 125–28.

Allen, M. & Friston, K. J. 2018. “From Cognitivism to Autopoiesis: Towards a Computational Framework for the Embodied Mind.” *Synthese* 195 (6): 2459–82.

Andrews, Mel. 2021. “The Math Is Not the Territory: Navigating the Free Energy Principle.” *Biology & Philosophy* 36 (3): 30.

Albarracin, M., Riddhi, J. P., Maxwell, J. D. R., & Yoshimi, J. 2022. “Mapping Husserlian Phenomenology onto Active Inference.” arXiv. https://doi.org/10.48550/arXiv.2208.09058.

Adriaans, P., & van Benthem, J. 2008. *Handbook of Philosophy of Information*. Elsevier.

Adriaans, P. 2020. “Information.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/information/.

Bruineberg, J., Dolega, K., Dewhurst, J., & Baltieri, M. 2022. “The Emperor’s New Markov Blankets.” *Behavioral and Brain Sciences*, October, 1–63.

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A. & Hohwy, J. 2019. “The Hierarchically Mechanistic Mind: An Evolutionary Systems Theory of the Human Brain, Cognition, and Behavior.” *Cognitive, Affective, & Behavioral Neuroscience* 19 (6): 1319–51.

Bogacz, R. 2017. “A Tutorial on the Free-Energy Framework for Modelling Perception and Learning.” *Journal of Mathematical Psychology* 76: 198–211.

Buckley, C. L., Kim, C. S., McGregor, S. & A. K. Seth. 2017. “The Free Energy Principle for Action and Perception: A Mathematical Review.” *ArXiv:1705.09156 [q-Bio]*, May. http://arxiv.org/abs/1705.09156.

Capurro, R., 1978. *Information. Ein Beitrag zur etymologischen und ideengeschichtlichen Begründung des Informationsbegriffs* (Information: A contribution to the foundation of the concept of information based on its etymology and in the history of ideas), Munich, Germany: Saur.

Chalmers, D., 1995. "Facing Up to the Problem of Consciousness" Journal of Consciousness Studies, 2(3):200-19, 1995.

Clark, A. 2013. "Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." Behavioral and Brain Sciences 36 (3): 181–204. https://doi.org/10.1017/S0140525X12000477.

Clark, A. 2015. "Radical Predictive Processing." *The Southern Journal of Philosophy* 53 (S1): 3–27. https://doi.org/10.1111/sjp.12120.

Clark, A. 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford, New York: Oxford University Press.

Colombo, M., & Wright, C. 2018. "First Principles in the Life Sciences: The Free-Energy Principle, Organicism, and Mechanism." *Synthese*, https://doi.org/10.1007/s11229-018-01932-w.

Colombo, M., Elkin, L., & Hartmann, S. 2021. "Being Realist about Bayes, and the Predictive Processing Theory of Mind." *The British Journal for the Philosophy of Science* 72 (1): 185–220. https://doi.org/10.1093/bjps/axy059.

Carhart-Harris, R. L., and K. J. Friston. 2010. "The Default-Mode, Ego-Functions and Free-Energy: A Neurobiological Account of Freudian Ideas." *Brain* 133 (4): 1265–83. https://doi.org/10.1093/brain/awq010.

Carhart-Harris, R. L., and K. J. Friston. 2019. "REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics." *Pharmacological Reviews* 71 (3): 316–44. https://doi.org/10.1124/pr.118.017160.

Chipot, C., Pohorille, A. eds. 2007. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*. Vol. 86. Springer Series in CHEMICAL PHYSICS. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-38448-9.

Danziger, K. 1980. "The History of Introspection Reconsidered". *Journal of the History of the Behavioral Sciences*. 16 (3): 241–262.

Demekas, D., Parr, T. & Friston, K. J. 2020. "An Investigation of the Free Energy Principle for Emotion Recognition." *Frontiers in Computational Neuroscience* 14. https://doi.org/10.3389/fncom.2020.00030.

De Langhe, Rogier. 2014. “A Unified Model of the Division of Cognitive Labor.” *Philosophy of Science* 81 (3): 444–59.

