UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE E TECNOLOGIE

DIPARTIMENTO DI INFORMATICA
GIOVANNI DEGLI ANTONI

Corso di Laurea Magistrale in Informatica

A COMPUTATIONAL APPROACH TO EMOTIONS: MODELING VALENCE AND AROUSAL IN ACTIVE INFERENCE

Relatore: Giuseppe Boccignone
Correlatore: Sabrina Patania

Tesi di Laurea di:
Luca Annese
Matricola 18805A

ANNO ACCADEMICO 2023-2024

Abstract

This thesis advances the field of affective computing by introducing a novel computational framework that integrates the Active Inference approach with the constructs of valence and arousal to model emotional dynamics. While widely studied as a framework for perception, action, and learning, its application to emotions remains under explored, particularly regarding the simultaneous integration of valence and arousal. The work presented here addresses this gap by extending the Active Inference framework to accommodate both affective dimensions.

This thesis demonstrates that valence and arousal can be incorporated into the Active Inference framework to capture the dynamic influence of emotions on decision-making. The integration of these affective states provides a biologically plausible model of emotional processes and underscores the role of emotions in guiding behavior in uncertain environments. The findings pave the way for more comprehensive models of emotion-driven agency and open new avenues for research in affective computing and autonomous systems.

Keywords: active inference, arousal, valence, emotions, affective computing, free energy principle

Acknowledgments

I would like to begin by thanking my supervisor Professor Boccignone who introduced me and guided me through the topics that I now so deeply value.

I would also like to extend my sincere gratitude to my co-supervisor Sabrina, for her enthusiasm and expertise, that guided me in the development of this thesis and gave me so many opportunities.

After having paid the proper tributes to those who actually deserve it, I would also love to thank the many people who, whether they were aware of it or not, have given me so much in these years.

To my family, who accompanied and supported me until reaching this great goal.

To Massi, Fede and Ale, my second family, whose constant presence keeps me grounded and put up with me and my stupid ideas.

To those I have started this journey with, Ariana, Lia, and Pietro, with whom I have endured the hardships and experienced the beauty of (university) life.

And to those I am ending it with: Gabriele, whose intelligence is matched only by his stupidity, the best study partner I could ever ask for, and Giulia, whom I unfortunately met too late, but with whom I spent the most significant time, who made these last months the ones I will look back with the most nostalgia.

To my Quinto Romano boys: Sangio, who kept me company during all these university days and the following evenings out, Luca, who understands me like a brother, Matteo, who is always ready for a chat and share a drink, and Francesco, with whom I've grown up and will keep growing.

To the PhuseLab people who have accompanied me during this last year and brighten my days in the laboratory. In particular to Jacopo, who lights up each day with his wonderful craziness.

To you all, and many others that I have met in these years, we will meet again at the vending machines.

Explicit expliceat,
bibere scriptor eat

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AcI	Active Inference
AC	Affective Charge
BET	Basic Emotion Theory
BMR	Bayesian Model Reduction
CAT	Conceptual Act Theory
DL	Deep Learning
EFE	Expected Free Energy
ER	Emotion Recognition
FEP	Free-Energy Principle
InG	Information Gain
KL	Kullback-Leibler
ML	Machine Learning
POMDP	Partially Observable Markov Decision Process
RL	Reinforcement Learning
VFE	Variational Free Energy

Introduction

The aim of this thesis is to push the boundaries of current affective computing approaches by incorporating a more comprehensive computational framework that would bridge agency and emotions.

More specifically, the work presented here grounds in the Active Inference framework [1] as the computational substance of agency and enhances such framework with the construct of valence and arousal, on which emotions are defined.

In individuals, any sensory modality induces an aroused and valenced state. Valence tend to signals prospects of satisfaction or anxiety, while arousal is more prone into driving behavior. Both are dimensions of the core affect [2], which refers to the fundamental, conscious experience of feeling good or bad (valence-related), energized or fatigued (arousal-related), which serves as a basis for more complex emotional experiences, and cognition in general.

Active Inference is a novel theoretical approach that integrates action, perception, and learning, positing that sentient behavior arises from the brain's use of internal models to predict and guide actions [3]. This concept is rooted in the Free Energy Principle which provides a systematic and functional strategy for building an agent that incorporates requirements that meet our purposes.

Active Inference is indeed a complex theory that encompasses many fields of application. Primarily, it is a process theory of brain function [4], whose predicted electrophysiological responses resemble empirical measurements and extends earlier theories such as predictive coding. Its principles are being exploited: in robotics, demonstrating the potential for creating autonomous systems that mimic human-like behavior; in psychology, aiding in understanding both healthy and pathological behaviors; even in philosophy and consciousness research [5, 6].

Even though Active Inference and all its derivatives have been employed and tested as possible models of emotions, many fail to integrate such vision into the framework as it is.

Cogent for this Thesis, to the best of our knowledge, no Active Inference agent has been designed and developed that includes both valence and arousal in its generative model.

Previous work related to valence can be found primarily in the reinforcement

learning literature, specifically linking reward and punishment to positive and negative valence. More recently, models that embed a mood parameter have been proposed. However, these models had a targeted scope and did not aim at accounting for the broader, general role of emotions.

The model we propose here makes a step forward in such direction. It combines the hierarchical Bayesian model and valence evaluation of [7] with the free energy model of arousal potential formalized by Yanagisawa [8–10].

Model simulations are carried out in a paradigmatic environment of the reinforcement learning literature: a synthetic agent (rat) explores a T-shaped maze in search of a reward (avoiding punishment). Through unexpected changes of the reward positions, we gauge the agent's responses with respect to the affective part, in particular the combination of valence and arousal.

In the sequel of this Thesis, we start by unfolding the state of the art and the achievements in affective computing; we discuss how most approaches rely upon the outdated Basic Emotion Theory paradigm. We then move on to overview the different theories of emotions, in particular the constructivist view of emotion, on which this thesis is conceptually based. This view motivates the hypothesis that the brain can be seen as a predictive machine.

We then formally introduce the Active Inference framework. The model is then extended in order to account for both the core affective dimensions of valence and arousal. Eventually, we present results from our simulations and show how the affective component can influence the behavior of the agent.

Some mathematical/technical details are left to the Appendix. There are also presented the bare essentials of the Python framework pyMDP conceived to support and solve Active Inference and which has been extended to include the features introduced in the present dissertation.

Chapter 1

State of the art

Affective computing is the interdisciplinary field of study concerned with developing systems and devices that can recognize, understand, interpret, simulate, and stimulate affective states [11]. Since its introduction by Rosalind Picard [12] , Affective Computing stands as an intermediary of computer science, psychology, neuroscience, philosophy, and industry. The growing interest of both technology giants such as Meta, Apple, Amazon and Google as well as hundreds of small companies shows how much of an impact these methods have [13, 14]. Its applicability ranges from healthcare [15], as it can assist in early diagnosis of mental disorders, in robotics to build more human-like androids and even in marketing to predict or influence consumer behavior [16].

1.1 Defining the landscape

Affective computing was firstly defined by Picard as

> computing that relates to, arises from, or influences emotions.

Such definition is quite broad and general, but from it we can distinguish the three main topics Affecting computing encompasses and we desire artificial agents to have

1. recognise emotions
2. express emotions,
3. have emotions,

the latter point being the hardest one, is still a matter of debate for the philosophers of mind. The above definition stands when all its parts are well defined. The main problem here, is that an actual definition of what an emotion is that everyone

agrees on does not exist; in fact, it is still being discussed and, it is a rather controversial issue. For the time being, we assume a broad definition of emotion, that will suffice:

> *A coordination or composite of experiences, behavioral expressions and physiological/neurological components, with varying duration in time.*

Not only is the fundamental quantity of this field hard to define to tune in all, but even the field itself is often mislabeled [17], making Affective Computing a slippery terrain.

1.2 Emotion detection & recognition

The emotion detection/recognition task presently represents the most explored region of the Affective Computing landscape.

Emotions are an inseparable component of human life. These emotions influence human decision-making and help us communicate to the world in a better way. Emotion detection, also known as emotion recognition, is the process of identifying a person's various feelings or emotions (for example, joy, sadness, or fury) [18].

From a practical view, emotion detection is a cornerstone of affect-aware interfaces that aim to automatically detect and intelligently respond to users' affective states in order to increase usability and effectiveness [19]; instead of simply responding to the user's commends, human-centered interfaces must have the ability to detect subtleties of and changes in the user's behavior, especially his/her affective behavior, and to initiate interactions based on this information [20]. On the other hand, from a theoretical standpoint, emotion detection arise from a digital signal processing and machine learning problem because it involves the development of a classifier or regressor to detect an ill-defined phenomenon (affect) from observable signals.

The modalities on which it is defined make the emotion recognition task complex: the problem is extremely challenging because the affective states that we intend to capture and measure are not directly observable and are imbued with social and individual differences [21], that change even according to the context, be it cultural or political. In addition, affective states are represented as multicomponential, encompassing neurobiological changes, physiological responses, bodily expressions, action tendencies, and cognitive, metacognitive, and phenomenological states.

Figure 1.1 illustrates the conceptual foundation of ER. Note that there is no box or label titled emotion. Instead, the figure lists the componenets that most reasearches would agree as being relevant for emotion without getting into wheter

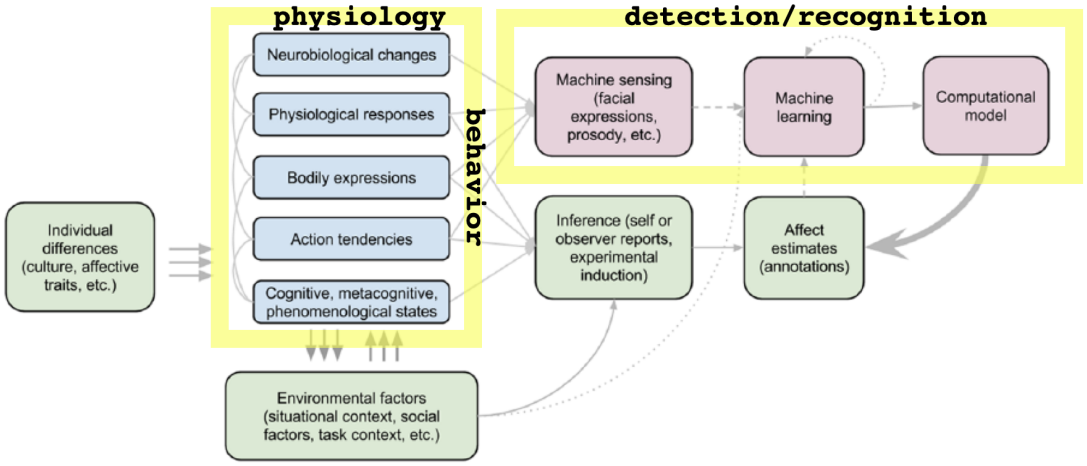

Figure 1.1: Conceptual and computational foundation of the ER approach. Modified from [22]

the subjective experience is the emotion, or the whole complex. We refer to this abstract representation as the affect estimate [22]. Specifically, emotions are not directly conveyed by some physiological or behavioral signal, but rather are psychological phenomena that are not directly accessible to the outside observer and that constitute the information that perceivers are eager to infer from the perception of visible behavior and actions [23]. Emotion is not the only factor that underlying nonverbal communication. Interpersonal attitudes, personality, cultural and even situational contexts are also to be considered among the cues. Thus, affect measurement requires inference, or construction, to produce an affect estimate, such as the amount of anger for self-reports, judged anger for observer reports, or elicited anger when experimentally induced.

Sensors can measure bodily/physiological signals (e.g., infrared, EEG, electrodermal activity), extending beyond what can be easily perceived by humans (e.g., facial expressions). However, they cannot infer an affect estimate from the measurements themselves. Affective computing assumes there is a link between an affect estimate and machine-readable bodily/physiological signals. Trope [24] proposed two steps for the social attribution processes: identification and inference. Identification is the formation of a first representation of the perceived stimulus in terms of meaningful relevant categories. This representation then serves as input for the inference process. AC models the before mentioned link likewise: firstly, abstractions (called descriptors or features) are extracted from raw signals recorded by sensors. For example, if the signal is a facial video recorded from a camera, the descriptors might be the activation of facial action units (AUs) or facial textures.

Computer vision-based techniques are needed to automatically compute facial descriptors from video. In general, computing descriptors from signals, entails also correction for measurement errors, interpolation techniques for missing data, and methods for signal filtering and denoising.