Di Paolo, E., Thompson, E. & R. Beer. 2022. “Laying down a Forking Path: Tensions between Enaction and the Free Energy Principle.” *Philosophy and the Mind Sciences* 3. https://doi.org/10.33735/phimisci.2022.9187.

Di Paolo, E. 2005. “Autopoiesis, Adaptivity, Teleology, Agency.” *Phenomenology and the Cognitive Sciences* 4 (4): 429–52. https://doi.org/10.1007/s11097-005-9002-y.

Foerster, 2003 p.289. “Ethics and Second-Order Cybernetics.” In *Understanding Understanding*, by Heinz von Foerster, 287–304. New York, NY: Springer New York.

Friston, K. J., & Stephan, K. E. 2007. “Free-Energy and the Brain.” *Synthese* 159 (3): 417–58. https://doi.org/10.1007/s11229-007-9237-y.

Friston, K. J. 2009. “The Free-Energy Principle: A Rough Guide to the Brain?” *Trends in Cognitive Sciences* 13 (7): 293–301. https://doi.org/10.1016/j.tics.2009.04.005.

Friston, K. J. 2010. “The Free-Energy Principle: A Unified Brain Theory?” *Nature Reviews Neuroscience* 11 (2): 127–38. https://doi.org/10.1038/nrn2787.

Friston, K. J., Tamara Shiner, Thomas FitzGerald, Joseph M. Galea, Rick Adams, Harriet Brown, Raymond J. Dolan, Rosalyn Moran, Klaas Enno Stephan, and Sven Bestmann. 2012. “Dopamine, Affordance and Active Inference.” Edited by Olaf Sporns. *PLoS Computational Biology* 8 (1): e1002327. https://doi.org/10.1371/journal.pcbi.1002327.

Friston, K. J. 2013. “Life as We Know It.” *Journal of The Royal Society Interface* 10 (86): 20130475. https://doi.org/10.1098/rsif.2013.0475.

Friston, K. J. Klaas Enno Stephan, Read Montague, and Raymond J Dolan. 2014. “Computational Psychiatry: The Brain as a Phantastic Organ.” *The Lancet Psychiatry* 1 (2): 148–58. https://doi.org/10.1016/S2215-0366(14)70275-5.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T. & Pezzulo, G. 2015. “Active Inference and Epistemic Value.” *Cognitive Neuroscience* 6 (4): 187–214. https://doi.org/10.1080/17588928.2015.1020053.

Friston, K. J. 2019. “A Free Energy Principle for a Particular Physics.” ArXiv:1906.10184 [q-Bio], June. http://arxiv.org/abs/1906.10184.

Floridi, L. 1999, “Information Ethics: On the Philosophical Foundation of Computer Ethics”, *Ethics and Information Technology*, 1(1): 33–52. doi:10.1023/A:1010018611096

Floridi, L. 2002, "What Is the Philosophy of Information?" *Metaphilosophy*, 33(1–2): 123–145. doi:10.1111/1467-9973.00221

Floridi, L. (ed.), 2003, The Blackwell Guide to the Philosophy of Computing and Information, Oxford: Blackwell. doi:10.1002/9780470757017

Floridi, L. 2010, "The Philosophy of Information as a Conceptual Framework", *Knowledge, Technology & Policy*, 23(1–2): 253–281. doi:10.1007/s12130-010-9112-x

Floridi, L. 2011, The Philosophy of Information, Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199232383.001.0001

FitzGerald, T. H. B., Dolan, R. J. & Friston, K. J. 2015. "Dopamine, Reward Learning, and Active Inference." *Frontiers in Computational Neuroscience* 9. https://doi.org/10.3389/fncom.2015.00136.

Gershman, S. J. 2019. "What Does the Free Energy Principle Tell Us about the Brain?" ArXiv:1901.07945 [q-Bio]. http://arxiv.org/abs/1901.07945.

Herfeld, C. & Lisciandra, C. 2019. "Knowledge Transfer and Its Contexts." Studies in *History and Philosophy of Science Part A* 77 (October): 1–10. https://doi.org/10.1016/j.shpsa.2019.06.002.