From the features extracted, affect estimates are to be produced. The most common approach uses techniques from a subfield of machine learning called supervised learning. Supervised learning uses training data consisting of descriptors that are temporally synchronized with researcher provided affect annotations (from self-reports, observer judgments, or elicited condition) to model (learn) the relationship between the two. Ideally, the models should also include contextual information, which can be divided into: environmental factors and internal context. Note that the term model entails different meanings: computationally it refers to an output of a supervised classifier and can take many forms, according to the supervised learning method applied. The model can take another meaning if we focus on the definition of emotion: there are two types of generic emotion models in affective computing, namely discrete emotion model [25] and dimensional emotion model (or continuous emotion model) [26].

These approaches rely on cleverly handcrafted features and machine learning models to derive such feature representations. Unfortunately, the steps involved in designing handcrafted features can be work-intensive and error-prone. Motivated by the success of deep learning in other AI-relevant tasks based on images and speech, researchers have started to use deep models for representation learning in affect recognition [27]. Nowadays, DL-based models have become hot spots and outperformed ML-based models in most areas of affective computing. A comprehensive review can be found at [28].

1.3 What is an emotion

In 1884, William James, the American psychologist, famously posed the question: what is an emotion? [29] After more than a century of scientific inquiry, however, emotions remain essentially contested concepts: scientists disagree on how they should be defined, on where to draw the boundaries for what counts as an emotion and what does not, on whether conscious experiences are central or epiphenomenal, and so on. Such disputes have sown great discord among scientists, leaving the field in perpetual upheaval, and without a unified framework for guiding scientific inquiry and accumulating knowledge [30].

The first to scientifically explore emotions was Charles Darwin with his publication of *The Expression of the Emotions in Man and Animals* in 1872. Darwin reasoned that if humans and other mammalians species share a common ancestor, then humans behavior can give evidence of such connection. He noticed that some

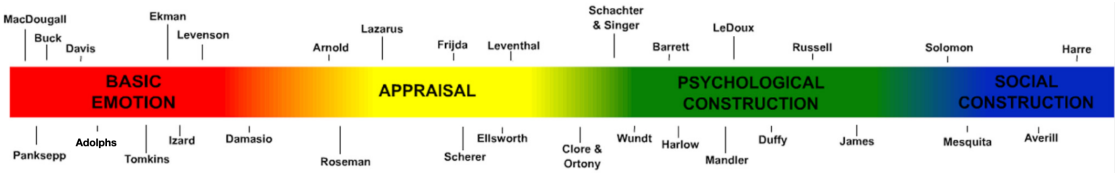

Figure 1.2: Navigating emotion Theories. Theories are loosely arranged along a spectrum, which could be defined in terms of a gradient of essentialism, the highest degree located at the left. Four "zones" are distinguished: (1) Basic Emotion Theories (BET); (2) Appraisal Theories; (3) Psychological Construction Theories; (4) Social Construction Theories. Adapted from [31, 32]

bodily and facial expressions of humans were similar to those of other animals, and concluded that behavioral correlates of emotional experience were the result of evolutionary processes. An important aspect of Darwin's theory, is that emotion expressions are not unveiling an internal mental state, but they are rather vestiges of our evolutionary past [33]. Although Darwin did not intended to craft a model of emotion, Silvan S. Tomkins [34] before, and Paul Ekman after [35, 36], used it as inspiration to define what is known as the Basic Emotion Theory. The former posited the existence of a limited number of discrete, primary, or "basic emotions" as part of a universal human nature. These were held to be characterized by signature facial expressions and specific patterns of behavioral and autonomic responses. The latter focused on facial movements claiming that very specific configurations of facial muscle movements correspond to different emotion categories in a one-to-one manner.

In contrast to preceding philosopher-psychologists who believed that emotions were mental states that manifested via physical changes, Willliam James, and simultaneously but independently Carl Lange, proposed that emotions were perceptions of bodily changes. James wrote, "My thesis on the contrary is that the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur IS the emtotion". Along with James, Wilhelm Wundt is one the major contributor the constructionist approach to emotion. According to Wundt [37], valence, arousal and intensity define a multidimensional affective space that people inhabit and are descriptive features of a unified state. He suggested that affective feelings were as influenced by externally-driven sensations (vision, hearing, touch, and so on) as by internally-generated sensations. Following these starting points, a key component in common between all constructionists' theories, is that emotions are physical compounds that are constructed out of more basic psychological, contextual and social ingredients that are not themselves specific to emotion [31, 38].

Fundamental questions	**Basic**	**Appraisal**	**Psychological Construction**	**Social Construction**
Are emotions unique mental states?	Yes	Yes	No	Varies by model
Are emotions caused by special mechanisms?	Yes	Varies by model	No	No
Is each emotion caused by a specific brain circuit?	Yes	No	No	No
Do emotions have unique manifestations (in face, voice, body state)?	Yes	Varies by model	No	No
Does each emotion have a unique response tendency?	Yes	In most models	No	No
Is experience a necessary feature of emotion?	Varies by model	Yes	Yes	No
What is universal?	Emotions are universal	Appraisals are universal	Psychological ingredients are universal	Influence of social context is universal
How important is variability in emotions?	Epiphenomenal	Varies by model	Emphasized	Present, but not central
Are emotions shared with non-human animals?	Yes	Some appraisals are shared	Affect is shared	No
How did the evolution shape emotions?	Specific emotions evolved	Cognitive appraisals evolved	Basic ingredients evolved	Cultural and social structure evolved

Table 1.1: Core assumptions of four emotion perspectives. Adapted from [32]

As illustrated in 1.2, between these two conflicting theories sits the appraisal theory. Firstly posited by David Irons [39], states that the essence of emotions is situated in a meaning analysis, which intervenes between the object and the resulting physical changes. This analysis is specific to a particular emotion, in contrast to Jame's idea that there are only general processes. Appraisal theories as we know them today, are attributed to the work of Arnold [40]. Appraisal models of emotion propose that emotions or emotional components are caused and differentiated by an appraisal of the stimulus as mis/matching with goals and expectations, as easy/difficult to control, and as caused by others, themselves or impersonal circumstances.

1.4 Which is the model

As witnessed by a recent review [41] a large number of papers (75%) that have been published in the field deals with the detection problem (Figure 1.5, top panel).

A study [42] has shown that human emotions are expressed mainly through facial expression (55%), voice (38%), and language (7%) in daily human communication. In fact, the majority of models focuses on behavioral signals, such as facial expression and voice expression. However, physical-based affect recognition may be ineffective due to the very involvement of active humans, who can conceal or tamper their real emotions, physiological signals (e.g., EEG, ECG, SCR) can generate more objective predictions. Recent works have shifted the attention on multi-modal fusion model [43,44], that integrates more than one signal to produce better descriptors.

As to the psychological models adopted, basically 50% of the works adopt variants of BET. These model adopt the discrete emotion model, also called categorical emotion model, which classifies emotions into limited emotion categories. The two most used models are the Ekman's six basic emotions and Plutchik's emotional wheel [45], which involves eight basic emotions and how these are related. Other works have adopted, in contrast, the concept of a continuous multi-dimensional model. This psychological approach establishes that emotions can be differentiated on the basis of dimensional parameters, such as arousal. and valence. Russell suggested the concept of *core affect* to explain and represent the combination of these two dimensions. Another instace of this approach is the three-dimensional framework proposed by Russell and Mehrebian [46], which includes the dimension of dominance.

Automated emotion recognition systems involve several steps for predicting accurate emotional states. The schematic view of the pipeline is depicted in 1.4. The detection problem is therefore solved through the classic pattern recognition pipeline:

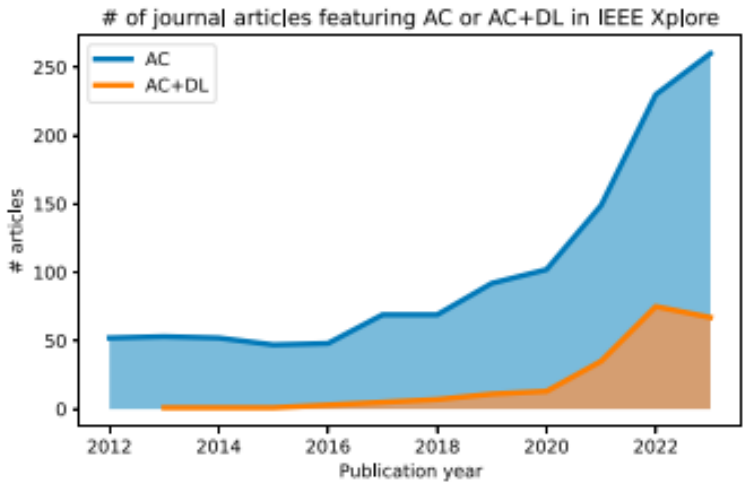

Figure 1.3: Number of journal publications appearing in IEEE Xplore between 2012 and 2023 featuring the terms "affective computing" (AC) or "affective computing" AND "deep learning" (AC + DL). Adapted from [48]

`signal` → `feature extraction` → `classification / regression`

where, more recently, supervised deep learning techniques applied to very large datasets are overcoming ML-based model, by joining the feature extraction and classification steps. The ML-based pipeline consists of pre-processing or raw signals, hand-crafted feature extractor, and well-designed classifier [28]. ML-based techniques for affective analysis are hard to be reused across similar problems on account of their task-specific and domain-specific feature descriptors. Due to their strong ability of feature representation learning, DL-based models have taken the spotlights. However, DL-based approaches have not yet had a huge impact on physiological emotion recognition, if compared with ML-based models [44].

Although public interest in this field continues to grow along with academic interest, as shown in 1.3, still a lack of trust in these automated models permeates the field. As a result, stakeholders, specialists, and physicians are hesitant to rely on existing models to make decisions. This is due to the inability of present emotion identification techniques to explain the predictions provided by decision support systems [47].

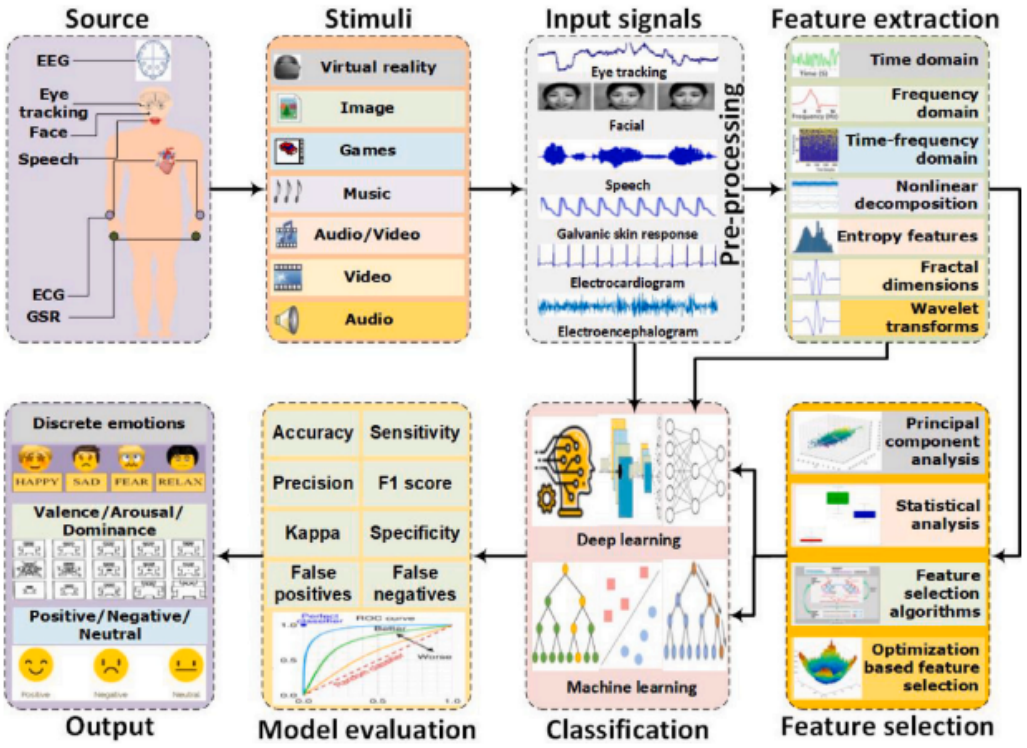

Figure 1.4: Schematic diagram of steps involved in an automated emotion recognition system. Adapted from [47]

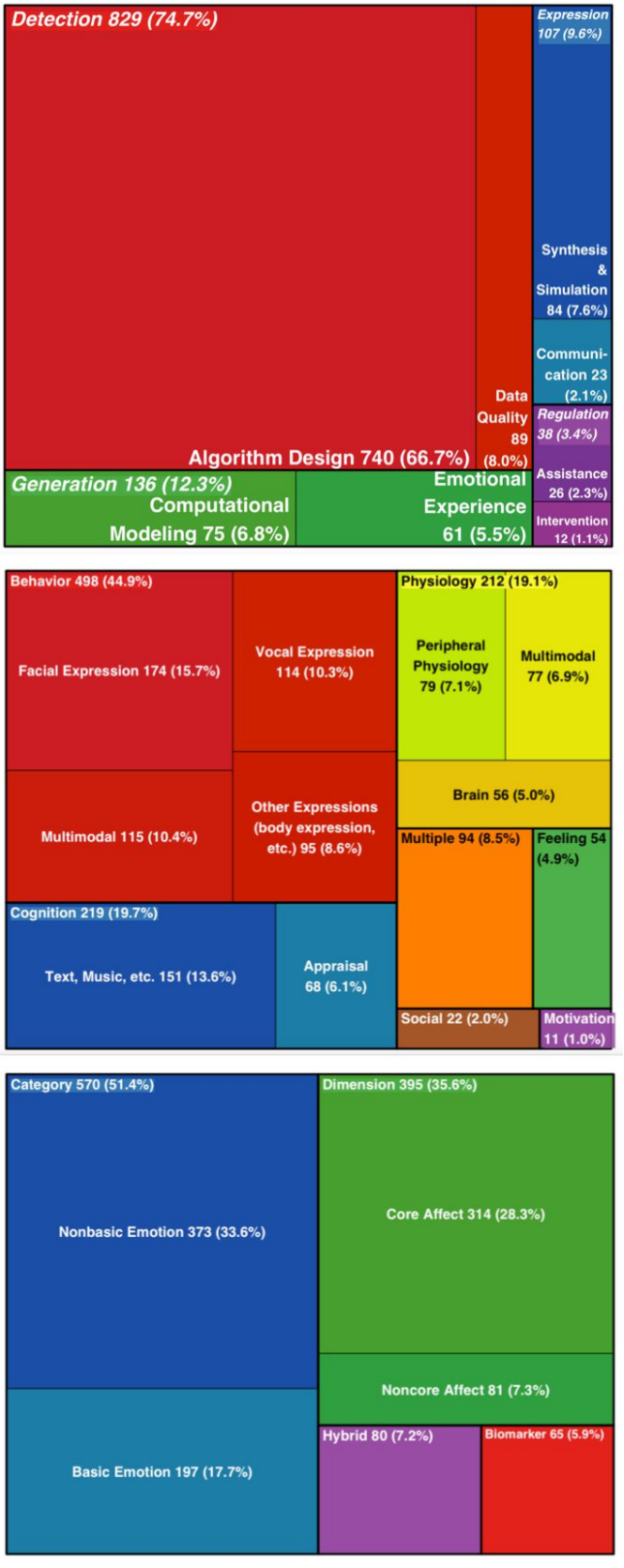

Figure 1.5: Distribution of papers over topics (top panel), modalities (centre), psychological models (bottom). Adapted from [41]

Chapter 2

Emotion theory

Constructionism, as briefly introduced in 1, is an emotion theory that challenges the classical view of emotions as physiologically innate responses. Instead, it argues that emotions are formed by a combination of psychological processes, such as social norms and personal experience. According to this view, emotions are stripped of their biological foundations, as pre-determined responses to stimuli, but rather are seen as dynamic phenomena that emerge from how individuals perceive and make sense of their internal and external contexts. The constructionist view of emotion includes social construction theories [49, 50], psychological construction theories [2,51] and descriptive appraisal theories [52] as well as the theory of constructed emotion that integrates social construction and psychological construction, as well as neuroconstructive and rational constructionist perspectives.

2.1 Constructivism overview

Psychological construction theories argue that emotions are not special mental states, generated by their own mechanisms, with their unique function or form, but are the same as other mental states such as cognition and perception. The starting assumption is that there are no subsystems in the brain dedicated to any emotion category or even emotion itself: the systems that realize emotions also realize all other mental states. By some psychological construction accounts, emotions (like all mental states) are categorizations of the self, based on multiple subpersonal systems such as perception, working memory, and affective state [53], making these views continuous with descriptive appraisal accounts found to the very right of the yellow zone 1.2.

The most prominent proposal here is currently represented by the Conceptual Act Theory [51]. According to the conceptual act theory, an emotional instance

is created when physiological changes in the body (or the accompanying affective experiences) are given psychological significance by being connected to or resulting from a situation in the outside world. The affective component reflects the *core affect* as posited by Russell, which entails the ever-present sense of feeling, the way via which we have access to out inner bodily sensations "which are themselves the representation and utilization of . . . the relevant statistical regularities . . . of the internal milieu" [54]. The core affect is an abstract representation of a psychologically primitive state develops along two dimensions:

1. pleasure vs. displeasure, spanning in a continuous scale from positive to negative.

2. high arousal vs. low arousal, measured along a continuous scale.

What people categorize as "fear", "anger", or "happiness" arise from the conceptualization of core affect through the lens of knowledge of emotion, or more specifically, an individual's understanding of the categories of fear, anger happiness, and so on. Within a moment, conceptualization occurs efficiently and automatically, integrating situation-specific emotion knowledge from past experience with internal sensory data from the body to create a psychologically meaningful state [56].

On the far-right part of 1.2 lie social constructionists theories of emotions. The thesis of social constructionism is that emotions are shaped by culture and society. According to this theory, emotions are not discrete entities but rather the result of the interaction of multiple factors (cognitive, motivational, and physiological). Furthermore, and in line with cognitive theories of emotion, it is believed that an individual's evaluation of the circumstance organizes the other elements of emotion. Some social construction models posit the "two social matters that impinge heavily on the personal experience of emotion" [57] are local language together with the local moral order. As implied by their choice of the term "local", the social constructionist perspective places a great emphasis on the cultural relativity of emotions. Emotions are performances of culture, rather than internal mental states. Whether a socially constructed event is seen as an emotion (as opposed to some other kind of psychological event) depends on the network of social consequences it produces.This prospective is well-summarized in:

> In the case of emotions, the overlay of cultural and linguistic factors on biology is so great that the physiological aspect of some emotional states has had to be relegated to a secondary status, as one among the effects of the more basic sociocultural phenomena. [57, p. 4]

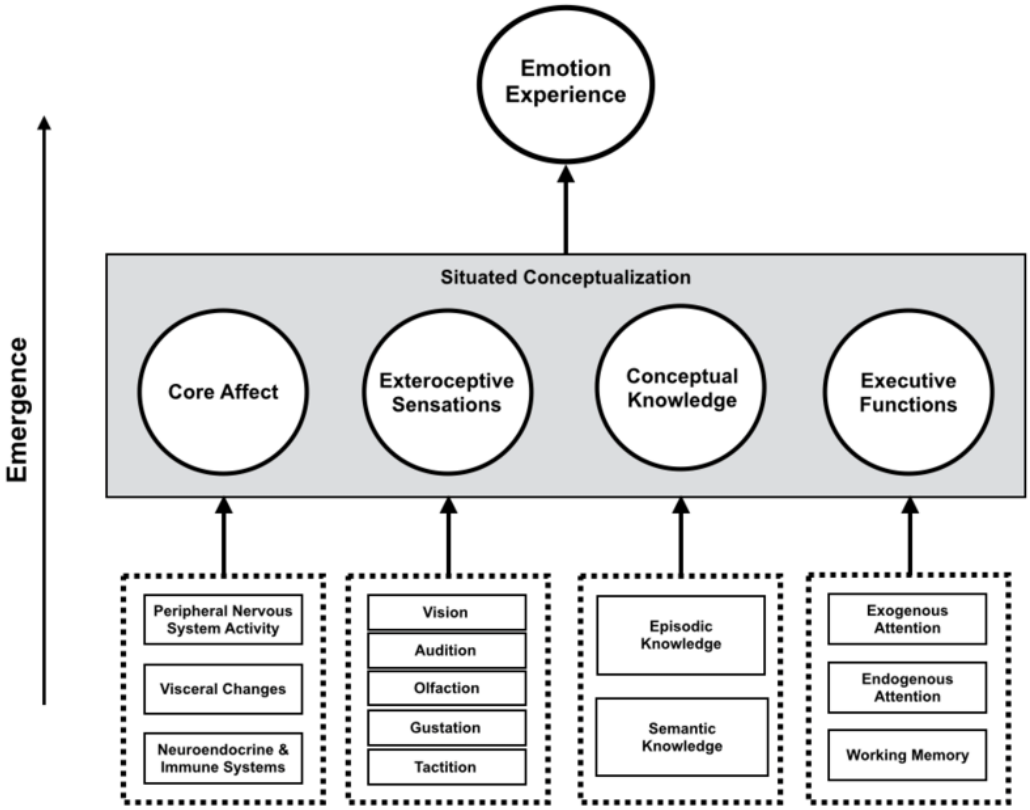

Figure 2.1: Measurable phenomena such as peripheral nervous system activity, visual sensations, episodic knowledge, exogenous attention, contribute to the more basic psychological components of emotion such as core affect, exteroceptive sensation, conceptual knowledge, and executive function. These components combine in a given context to create an emergent emotion experience. Adapted from [55]

2.2 Emotion as a conceptual act

Philosophers and psychologists believed that the mind is organized as a typology that includes Platonic emotional categories like fear, grief, rage, and so on. Emotions are thought to be fundamental components. Along this line of thought, scientists have searched for the corresponding physical essences of these kinds of emotions in patterns of response of the peripheral nervous system, in facial muscle movements, and in the structure or function of the mammalian brain, attempting to identify the "natural joints" that distinguish one type of emotion from another [51]. In the science of emotion, the majority of analytical techniques rely on stimulus-driven models, such as the stimulus → organism → response model, which states that an emotion's causal mechanism is "off" until it is turned "on" by a stimulus's (physical or appraised) characteristics. By these assumptions, BET strives to be one of the most used models in the affective computing economy.

As a matter of fact, a constructive analysis is more coherent with how the brain actually works: there is not underlying mechanism that follows a deterministic pattern of responses, but rather an ongoing, continually modified constructive process during which stored knowledge makes incoming sensory inputs meaningful. The process of synthesizing sensory information and stored knowledge is automatic and ongoing, with no sensation of agency or control over the visual experience.

Because our brains are hardwired to the facts of the physical world, there comes a period in our ontogeny when, as infants immersed in sensory information, the outside world seeds our earliest concepts. With the growth of our brains and the acquisition of language, we establish connections with others and ourselves, leading to the creation of purely mental concepts through cooperative communication.

A *concept* can be viewed as aggregated memories that accumulate for a category across experiences with its instances.

To put it simply, a *category* or *kind* is a collection of items (the category of "horse"). From a cognitive perspective, it specifies a collection of things or events that are considered comparable for a specific purpose or function in a given circumstance; consequently, a group of experiences with shared features/goal/action. One of the main functions of the brain is to categorize. According to Barrett [51, p. 89], categorization can be view as "comprising two processes":

- Accessing and activating a relevant category representation and binding it to a perceived instance;
- Drawing inferences from knowledge associated with the category and applying them to the instance.

To develop a categorization, or situated conceptualization, that best suits the circumstances and directs behavior, the brain draws on prior experiences. In

order to identify sensory inputs, establish a causal explanation for their cause, and generate action plans for addressing them, the brain continuously builds concepts and categorizes them. When something is categorized as part of one category and not another, it gains meaning. This is achieved by applying conceptual knowledge. At that point, it becomes feasible to make plausible inferences about it, determine the best course of action, and share our experiences with others [58]. For example, people who have only seen horses as part of a bucolic scenes, or as docile pets that are prone to be mounted by humans, the image of a horse is calming and cute. For these people, a normal course of actions might be approaching the horse. For others who witnessed only the wildest part of a horse, who have been bitten with the resultant pain and swelling, the same image might evoke in them a fearful reaction.

One of the main functions of the brain is to categorize its internal and external milieu. Since categorization is enactive and prepares you for the most suitable course of actions, it always produces some kind of automatic changes in the body.

Exteroception represent sensory changes in the external world. External perceptions are gathered through sight hearing and touch or through the proprioception of self-movement and body position (often used in robotics). However, along with exteroceptive processing, interoceptive processing give birth to agents' feelings of affect, and influences every action the agents performs.

Interoception here denotes the sensory data that collectively describe the constantly changing physiological state of the body, arising from the allostatic regulation of various bodily systems, including the autonomic nervous system, the endocrine system, and the immune system [59]. Interoception enables the agent's brain to construct the environment in which the agent lives and, eventually, to give meaning to words. Deprived of interoception, without affect and feelings, the agent would be unlikely to survive for long [54].