Helmholtz, H. 1882. Die Thermodynamik chemischer Vorgänge. Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin, 1882 - Erster Halbband (Berlin: Verl. d. Kgl. Akad. d. Wiss., 1882)

Hohwy, J., Roepstorff, A. & Friston, K. 2008. "Predictive Coding Explains Binocular Rivalry: An Epistemological Review." Cognition 108 (3): 687–701. https://doi.org/10.1016/j.cognition.2008.05.010.

Hume, D. 1748. An Enquiry concerning Human Understanding. Reprinted in Enquiries Concerning the Human Understanding and Concerning the Principles of Morals, 1777 which was reprinted, L.A. Selby-Bigge (ed.), Oxford: Clarendon Press, 1888 (second edition 1902).

Kersten, D., Mamassian, P. & A. Yuille. 2004. "Object Perception as Bayesian Inference." *Annual Review of Psychology* 55 (1): 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005.

Humphreys, P. 2002. "Computational Models." *Philosophy of Science* 69 (3): 1–11. https://doi.org/10.1086/341763.

Hutto, D. & Myin, E. 2017. Evolving Enactivism: Basic Minds Meet Content. MIT Press.

Noichl, M. 2019. “Modeling the Structure of Recent Philosophy.” *Synthese* 198 (6): 5089–5100. https://doi.org/10.1007/s11229-019-02390-8.

Jackson, F. 1986. "What Mary Didn't Know". *The Journal of Philosophy*. 83 (5): 291–295. doi:10.2307/2026143. JSTOR 2026143.

Kersten, D., Mamassian, P. & Yuille, A. 2004. “Object Perception as Bayesian Inference.” Annual Review of Psychology 55 (1): 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005.

Husserl, E., & W. Biemel. 1976. Die Krisis der Europäischen Wissenschaften und die Transzendentale Phänomenologie: Ein Einleitung in die Phänomenologische Philosophie. Husserliana: Edmund Husserl – Gesammelte Werke. Springer Netherlands. https://www.springer.com/de/book/9789024702213.

Kauffman, L. H. 2017. “Mathematical Work of Francisco Varela.” *Constructivist Foundations* 13 (1): 11–17.

Khachouf, O., Poletti, S. & Pagnoni, G. 2013. “The Embodied Transcendental: A Kantian Perspective on Neurophenomenology.” *Frontiers in Human Neuroscience* 7. https://www.frontiersin.org/article/10.3389/fnhum.2013.00611.

Korzybski, A. 1933. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. International Non-Aristotelian Library Publishing Company.

Kolmogorov, A.N., 1965, “Three Approaches to the Quantitative Definition of Information”, Problems of Information Transmission, 1(1): 1–7. Reprinted 1968 in International Journal of Computer Mathematics, 2(1–4): 157–168.

Landauer, R. 1961. “Dissipation and Hear Generation in the Computing Process. IBM Journal of Research and Development,” no. 5: 183–91.

Limanowski, J., & Blankenburg, F. 2013. “Minimal Self-Models and the Free Energy Principle.” Frontiers in Human Neuroscience 7. https://doi.org/10.3389/fnhum.2013.00547.

Locke, J. 1823. *The Works of John Locke.* London: Oxford University Press.

McCulloch, W. S. & Pitts, W. H. 1943, “A Logical Calculus of the Ideas Immanent in Nervous Activity” *Bulletin of Mathematical Biophysics* 7, 115–133. Reprinted in McCulloch 1964, pp. 16–39.

Nagel, T. 1974. "What Is It Like to Be a Bat?". *The Philosophical Review*. 83 (4): 435–450. doi:10.2307/2183914.

Nagel, Thomas. 1989. *The View from Nowhere*. Oxford, New York: Oxford University Press.

Pearl, J., D. Geiger, and T. Verma. 1989. "Conditional Indipendence and Its Representations." *Kybernetika* 25 (7): 33–44.

Ramstead, M. J. D., Seth ,A. K., Hesp,C., Sandved-Smith, L., Mago, J., Lifshitz, M., Pagnoni, G. 2022. "From Generative Models to Generative Passages: A Computational Approach to (Neuro) Phenomenology." *Review of Philosophy and Psychology*, https://doi.org/10.1007/s13164-021-00604-y.

Ramstead, M. J. D., Kirchhoff, M. D. & Friston, K. J. 2020. "A Tale of Two Densities: Active Inference Is Enactive Inference." *Adaptive Behavior* 28 (4): 225–39.

Ramstead, M. J. D., Badcock, P. B. & Friston, K. J. 2018. "Answering Schrödinger's Question: A Free-Energy Formulation." *Physics of Life Reviews* 24 (March): 1–16. https://doi.org/10.1016/j.plrev.2017.09.001.

Ramstead, M. J. D., Kirchhoff, M. D. & Friston, K. J. 2020. "A Tale of Two Densities: Active Inference Is Enactive Inference." *Adaptive Behavior* 28 (4): 225–39. https://doi.org/10.1177/1059712319862774.

Rao, R. P. N., & Ballard, D. H. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2, 79–87.

Rosenblatt, F. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review*. 65 6, 386-408 .

Rumelhart, D.E., and G.E. Hinton. 1986. "Learning Representations by Back-Propagating Errors." *Nature*, 323: 533–36. https://doi.org/10.7551/mitpress/4943.003.0042.

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.

Sandved-Smith, L., Hesp, C., Mattout, J., Friston, K., Lutz, A. & M. J. D. Ramstead. 2021. "Towards a Computational Phenomenology of Mental Action: Modelling Meta-Awareness and Attentional Control with Deep Parametric Active Inference." *Neuroscience of Consciousness* 2021 (1): 18. https://doi.org/10.1093/nc/niab018.

Shannon, C E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal*, 27: 379–423.

Shannon, C. E., & Weaver, W. 1949. “The Mathematical Theory of Communication.” Urbana, IL: *The University of Illinois Press*, 1-117.

Shannon, C. E. 1993. “Collected Papers” N. J. A. Sloane and A. D. Wyner (eds), Los Alamos, CA: *IEEE Press*, 177–179 and 180–183. doi:10.1109/TIT.1953.1188572

Sims, M., & Pezzulo, G. 2021. “Modelling Ourselves: What the Free Energy Principle Reveals about Our Implicit Notions of Representation.” *Synthese,* 199 (3-4):7801-7833 https://doi.org/10.1007/s11229-021-03140-5.
Srinivasan, M. V., Laughlin S. B. & Dubs A. “Predictive Coding: A Fresh View of Inhibition in the Retina” *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 216, 427-459

Tarski, A., 1944, “The Semantic Conception of Truth: And the Foundations of Semantics”, Philosophy and Phenomenological Research, 4(3): 341–376. doi:10.2307/2102968

Thompson, E. 2007. *Mind in Life*. Harvard University Press.

Varela, F. J., Rosch, E. & Thompson, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA, USA: MIT Press.

Varela, F.J. 1996. “Neurophenomenology: A Methodological Remedy for the Hard Problem.” *Journal of Consciousness Studies* 3 (4): 330–49.

Veissiere, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J. & L. J. Kirmayer. 2019. “Thinking through Other Minds: A Variational Approach to Cognition and Culture.” The Behavioral and Brain Sciences 43: e90. https://doi.org/10.1017/S0140525X19001213.

Wade, N. J. 1996. “Descriptions of Visual Phenomena from Aristotle to Wheatstone.” Perception 25 (10): 1137–75. https://doi.org/10.1068/p251137.

Weisberg, M. 2007. “Who Is a Modeler?” *The British Journal for the Philosophy of Science* 58 (2): 207–33. https://doi.org/10.1093/bjps/axm011.

Wheeler, J. A. 1990, “Information, Physics, Quantum: The Search for Links”, in Complexity, Entropy and the Physics of Information, Wojciech H. Zurek (ed.), Boulder, CO: Westview Press, 309–336.

Whitehead, A. N. & Russell, B. 1963. Principia Mathematica. Cambridge: Cambridge University Press.