The concepts used during categorization can be thought of as tools to modify and regulate the body, to create feelings and dispositions towards actions. The brain evolved this way to efficiently maintain energy regulation in the body. Such regulation is provided by two complementary processes: homeostasis and allostasis.

Homeostasis is the ongoing maintenance and defense of vital physiological variables such as blood pressure and blood sugar [60, 61]. This was a reactive method in which a controlled variable was recognized to have been perturbed from its optimal level, leading to the elicitation of corrective responses that acted to bring the variable back to its pre-perturbation levels.

Allostasis is the elementary ongoing process of optimizing the body's internal milieu: the brain anticipates the needs of the body and attempts to meet those needs before they arise; more precisely, it is the capacity to vary physiological systems flexibly according to predicted energy demands. Energy regulation (e.g,

metabolism) is likely to be at the core of the human mind, regardless of whether a person is thinking, feeling or perceiving.

The concept of allostasis—the process by which the brain regulates the body's internal state to meet changing demands—relies on the brain's ability to anticipate and respond to shifting physiological needs. As an animal's physiological state fluctuates throughout the day, its recent experiences shape which sensory inputs are prioritized, determining what aspects of the environment are most relevant in the present moment. This continual adjustment influences the animal's future interactions with its environment, guiding behavior in ways that optimize survival.

At the core of this process is the *predictive brain*. The brain doesn't just react to immediate sensory inputs but also uses past experiences and internal representations to predict future states. By continuously updating its predictions, the brain fine-tunes its understanding of both the internal body and external world, ensuring that its responses are timely and adaptive. Thus, the brain's ability to regulate through allostasis is tightly linked to its predictive capacity [62, 63].

For a brain to effectively regulate the body within its environment, it maintains an internal model that integrates the body's physiological state with the outside world [1]. This model is shaped by the body's needs, meaning the brain interprets the world through the lens of maintaining physiological balance. As a result, the brain not only tracks the statistical regularities of the external environment, but also monitors the consistent patterns of its internal state. In brief, the brain's internal model runs on concepts that categorize sensations to give them meaning. In assembling populations of predictions, each carrying a certain probability of being the best match for the current circumstances (known as Bayesian priors), the brain is essentially constructing concepts—what Barsalou refers to as "ad hoc" concepts [65]. These are flexible, context-specific groupings that help the brain make sense of the present moment. Incoming sensory information acts as prediction error, serving to either refine or adjust these predictions. Certain predictions will align better with the sensory input because they are based on stronger priors, meaning they are more likely to match past experiences. As a result, the brain categorizes new sensory events as similar to previous encounters, allowing it to make sense of the world by referencing its internal models [66].

This formulation of the brain being a predictive machine has gained popularity in cognitive and theoretical neuroscience against the traditional approach of it being a deterministic framework. Brain's simulations function as a Bayesian filters for incoming sensory inputs. According to these theories, the brain is comparable with a probabilistic machine, which, using past experiences as a guide, prepares

[1] There is a well-known principle in cybernetics: every good regulator of a system must be a model of that system. Thus, anything that regulates a system, like a brain with its body, must contain an 'internal model' of that system [64].

multiple competing simulations that answer the ever-present question "what is this new sensory input most similar to?" [67]. According to Bayesian logic, the brain uses pattern completion to decide among simulations and decides which to implement, based on the impact on physiological efficiency [54]. In every waking moment, the brain gives sensations, either exteroceptive or interoceptive, meaning. When, we focus on some of those sensations that are interoceptive ones, the resulting meaning can be an instance of emotion.

The interoceptive network generates predictions about the body's internal state, tests these simulations against actual sensory input from the body, and updates the brain's model of the body in the world accordingly. Interoceptive sensations are typically experienced as lower-dimensional feelings of affect, specifically in terms of valence (pleasantness/unpleasantness) and arousal (intensity) [51]. If interoception plays a central role in allostasis, and allostasis forms the foundation of the brain's computational architecture, then affective properties—valence and arousal—should be considered fundamental aspects of consciousness itself, rather than merely features of emotion. This suggests that affect is intrinsic to how the brain regulates the body and experiences the world [68].

This is exactly the take of the CAT view: an instance of emotion is constructed the same way as all other perceptions are constructed. There are clearly some differences between the categorization involved in seeing or hearing, and the categorization of an emotion. The former include an object which the perception is about, and this object existed before and is independent of the perception. In the case of emotions, the input to the categorization process does not include anything that counts as an emotion before the act of categorization [53]. In fact, emotions are objects created by the process of categorization. That is, when one categorizes himself as afraid, he is creating an object, fear. The basic point here, is that category knowledge and categorization constitute emotions by adding novel epistemic properties to sensory inputs and actions. The knowledge behind has been learned through language, socialization, stories and other cultural artifacts within the person's day-to-day experience. On this basis, the Conceptual Act Theory of emotion can be encapsulated as follows [51]

- Sensory inputs are categorized using conceptual knowledge from past experience. The process of combining incoming sensory input with stored knowledge is obligatory and automatic.

- The prior knowledge used in perception is enactive, as a consequence of predictive coding the brain performs perceptual inference.

- The inferential process induced by categorization prepares for situated action.

- Categorization, being enactive and preparing for specific actions, produces some kind of automatic change in the physical state of the agent. In the same way of perception, the brain makes predictions and meanings of bodily sensations.

- The process meaning making rarely happens because of a deliberate, conscious goal to figure things out.

The conceptual act is thus the process of applying prior knowledge to incoming sensory input. This process is an active one, rather than a passive event, because the agent is not simply detecting or experiencing what is happening in the world or within the body. Instead, the agent's prior knowledge actively contributes to shaping the momentary experience. Thus, every conceptual act is embodied, as prior experiences, stored as category knowledge, are activated through sensory and motor neurons. This activation not only influences bodily states but also modulates their representations and sensory processing, making the experience deeply intertwined with the body's physiological state and past interactions [54]. Physiological changes occur all the time in the body: blood pressure goes up and down, temperature is high and low or voluntary muscles contract. However, only sometimes these changes are perceived as being causally related to surrounding events, and as this happens, an emotion is constructed. The Conceptual Act Theory, therefore, introduce in the equation of emotion emergence the interoceptive sensations, as parallel to the exteroceptive sensation. The brain, at the same time, tries to tie together the inner and outer world, so that a person can make sense of both, equally.

Interoception provides the brain with continuous sensory feedback about body's internal physiological states. These signals are then integrated and processed, resulting in the psychological primitive called core affect. Core affect reflects the basic, embodied dimensions of valence (pleasure/displeasure) and arousal (activation levels). This kind of affect is referred to as "core" because it is grounded in the internal milieu, an integrated sensory representation of the physiological state of the body: the somatovisceral, kinesthetic, proprioceptive, and neurochemical fluctuations that take place within the core of the body. Affect can be seen as the basic "building block" of emotion, or any other mental state. Affect provides the underlying feeling tone, while emotions are the more refined, structured responses that integrate affect with cognitive and contextual factors. Therefore, while affect contributes to emotion, it is not synonymous with it. As [38] put it:

> Core affect [...] represents a basic kind of psychological meaning. The basic acoustical properties of animal calls (and human voices) directly act on the nervous system of the perceiving animal to change its affective state and in so doing conveys the meaning of the sound [...] All

> words (regardless of language) have an affective dimension of meaning, so that people cannot communicate without also (often inadvertently) communicating something about their affective state. Learning a new language fluently does not merely require making a link between the phonological forms of words and their denotation, but a connection to affective changes must also be forged.

In this perspective, while Conceptual Act Theory is often described as a psychological construction theory of emotion, it positions emotion on equal footing with cognition and perception: the hypothesis about ad hoc concepts is not specific to emotion, but holds for every action one might take and every experience one might have, whether mental states it is classified as. The same processes that create in the brain the instance of an emotion, are at work when creating non-emotions states [30]. Any instance of emotion encompasses various features, such as feelings of pleasure or displeasure, comfort or discomfort, and arousal or sleepiness. In my work, we refer to these experiential features as affect, which arise from the brain's ongoing allostatic regulation of the body. Affect reflects the brain's constant effort to maintain physiological balance, meaning it is a broader phenomenon that is not limited to specific emotional experiences. Rather, affect serves as the foundational background of bodily sensations that accompany all states of consciousness, including but not exclusive to emotions. *Perception* refers to psychological moments when the focus is on interpreting externally driven sensations to understand what they signify in the world. *Cognition*, on the other hand, refers to moments when the brain focuses on retrieving and reinstating prior experiences. If this mental activity involves recalling specific past experiences, it is called *memory*; if not, it is referred to as *thinking*. When the mental activity is directed toward future possibilities, it is known as *imagining*. *Emotion* represents psychological moments in which the focus is on interpreting internal sensations from the body, aiming to understand what these internal signals represent in relation to one's physiological and affective state. Core affective changes, conceptualization and executive attention are only but three psychological primitives that are constantly in play, continually shaping one another as they combine like ingredients to make a variety of mental states [56].

Indeed, the novel perspective of the predictive brain, is the most suitable to account for the overall framework on language and emotions that we have outlined in previous chapters. The basic assumptions of such perspective can be summarised as follows [66, 69]:

1. *The brain is not for thinking (or loving).* The brain's most important function is not centered on rationality, emotion, or imagination, but on controlling the body to manage allostasis. By generating predictions about the body's future needs, the brain allows the agent to act efficiently, ensuring

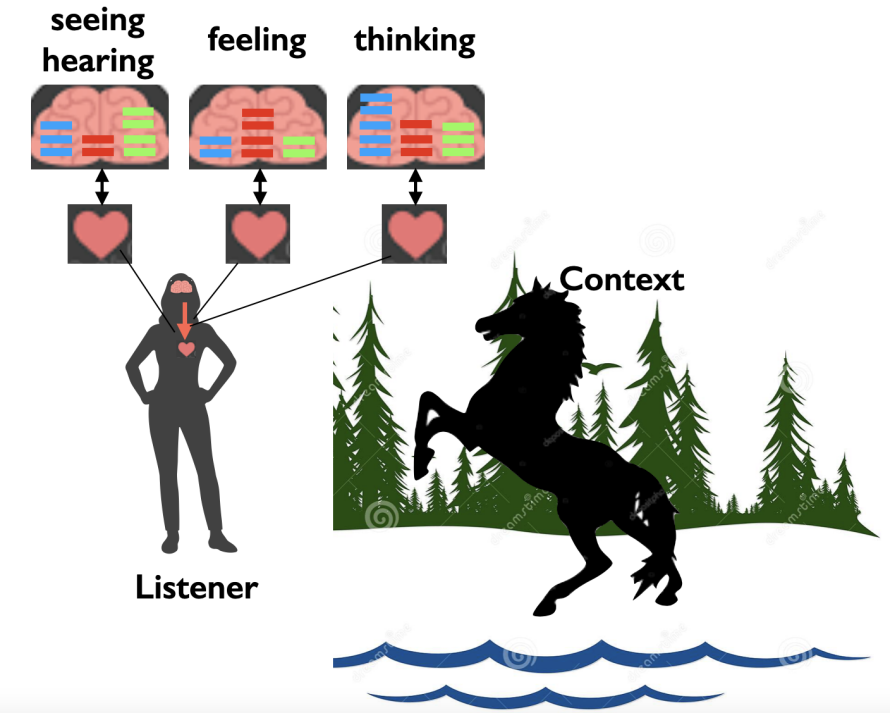

Figure 2.2: Mental states. Mental states, iconically depicted as brain states, comprised of different combinations of the same three psychological primitives (represented in red, green, and blue). Depending on the recipe (the combination and relative weighting of psychological primitives in a given instance) and a psychologist's interest, mental states are called seeing/hearing or thinking or feeling.

survival. This predictive regulation helps the brain maintain homeostasis, optimizing the body's internal balance in response to both external challenges and internal physiological shifts.

2. *There is one and only brain.* It has not phylogenetically evolved by adding layers of specific functions with increasing complexity, as postulated by MacLean. But evolution does not work in the name of perfection, it is not a ladder with human beings at the top of it [70].

3. *The brain is a network.* Basically, a network of 128 billion neurons, by and large, connected as a single, massive flexible structure, resulting in trillions of activity patterns, where connections become stronger or weaker depending on what is happening in the world and in the agent's body throughout life. No single neuron or area is the locus of a single psychological function (vision, touch, reasoning, memory, etc.).

4. *The brain develops by wiring itself to its world.* Ontogenetically, infant's genes carrying neural wiring instructions, are guided and regulated by the surrounding physical and social environments that help in tuning and pruning neural connection in order to manage the body budget. In such endeavour, a brain becomes optimized for the particular environment in which it develops.

5. *The agent's brain predicts what the agent does.* Moment by moment, predictions are exploited to test conceptual representations against the incoming, buzzing sensory evidence - from the external world and from the body - to categorize it according to past experience (prior knowledge/ memories/learned representations), in the effort of anticipating body's needs and preparing the optimal "actions" to satisfy those needs before they arise.

6. *A brain works with other brains.* In social species, agents regulate one another's body budgets through their (inter)actions. Humans are unique in the animal kingdom, because they can afford regulation with words. Many regions involved in language also control the proper body (e.g., areas of the "language network" are involved in heart regulation). This kind of regulation is a powerful one since it can be performed across distances and time (e.g., a phone call or reading an ancient text)

7. *One brain makes more than one kind of mind.* Agents come into the world with a basic brain plan that can be wired in a variety of ways. Beyond the individual, micro-wiring is tuned and pruned by social groups and culture.

8. *Brains can create reality.* Boundaries between social and physical realities is porous (e.g., studies showing that people judge wine as tasting better when expensive). Brains do not just select information from the environment, but by creating categories add new functions to the world. These are communicated and shared with other brains and weaved into the world to become part of the social environment, which, in turn, will help to wire novel brains.

2.2.1 A primitive: core affect

We have previously introduced the psychological primitive "core affect". In modern psychological usage, "affect" refers to the mental counterpart of internal bodily representations associated with emotions, actions that involve some degree of motivation, intensity, and even personality dispositions. In the science of emotion, "affect" is a general term that has come to mean anything emotional. A cautious term, it allows reference to something's effect or someones internal state without specifying exactly what kind of an effect or state it is.

Russell [71] defined core affect as "a neurophysiological state that is consciously accessible as a simple, nonreflective feeling that is an integral blend of hedonic (pleasure-displeasure) and arousal (sleepy-activated) values". At any given moment, the consciuous experience is a floating point in the core affect space (2.3).

The horizontal dimension, pleasure-displeasure, ranging from one extreme (e.g., agony) to its opposite (e.g., ecstasy) is the assessment of one's current condition. The vertical dimension, arousal, ranges from sleep to energetic activation, passing through various stages of alertness, is the sense of one's mobilization and energy. A person always has a core affect: it can be neutral, moderate or extreme. Changes can be short or long lasting. Intense core affect can dominate consciousness, becoming the central focus of a person's experience. In contrast, milder core affect typically exists as part of the background, subtly influencing one's conscious world without demanding full attention. When core affect changes rapidly or significantly, it moves to the forefront of awareness, capturing conscious focus. As the intensity of core affect weakens or stabilizes, it gradually fades into the background. In moments when core affect is neutral and stable, it may even disappear from consciousness altogether, allowing other aspects of experience to take precedence [71].

The process of changing core affect is not fully understood. There exist many theories regarding what valence and arousal entail, and how they are embodied and how these these two quantities are correlated, if not. However, many agrees that these two values are influenced by the brain's inner model of the body, and are strictly related to interoception. Arousal is usually themed with novelty [73]. According to this theory, the novelty of an event, thus how much that event was

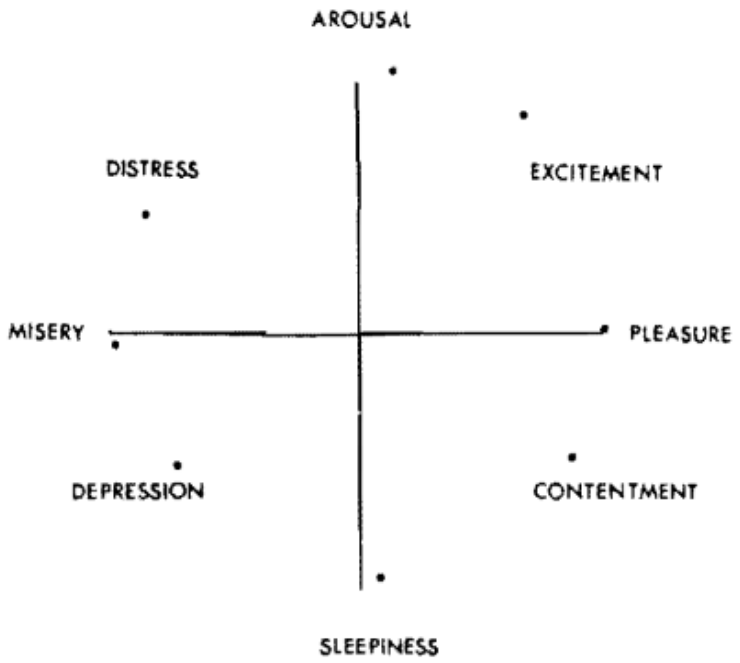

Figure 2.3: The circumplex of emotions. The horizontal (east-west) dimension in this spatial metaphor is the pleasure-displeasure dimension, and the vertical (north-south) dimension is arousal-sleep. [26]

unexpected by the perceiver as to his internal model, is a source of arousal potential. An appropriate level of arousal potential might induce a positive hedonic response, but an extreme arousal potential might induce an overwhelming feeling and a negative responses [72]. An event with no information causes no arousal, conversely, excessive information gain, such that one can hardly cope, causes discomfort. In [72, 74, 75] a mathematical formulation of arousal has been proposed.

On the other hand, valence, has been formalized as being related to the confidence one has on its internal model [7]. Valence has a similar derivation (yet, different computation) to arousal, yet distinct implication. When an event occurs that favors the agent's internal model, and thus, the agent is confident in reaching its goal, then it will feel a positive valenced state, otherwise, its confidence in his model, or confidence over all, plummets.

Valence and arousal are inherently linked, via the hedonic function of the arousal potential, the so-called *Wundt curve* [76](Fig. 2.4). In this formulation, valence is a function of arousal.

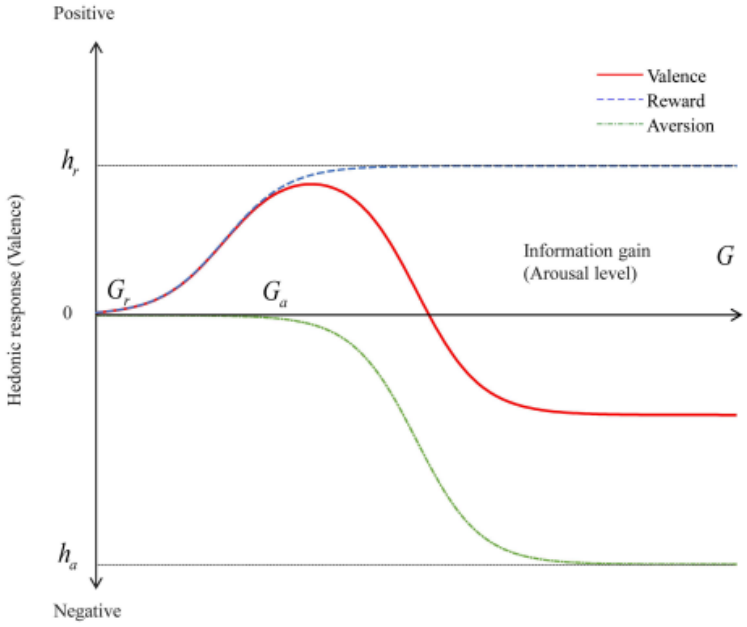

Figure 2.4: Valence as a function of information gain. The valence is modeled as a summation of two sigmoidal functions representing reward and aversion systems [72].

2.3 Constructing emotions

From the overall picture we've outlined, it becomes clear that searching for a specific neurobiological basis of emotion, in terms of pinpointing a defined set of brain areas or a distinct network that provides an emotion "fingerprint," is misguided. Emotion is not localized to a single brain region or fixed circuit. Instead, emotions are constructed dynamically through the brain's integration of interoception, core affect, past experiences, and contextual factors. On one side, we have sensations—both from the external world and from within the body; on the other side, we have a categorization process that drives the brain's ability to make predictions. The brain's core function is to generate these predictions to give sensations meaning, all in the service of allostasis, or maintaining the body's internal balance. Emotions, like fear or happiness, emerge depending on how sensations are categorized at a given moment and on how the brain weighs the relevance of the core affect state, characterized by valence (pleasantness) and arousal (intensity). These affective states are shaped by ongoing interoceptive (internal) and exteroceptive (external) inputs. Ultimately, emotion is simply the label for psychological moments when the brain focuses on interpreting what these internal sensations from the body represent.

This way researchers can talk about emotion in a theory-neutral fashion. Under such circumstances, if one observes the "neural reference space" of core affect (Fig. 2.5), this might be considered as the neural underpinning of emotion.

This neural reference space can be subdivided into two related functional networks [38]:

- *Sensory integration network*: establishes an experience-dependent, value-based representation of an object that includes both external sensory features of an object along with its impact on the homeostatic state of the body. It includes the cortical aspects of the amygdala (specifically, the basolateral complex (BL)), the central and lateral portions of OFC, as well as most of the adjacent agranular insular areas. The sensory integration network has robust connections with unimodal association areas of many sensory modalities, including the anterior insula that represents interoceptive sensations.

- *Visceromotor network*: it is part of a functional circuit that guides autonomic, endocrine, and behavioral responses to an object. It includes the medial portions of the OFC (extending into what is sometimes called the vmPFC), as well as subgenual and pregenual areas of the ACC, with robust reciprocal connections to all limbic areas (including many nuclei within the amygdala, and the ventral striatum), as well as to the hypothalamus, midbrain, brainstem, and spinal cord areas that are involved in internal-state regulation. These areas modulate changes in the viscera associated with the autonomic nervous system (including tissues and organs made of smooth muscle, such as the heart and lungs) and neuroendocrine changes that affect the same organs by way of the chemicals released into the bloodstream via hypothalamic regulation of the pituitary gland. In addition, the visceromotor network (particularly the vmPFC) is important for altering simple stimulus-reinforcer associations via extinction or reversal learning and appears to be useful for decisions based on intuitions and feelings rather than on explicit rules, including guesses and familiarity based discriminations.

To sum up, parts of the affective circuitry are intricately interconnected with sensory cortical areas, while others are closely linked to regions that control autonomic and hormonal responses necessary for regulating the body's homeostasis. The highly re-entrant nature of neural activity—where signals loop and feedback between brain areas and the body—makes it challenging to establish clear cause-and-effect relationships. This complexity blurs the lines between sensory and affective processing, as well as between the brain's influence on the body and the body's feedback to the brain. As a result, the interaction between sensory inputs and affective states is highly dynamic and reciprocal.

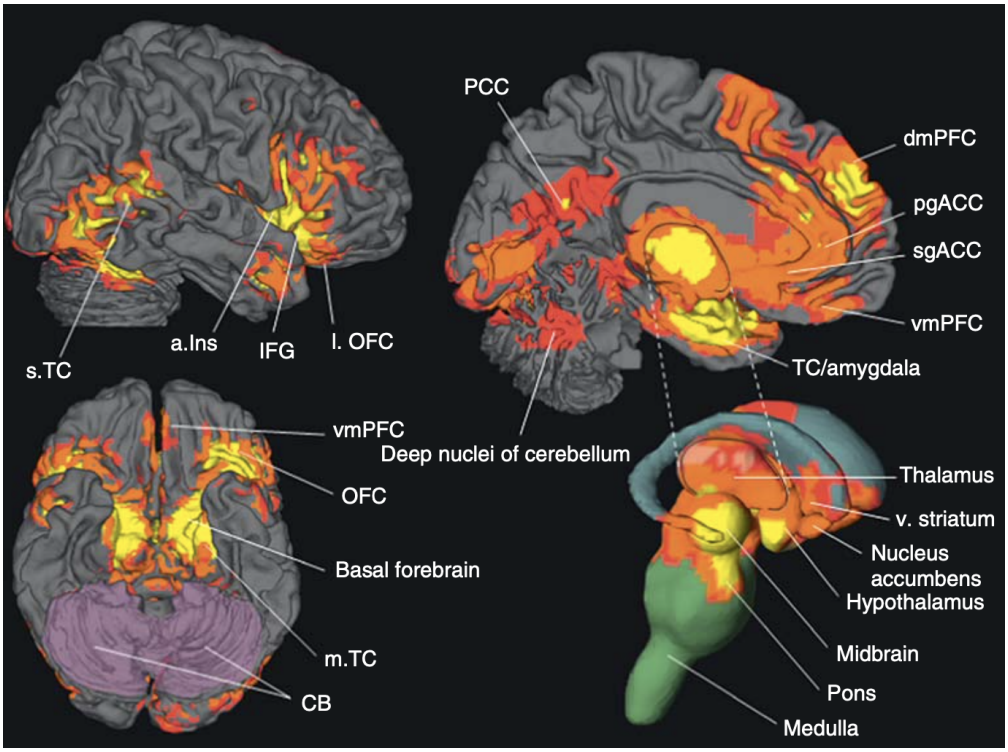

Figure 2.5: Neural reference space for core affect. 165 neuroimaging studies of emotion (58 using PET and 107 using fMRI) summarized in a multilevel meta-analysis to produce the observed neural reference space for emotion. These areas include (from top left, clockwise) anterior insula (aIns), lateral OFC (lOFC), pregenual cingulate cortex (pgACC), subgenual cingulate cortex (sgACC), ventral medial prefrontal cortex (vmPFC), temporal cortex/amygdala (TC/Amygdala), thalamus, ventral striatrum (v Striatum), nucleus accumbens, hypothalamus, midbrain, pons, medulla, OFC, and basal forebrain. Other areas shown in this figure (e.g., inferior frontal gyrus (IFG), superior temporal cortex (sTC), dorsal medial prefrontal cortex (dmPFC), posterior cingulate cortex (PCC), medial temporal cortex (mTC), and cerebellum (CB)) relate to other psychological processes involved with emotion perception and experience. From [38]

The key concept here is that the circuitry within the neural reference space for core affect binds sensory information from the external world to sensory information from the body, so that every mental state is intrinsically infused with affective content.