Appendix

Code 1 Bayesian Belief Updating

```
# Example 1
# value of 1 variable is inferred from 1 observation
# organism infers size of food v from light intensity g = v**2

import numpy as np
from matplotlib import pyplot as plt
from google.colab import files

u = 2
var_u = 1

v = np.linspace(0.01, 5, 100)
v_p = 3
var_p = 1

prior = 1/np.sqrt(2*np.pi*var_p)*np.exp(-(v-v_p)**2/(2*var_p))
likelihood = 1/np.sqrt(2*np.pi*var_u)*np.exp(-(u-v**2)**2/(2*var_u))

posterior = prior*likelihood

plt.figure(0)
plt.plot(v,prior)
plt.xlabel('v')
plt.ylabel('Probability')
plt.savefig('prior1.png',dpi=300)
```

```
plt.figure(1)
plt.plot(v,likelihood/likelihood.sum())
plt.xlabel('v')
plt.ylabel('Probability')
plt.savefig('likelihood1.png',dpi=300)

plt.figure(2)
plt.plot(v,posterior)
plt.xlabel('v')
plt.ylabel('Probability')
plt.savefig('posterior1.png',dpi=300)

# with bigger noise (var_u=10) in the observation

u = 2
var_u = 10

v = np.linspace(0.01, 5, 100)
v_p = 3
var_p = 1

prior = 1/np.sqrt(2*np.pi*var_p)*np.exp(-(v-v_p)**2/(2*var_p))
likelihood = 1/np.sqrt(2*np.pi*var_u)*np.exp(-(u-v**2)**2/(2*var_u))

posterior = prior*likelihood

plt.figure(3)
plt.plot(v,prior)
plt.xlabel('v')
plt.ylabel('Probability')
plt.savefig('prior1.png',dpi=300)
```

```
plt.figure(4)
plt.plot(v,likelihood/likelihood.sum())
plt.xlabel('v')
plt.ylabel('Probability')
plt.savefig('likelihood2.png',dpi=300)

plt.figure(5)
plt.plot(v,posterior)
plt.xlabel('v')
plt.ylabel('Probability')
plt.savefig('posterior2.png',dpi=300)

#files.download('prior1.png')
#files.download('likelihood1.png')
#files.download('likelihood2.png')
#files.download('posterior1.png')
#files.download(‘posterior2.png')
```

Code 2 Perception as Free Energy Minimisation

```
# Example 2
# Free energy minimisation for most likely value of phi
# instead of getting whole distribution
# gradient descent with eulers method

T = 10
dt = 0.001
t = np.linspace(0,T,10000)
v = np.zeros_like(t)

u = 2
var_u = 1
v_p = 3
var_p = 1

for i_t, n in enumerate(t):
        if n == 0:
            v[i_t] = v_p
        else:
            v[i_t] = v[i_t-1] + dt * ( (v[i_t-1]-v_p) / var_p + 2*(v[i_t-1]) * (u-(v[i_t-1])**2) /
var_u )
          #v[i_time] = v[i_time-1] + dt * ( (v[i_time-1] -v_p) / var_p + dg(v[i_time-1]) *
(u_obs-g(v[i_time-1] )) / var_u )

# Example 3
# Calculating prediction error from
# free energy minimisation

v = np.zeros_like(t)
```

```
eps_u = np.zeros_like(t)
eps_p = np.zeros_like(t)

for i_t, n in enumerate(t):
    if n == 0 :
        v[i_t], eps_u[i_t], eps_p[i_t] = v_p, 0., 0.
    else:
        v[i_t] = v[i_t-1] + dt * ( - eps_p[i_t-1] + 2*(v[i_t-1]) * eps_u[i_t-1] )

        eps_p[i_t] = eps_p[i_t-1] + dt * ( (v[i_t] -v_p) - var_p * eps_p[i_t-1] )
        eps_u[i_t] = eps_u[i_t-1] + dt * ( (u-(v[i_t])**2) - var_u * eps_u[i_t-1] )

plt.figure()
plt.subplot(211)
plt.plot(t,v)
plt.xlabel('time')
plt.ylabel('size')

plt.subplot(212)
plt.plot(t,eps_u, linestyle='--')
plt.xlabel('time')
plt.ylabel('prediction error')
plt.savefig('PredictionError.png',dpi=300)

#files.download('PredictionError.png')
```