In Wundt's theoretical framework, affect is conceptualized as a fundamental feeling state, representing a psychological primitive and a core component of human consciousness. As one of the basic elements of mental life, affect serves as a foundational building block, integral to the broader structure of psychological experience.

In this context, core affect functions as a driver of attentional processes within the human brain, where attention is understood as any factor that modulates neuronal firing, either by increasing or decreasing it. However, affect, by itself, does not provide specific information regarding the nature of external changes, their location, or appropriate responses to them. Instead, it operates as a rudimentary mechanism, akin to a "sixth sense," signaling that something significant has occurred, without offering further detail or guidance on the nature of the event.

Chapter 3

A unifying theory

So far we have depicted a picture regarding the brain as being a tool for predictions rather than reactions. On this view, the brain is not merely an elaborate stimulus-response mechanism, but rather a sophisticated statistical organ that actively generates and tests hypotheses to explain the sensory stimuli it encounters.

Instead of passively reacting to sensory input, the brain constructs predictions about the world, which are continually evaluated against incoming sensory evidence. This process allows the brain to infer the causes of sensory data, forming a dynamic and ongoing cycle of hypothesis generation and testing, where perception and cognition are driven by the constant comparison between predicted and actual sensory experiences.

This perspective can be traced back to the Helmholtzian formulation of perceptive inference [77]. The underlying idea is that the brain maintains a model of the world that tries to optimize using sensory inputs. Remember that this idea is equally defined, but with intrinsically differences, to both the environments the brain has to deal with; thus perception and input, refer to exteroceptive and interoceptive stimuli.

In this view, the brain is an inference machine that actively predicts and explains its sensations. In this context, terms like "explanations", "hypotheses", and "beliefs" should not be interpreted as consciously held mental states, but rather as neuronally encoded probability distributions, commonly referred to as Bayesian beliefs, that represent the brain's estimations about the hidden causes of sensory signals. These beliefs are implicit, operating at the level of neural circuits, and involve the brain calculating the likelihood of different possible causes of incoming sensory data. Through this process, the brain is constantly updating these probabilistic representations in light of new evidence, allowing it to refine its predictions and reduce uncertainty in a Bayesian manner [78].

3.1 Predictive coding

Current formulations of Helmholtz's notion of perception as unconscious inference have become some of the most influential metaphors for neuronal processing. These ideas are now primarily considered within the framework of the Bayesian brain hypothesis, which suggests that the brain functions as a probabilistic machine, constantly predicting sensory input based on prior experiences and updating these predictions as new data becomes available. This approach is widely referred to as predictive coding, where the brain minimizes the difference between expected and actual sensory input—prediction error—through a process of Bayesian inference [78].

Predictive coding is a process theory in computational and cognitive neuroscience, which proposes a potential unifying theory of cortical function, namely that the core function of the brain is to minimize the prediction error, arising from the discrepancy between the predicted input and the input actually received [78–80]. This process of minimization can be achieved in multiple ways:

- Through immediate inference about the hidden state of the world, which can explain perception.
- Through updating the brain's internal model about the world to make better predictions, which could explain learning.
- Through the search for sensory data that realize the predictions, which could explain adaptive behavior, or in general action.

The core intuition behind predictive coding is that the brain is composed of a hierarchy of layers [81], which each make predictions about the activity of the layer immediately below them in the hierarchy. In this framework, perceptual and cognitive processes are understood as the outcome of a computational trade-off between two types of processing: top-down and bottom-up.

In this framework, descending predictions are generated at each level of the brain's hierarchical processing and are compared with representations at lower levels. These predictions are based on multi-level generative models of the environmental causes of sensory signals, and they represent the brain's expectations about what sensory data should be encountered. When there is a mismatch between these predictions and the actual sensory input, this difference is encoded as a prediction error.

Prediction errors, which signal the discrepancy between predicted and actual sensory information, are then fed back up the hierarchy to update higher-level representations. This bottom-up signal informs higher levels about where the predictions were inaccurate, prompting an adjustment in the brain's internal model.

Over time, this recurrent exchange between adjacent hierarchical levels facilitates a continuous updating process, with higher levels generating more abstract, global predictions (e.g., recognizing objects or scenes), and lower levels focusing on finer details and local variations in sensory data. By resolving prediction errors at multiple hierarchical levels, the brain develops a deep, hierarchically structured explanation for the incoming sensory inputs.

This process replaces traditional bottom-up models of perception, where sensory data was thought to accumulate evidence to form the percepted object. Instead, in predictive coding, perceptual content is primarily shaped by top-down predictive signals, which are continually updated and refined by bottom-up prediction errors as sensory evidence is compared to expectations across all levels of the hierarchy. [78, 82].

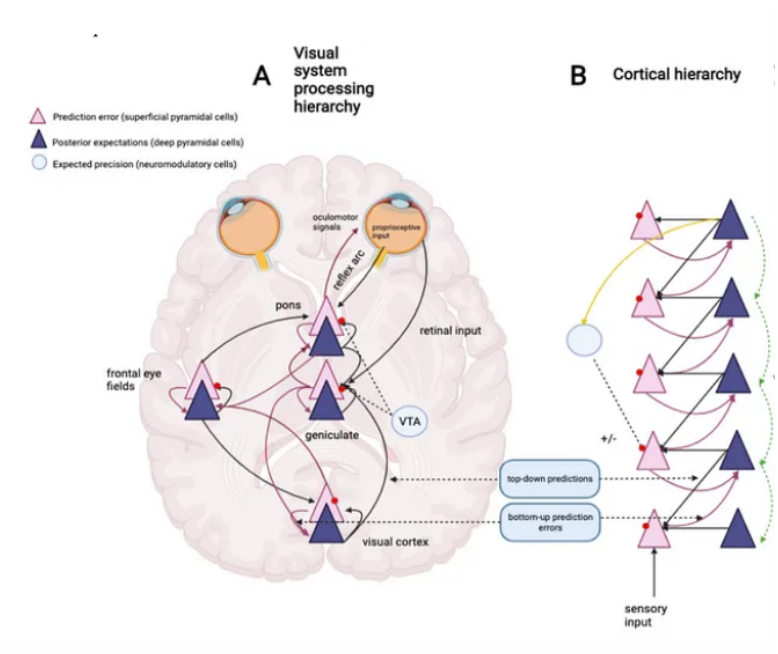

Figure 3.1: The simplified schematic of a cortical hierarchy (B) highlights hierarchical processing architectures, depicting expectation units and error units with prediction errors being fed forward and predictions reflecting the current perceptual hypothesis being fed back down the cortical hierarchy. Prediction errors are contextualised through precision modulation by neuromodulatory cells (blue circle). In the visual system (A) and other sensory cortices, we see morphological and functional asymmetry between forward and backward projections. Adapted from [83]

In computational terms, the activity of neuronal populations is understood to encode Bayesian beliefs, which are essentially probability distributions over the possible states of the world that cause sensory inputs. These beliefs represent the brain's probabilistic understanding of the external world, grounded in the notion that the brain must infer the hidden causes of sensory data. The simplest way to encode these beliefs is through the expected value (mean) of a hidden cause or expectation. These hidden causes refer to the underlying factors that generate sensory experiences but cannot be directly observed. Instead, they must be inferred from the sensory information they produce. As such, these causes remain hidden behind a sensory veil, meaning that the brain's knowledge of the world is always indirect, requiring it to continuously estimate and update these hidden states based on the sensory consequences it encounters [78].

3.1.1 Exact Bayesian Inference

In order to support adaptive behaviour, the brain has to overcome the ambiguous relationship between sensory data and their underlying hidden causes in the world. In predictive coding, the brain is understood as instantiating a generative model, which is generally speaking, a model of the process that generated the sensory data of interest.

The generative model is defined as the joint probability of the "observable" data (y) and an hypothesis (x) about these data (trees, birds, glasses etc.). In other words, a generative model is the product of $P(x)$ (priors over states) and $P(y|x)$ (likelihood of evidence probability if the hypothesis is true). This means that the generative model is a statistical model of how observations are generated (strictly speaking, a description of causal dependencies in the environment and their relation to sensory signal). It uses prior distributions $P(x)$ (which determine the probability of hypothesis before evidence) that the system applies to the environment about which it makes inferences [84].

Perception then becomes the process of inverting the causation process, thus, exploting the given sensory data to access the hidden causes that generated them (Fig. 3.2). In other words, an inference process. This inversion is the same as minimizing the difference between the recognition and posterior densities to suppress the prediction error.

To minimize prediction errors, the generative model continuously creates statistical predictions about what is happening or can happen in the world. The mechanism through which the brain updates its beliefs can be described using Bayesian inference. In Bayesian terms, the brain updates the likelihood of different hypotheses of sensory input by integrating new evidence (i.e., sensory data and prediction errors) with its prior beliefs. The Bayesian rule provides a formal framework for updating these probabilities, where priors (pre-existing beliefs) are

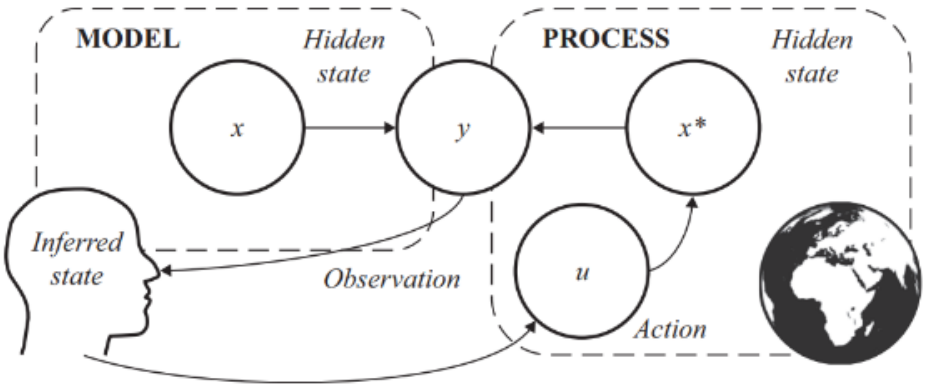

Figure 3.2: Generative process and generative model. Both represent ways in which sensory data (y) could be generated given hidden states (x) and are represented through arrows from x to y to indicate causality. The difference is that the process is the true causal structure by which data are generated, while the model is a construct used to draw inferences about the causes of data (i.e., use observations to derive inferred states). In other words, the models we use to explain our sensorium may include hidden states that do not exist in the outside world, and vice versa. Actions are here depicted as part of the generative process, as we said before, action are one of the possible way the agent has to minimize prediction error. Adapted from [1]

adjusted in light of new information, yielding more accurate or updated posterior beliefs.

Technically speaking, from the generative model

$$P(x, y) = P(y|x)P(x) \tag{1}$$

we can compute the true posterior directly via the Bayes rule

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \tag{2}$$

The generative model calculates the posterior probability $P(x|y)$, which in practice allows the system to assume the most probable hypothesis explaining the nature and causes of the sensory signal, taking into account the available sensory data. However, the normalizing factor $P(y)$ is often computationally intractable because $P(y) = \int P(x, y)dx$ requires an integration over all latent variable states x. The marginal $P(y)$ is often referred to as the *evidence*, since it effectively scores the likelihood of the data under a given model, averaged over all possible values of the model parameters [85].

3.1.2 Variational Inference

Exact inference is generally computationally intractable or even infeasible. As a consequence, the brain might model approximate Bayesian inference rather than computing it exactly [80].

Variational inference [86, 87] aims to approximate the exact posterior using an auxiliary posterior $Q(x|y;\theta)$ with parameters θ. The posterior distribution can be of any form and is under the control of the modeller. The goal is to fit the approximate posterior to the exact posterior minimizing the divergence between them, with respect to the parameters. Mathematically

$$Q^*(x|y;\theta) = \arg\min_{\theta} D_{KL}[Q(x|y;\theta)||P(x|y)] \tag{3}$$

the goal is to minimize the KL-divergence between the two distributions [2], according to the parameters. By doing this, we have replaced the inference problem of computing the posterior with an optimization problem of minimizing this divergence.

Note that equation 3 still contains the intractable true posterior. The strength of variational inference is that it instead optimizes a tractable upper bound on this divergence, the *variational free energy* (VFE)[3]. This way, the prediction error is always bounded by a (computable) quantity and by iteratively updating the approximate posterior (initially arbitrary), one can find a distribution that approximates the exact posterior.

3.2 Free-Energy principle

The seminal work of Rao and Ballard in 1999 [88] was impactful because it built on existing theoretical principles, specifically those identified by Mumford [89], to create a small-scale predictive coding network. Through their model, they empirically investigated the dynamic relationship between predictions and prediction errors in neural systems. This model provided an explanation for complex neurophysiological phenomena, particularly addressing extra-classical receptive field effects, such as the behavior of endstopping neurons, which had been previously difficult to explain. Their research demonstrated that a predictive coding network, which includes both bottom-up prediction error neurons and top-down predictive neurons, could replicate several "extra-classical" receptive field properties, such

[2]The exact solution, for which $D_{KL}[Q(x|y;\theta)||P(x|y)] = 0$, only exists when the family of variational posteriors considered includes the true posterior as a member. For example, if both true and approximate posterior are Gaussian

[3]In machine learning, this quantity is known as the (negative) evidence lower bound (ELBO). Instead of being minimized, it is maximized.

as endstopping. This was achieved through the hierarchical structure of the predictive coding network, where top-down predictions interact with sensory input and help refine the brain's understanding of external stimuli. Moreover, in this model, prediction error, value estimation, and weight updates follow from gradient descents on a single energy function [85].

This intuition, was later extended by Frinston [79,90], grounding it more firmly in the theoretical framework of variational inference algorithms, as well as integrating predictive coding with the broader free energy principle (FEP) [82] by identifying the above said energy function with the variational free energy of variational inference. This link, allow us to interpret the predictive coding update rule as performing approximate Bayesian inference.

The FEP posits that any self-organizing system in equilibrium with its environment must minimize its free energy. This principle serves as a mathematical formulation that explains how adaptive systems, such as biological agents (e.g., animals or brains), maintain their internal order, as evidenced by their homeostatic properties, and must therefore minimise the occurrence of atypical events in their living environment. In essence, by minimizing free energy, these systems are able to predict and adapt to sensory inputs from their environment in a way that reduces uncertainty or surprise. This adaptive process enables organisms to survive and function efficiently by aligning their internal models with the external world, effectively ensuring that their internal states remain within homeostatic bounds. As this theory explains, neuronal processes can implement free energy minimisation by modifying sensory input through action on the world, or by updating internal models through perception, with implications for understanding the dynamics and interactions between action, perception and learning.

The defining characteristic of biological systems is their ability to maintain their states and structure despite being in a constantly changing environment. From the brain's perspective, this environment encompasses both the external (sensory input from the world) and the internal (bodily or interoceptive signals) milieu. To preserve this stability, the probability distribution over the sensory states of such systems must exhibit low entropy. In mathematical terms, this implies that the system is likely to occupy only a small number of possible states, with a high probability, and is unlikely to be in the remaining states, which have a low probability.

Entropy in this context represents the average self-information or "surprise," which can be understood as the degree of unpredictability or uncertainty in sensory states. Formally, entropy is the negative logarithm of the probability of an outcome: states that are more surprising have lower probabilities and contribute more to the system's overall entropy. To maintain homeostasis, biological systems aim to minimize entropy, reducing surprise by predicting and controlling their

interactions with the environment [82].

As already stated, free energy is an upper bound on surprise. Mathematically speaking [78], this upper bound can be derived applying the Bayes rule directly to the KL-divergence in 3:

$$\begin{aligned} D_{KL}[Q(x|y;\theta)||P(x|y)] &= D_{KL}[Q(x|y;\theta)||\frac{P(y,x)}{P(y)}] \\ &= D_{KL}[Q(x|y;\theta)||P(y,x)] + \mathbb{E}_{Q(x|y;\theta)}[\ln P(y)] \\ &= D_{KL}[Q(x|y;\theta)||P(y,x)] + \ln P(y) \\ &\leq D_{KL}[Q(x|y;\theta)||P(y,x)] = \mathbf{F} \end{aligned} \tag{4}$$

Where in third line the expectation around $\ln P(y)$ vanishes since this quantity does not depend on x. Finally, VFE is an upper bound because $\ln P(y)$ is necessarily non-positive since $0 \geq P(y) \geq 1$. This means that minimising VFE is the same as maximizing model evidence, without getting tangled in infeasible computations. Importantly, $\mathbf{F}$ is a tractable quantity, since it emerges as the divergence between two quantities that we assume before.

3.2.1 A discussion on VFE

The problem of exact Bayesian inference now becomes a problem of optimization: the minimization of VFE. Variational free energy may seem at first an abstract concept, a reformulation of the starting problem, but its nature and the role it plays in Active Inference (as we will see in Ch.4) become more clear when decomposed into quantities that are more intuitive [1].

It can be expressed as a functional $F[Q, y]$ of the approximate posterior Q and a function of data y:

$$\begin{aligned} F[Q,y] &= -\mathbb{E}_{Q(x)}[\ln P(y,x)] - H[Q(x)] \\ &= D_{KL}[Q(x)||P(x)] - \mathbb{E}_{Q(x)}[\ln P(y|x)] \\ &= D_{KL}[Q(x)||P(x|y)] - \ln P(y) \end{aligned} \tag{5}$$

The first line of equation 5 decompose VFE in *energy* and *entropy*. It shows that minimizing with respect to Q requires consistency with the generative model while also maintaining a high posterior entropy. In other words, when precise prior beliefs are not available, we should adopt maximally be uncertain.

The second line emphasizes the interpretation of free energy minimization as finding the best explanation for sensory data, which must be the simplest explanation that is able to accurately account for the data. In fact, the first term is referred to as *complexity*, while the second accuracy. Inferring explanations that have minimal complexity has a cognitive importance: the updating of the model

entails a certain cognitive cost; hence, an explanation that diverges minimally from the prior is preferable.

The last line rephrases VFE as an upper bound on the evidence. This last decomposition shows, that perception and learning are not the only things an agent can do to lower its entropy, but it can also change the evidence part through acting to change sensory data.

The connection between predictive coding and the VFE provides insight into how the generative model reduces prediction errors through approximate Bayesian inference. This process can be interpreted as the neural information processing mechanisms effectively performing variational inference. Predictive mechanisms can be described in terms of the realization of variational principles. To implement a variational inference algorithm, it is necessary to define the forms of both the variational posterior and the generative model. In the context of predictive coding, this implies that the posterior probability densities are typically modeled as normal (Gaussian) distributions. With this assumption, free energy can be interpreted as the sum of the long-term average prediction error, which is thought to be associated with the FEP. In the context of predictive coding, this implies that the system minimizes long-term average prediction error by optimizing the statistics of the approximate posterior distribution through its generative model. This is an important observation for the very understanding of predictive processing because it allows us to think about the normative function of the predictive mechanisms, which is the long-term average precision-weighted error in terms of free energy minimization [84].

3.3 Neurobiological underpinnings of Predictive Coding

Although technically predictive coding is simply a variational inference and filtering algorithm under Gaussian assumptions, it has been claimed from the beginning to be a biologically plausible theory of cortical computation, and the literature has consistently drawn close connections between the theory and potential computations that can be performed in the brain.

Each cortical area such as V1, V2 or V4 is comprised of 6 internal layers: L1-L6. It is not yet clear how these six cortical layers are connected to each other, and may subtly vary between cortical regions and across species, but it can be assumed that they are organised according to the central microcircuit model of predictive coding [91]. In this scheme the six cortical layers can be split into:

- An input layer L4 which primarily receives driving excitatory inputs from the area below as well as from the thalamus.

- A deep feedback stream consisting of layers L5 and L6.
- A superficial feed forward flow, consisting of layers L1, L2 and L3.

The superficial layers of the cortex send excitatory connections forward to L4 of higher cortical areas, while the deep layers provide feedback connections, which can be either inhibitory or excitatory, to both the deep and superficial layers of lower areas. Within each cortical area, there is a well-established three-step feedback relay: input is received in L4, transferred to the superficial layers L2/3, and then projected forward to the next area in the hierarchy. From L2/3, the superficial layers send signals to the deep L5, which can either project to L6 or provide feedback to lower hierarchical regions. Interestingly, deep L5 and L6 are the only cortical layers which contains neurons which project to subcortical regions or the brainstem, and L6 especially appears to maintain precise reciprocal connectivity with the thalamus [78].

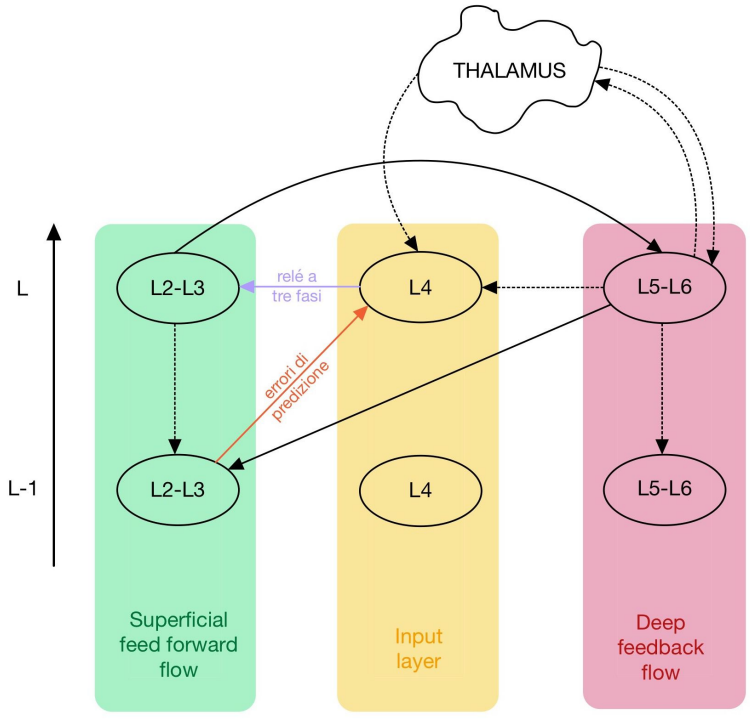

Figure 3.3: Canonical microcircuit proposed by Bastos et al. [91]

The fundamental mechanism of predictive coding involves sending predictions down the cortical hierarchy, while prediction errors are transmitted upwards. The Bastos microcircuit model 3.3 associates a layer of predictive coding, with a 6-level cortical region.

The inputs to L4 of the region are taken to be the prediction errors of the region below, which are then immediately passed upwards to the superficial levels L2/L3 where the prediction error and the value neurons are taken to be located. The predictions are taken to reside in the deep layers L5/6.

Overall, significant progress has been made in translating the abstract mathematical framework of predictive coding into neurophysiologically realistic neural circuits. However, many open questions and important challenges remain. The alignment with cortical microcircuitry is not perfect, and certain cortical pathways remain difficult to explain with current models. Despite these limitations, predictive coding theories offer some of the most compelling and general process models for linking cortical microcircuitry to an abstract computational framework, one that has demonstrated the ability to solve complex cognitive tasks.

Chapter 4

Unifying perception, action and learning

The discussion up to this point is common to every Bayesian brain theory. Variational inference provides a theoretical framework in which Bayesian optimality can be rewritten in terms of optimizing a certain quantity that the FEP identifies as the VFE. As we anticipated in chapter 3, this quantity not only depend on the model but on the data as well: by acting on the world to change the way in which data are generated, we can ensure a model is fit for purpose by choosing those data that are least surprising under our model [1].

In this picture, perception is well encapsulated and understood both as a tool and objective of surprise minimization. Active Inference (AcI) [92–95] extends predictive coding to consider generative models of actions.

In this perspective, AcI posits that perception and action not only have the same inferential nature, but that both serve the very same imperative: minimize VFE. Recalling from Eq. 5 that VFE can be decomposed as the sum of model evidence and the discrepancy between model and world: when an organism minimizes its divergence, through perception, then free energy becomes an approximation to surprise. When an organism, on the other hand, changes its sensory inputs, by action, gathers more observations to render them more similar to its preferences, in other words, minimizes surprise.

4.1 Active Inference

Active Inference offers a unified mathematical framework for modeling perception, learning, and decision-making. This framework treats each of these psychological processes and their interactions as interdependent forms of inference. AcI is based on the premise that every living organism has to endeavor to maintain a certain

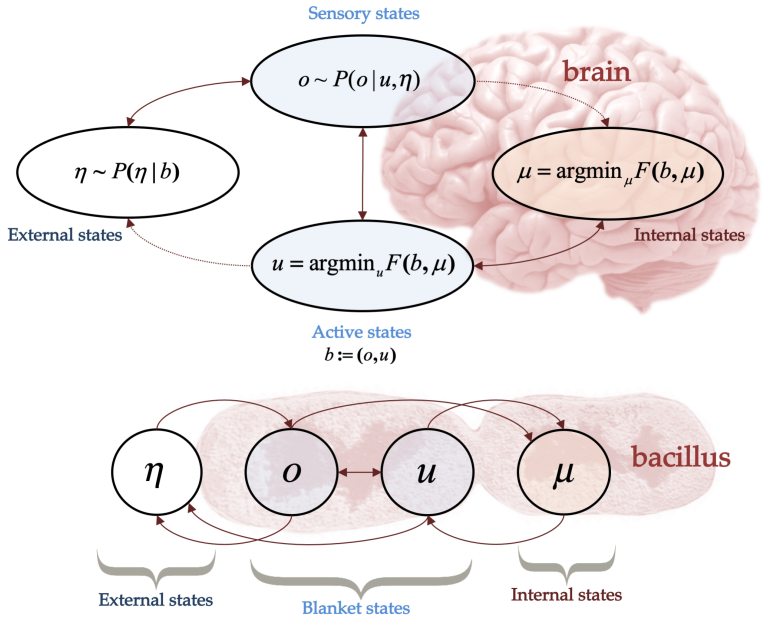

Figure 4.1: A Markov blanket (b) defines the boundary between internal (μ) and external (η) states of a system, mediating their interaction via sensory (o) and active (u) states, collectively known as *particular states*. Sensory states represent environmental inputs, while active states represent the system's influence on the environment. This structure, which forms a Bayesian network, applies to systems ranging from simple organisms to the brain. Adapted from [92]

order and predictability, despite the fact of being immersed in an ever-changing environment. A straightforward manifestation is homeostasis, which controls that fundamental physiological parameters remain within viable regions [1]. In other words, decision-making agents are assumed to infer the probability of various external states and events in the environment—including their own actions—by integrating prior beliefs with sensory input. In contrast to more classical interpretation of perception, which are *passive* (e.g., inferring the presence of an external object based on patterns of light that impinge on the retina), the inference process here is considered *active*, because the agent infers the actions most likely to generate preferred sensory out: it actively changes the environment (e.g., inferring that

opening the fridge will reveal available food options) [94]. This leads decision-making to favor actions that optimize a trade-off between maximizing reward and information gain.

Active Inference postulates that these processes can be understood as optimizing two complementary objective functions: variational free energy, which measures how well an internal model fits past sensory observations (Section 3.2.1), and expected free energy (EFE), which evaluates possible future actions based on prior preferences [92]. This dual optimization enables agents to align their behavior with both past experiences and future goals.

Active Inference is a *first principle* account, which describes the dynamics of systems that persist during a certain time scale of interest and that can be statistically segregated from their environment. The former condition means that there exists a steady-state probability density to which the system self-organizes and returns after perturbation, namely, its preferred states. The latter condition implies the presence of a Markov blanket [96] (see Figure 4.1): a set of random variables through which the internal and external states interact. Under these assumptions it can be shown that the states internal to the system parameterise Bayesian beliefs about external states and can be cast a process of variational free energy minimisation

4.2 The generative model

The generative model expresses how the agent represents the world. In simple terms, it is just a way of formalizing beliefs about the way outcomes are caused. Usually a generative model is specified in terms of the likelihood of each outcome, given their causes and the prior probability of those causes. The specific type of generative models considered here is a partially observable Markov decision process (POMDP). The term POMDP denotes two major concepts: the first is partial observability, which means that observations may only provide probabilistic information about the states of the world. The second is the Markov property, which simply means that beliefs about the current state of the world are all that matter for an agent when deciding which actions to take.

Another important characteristic of generative models used in AcI is that they are dynamic, in the sense that they evolve over time as new observations are gathered, and the observations that are added to the model depend (via action) on beliefs about variables in the model .

To formally frame this process, we need to account for the form of each distributions involved in the model (the square factors in Figure 4.2). The classic

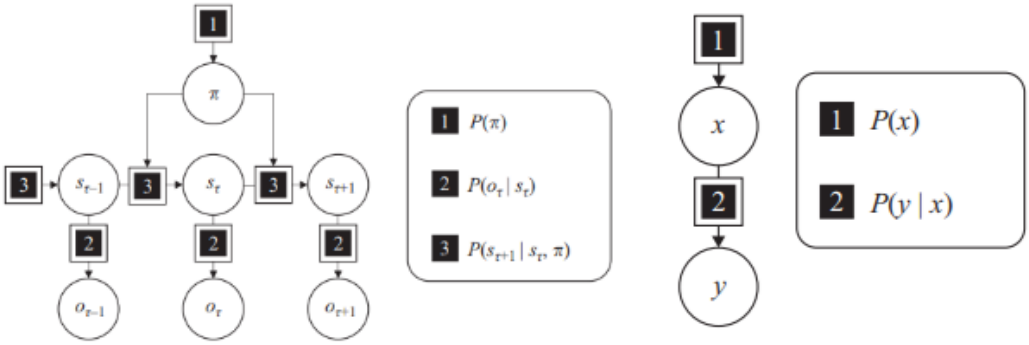

Figure 4.2: Differences between a static generative model (right) and a dynamic generative model (left). The models are represented as graphical probabilistic models: the circles represents random variables, the squares represent the probability distributions that describe the relationship between these quantity and the arrow represents the conditioning. Adapted from [1]

generative model is defined as follows:

$$P(o_{0:t}, s_{0:T}, \pi, A, B, D, \gamma) = P(\pi|\gamma)P(\gamma)P(A)P(B)P(s_0|D)P(D) \prod_{\tau=0}^{t} P(o_\tau|s_\tau) \prod_{\tau=1}^{T} P(s_\tau|s_{\tau-1}, \pi) \tag{6}$$

$$\begin{aligned} P(o_t|s_t) &= Cat(A) \\ P(s_\tau|s_{\tau-1}, \pi_{\tau-1}) &= Cat(B) \\ P(s_1|s_0) &= Cat(D) \\ P(\pi|\gamma) &= \sigma(-\gamma G(\pi)) \\ P(A) &= Dir(a) \\ P(B) &= Dir(b) \\ P(D) &= Dir(d) \\ P(\gamma) &= \Gamma(1, \beta) \end{aligned} \tag{7}$$

Here, $G(\pi)$ is is the free energy expected under each policy (see below). The other parameter descriptions can be seen in 4.1.

Active Inference agents perform inference by optimizing the parameters of an approximate (variational) posterior distribution rather than the true posterior. The variational distribution leverages independence between latent variables in

what is known as mean-field approximation[4]: the variational distribution is assumed to fully factorize, meaning that all the latent variables are considered independent, except for the hidden states and the policy. This leads to the following factorization:

$$Q(s_{0:T}, \pi, A, B, D, \gamma) = Q(\pi)Q(A)Q(B)Q(D)Q(\gamma)\prod_{\tau=0}^{T} Q(s_\tau|\pi) \tag{8}$$

$$\begin{aligned} Q(s_\tau|\pi) &= Cat(s_\tau^\pi) \\ Q(\gamma) &= \Gamma(1, \beta) \\ Q(\pi) &= Cat(\pi) \\ Q(A) &= Dir(a) \\ Q(D) &= Dir(d) \\ Q(B) &= Dir(b) \end{aligned} \tag{9}$$

In this model, observations depend only on the current state, while state transitions depend on a policy or sequence of actions. The random variable π represents all possible policies up to a given time horizon T and each policy is defined as a sequence of actions. Importantly, the policy π is a random variable that has to be inferred. In other words, the agent entertains competing hypotheses of its behavior in terms of policies. This contrasts with standard formulations in which a single state-action policy returns an action as a function of each state.

All the posterior probabilities over model parameters, including the initial state, are Dirichlet distributions. The Dirichlet distribution is used as a prior over the parameters of the categorical distribution due to its role as the conjugate prior for the categorical distribution. This property ensures that when a categorical distribution is multiplied by a Dirichlet distribution, and the result is normalized to obtain the posterior distribution over the parameters of the categorical distribution, the posterior remains a Dirichlet distribution. This allows the Dirichlet distribution to be reused as a prior in subsequent rounds of inference, enabling Active Inference agents to sequentially update their beliefs about model parameters as they receive new observations [94]. The sufficient statistics of these distributions are concentration parameters that can be regarded as the number of occurrences encountered in the past, and allow and easy update formulation.

[4]Assumes that the approximate posterior factorizes into the product of (independent) distributions: $Q(X) = \prod_i Q(X_i)$, where X is the set of all hidden variables, X_i represents the i^{th} hidden variable. This approximation often works well in practice, but it has the limitation of ignoring possible pairwise (or more complex) interactions between variables.

Table 4.1: Glossary of terms and notation

Notation	Description
S	Finite set of hidden states
s_τ	(Hidden) state at time τ
$s_{0:T}$	Sequence of hidden states $s_0, ..., s_T$
O	Set of all possible outcomes
o_τ	Outcome at time τ
$o_{0:t}$	Sequence of outcomes $o_0, ..., o_t$
T	Number of timesteps in a trial
U	Set of all possible actions
π	Policy
s_τ^π	Hidden states conditioned on each policy
Q	Approximate posterior distribution
F	Variational free energy
G	Expected free energy
Cat	Categorical distribution
Dir	Dirichlet distribution
Γ	Gamma distribution
A	Likelihood matrix
B	Matrix of transition probabilities
D	Prior expectation of the hidden state at the beginning of each trial
a, b, d, β	Parameters of prior
σ	Softmax function or normalised exponential

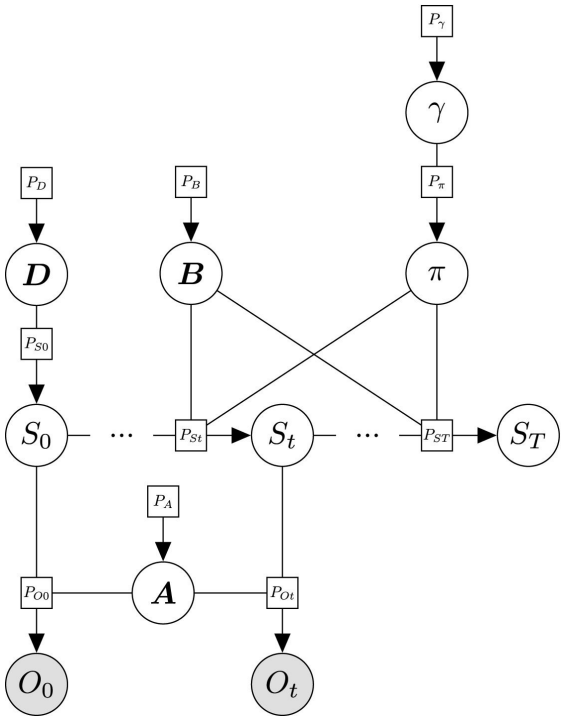

Figure 4.3: The variables highlighted in grey can be observed by the agent, while the remaining variables are inferred through approximate Bayesian inference. Adapted from [98]

Lastly, the precision parameter γ has been associated with the neuromodulator dopamine through what is called the "precision hypothesis" [97]. This parameter is a precision estimate for the expected free energy over policies. It can be thought of as encoding a prior belief about the confidence with which policies can be inferred.

4.2.1 Bayesian model reduction

An agent that navigate its environment does not only entertains a single model of the world, but to fully understand the causes of its sensations, it must compare different hypothesis about how sensory data are generated, and retain only the simplest and most compelling model. In Bayesian statistics, these processes are called Bayesian model comparison and Bayesian model selection: these correspond to scoring the evidence for various generative models in relation to available data and selecting the one with the highest evidence. Bayesian model reduction is a specific form of structure learning that formalizes after-trials hypothesis testing to simplify the generative model. It eliminates redundant explanations of sensory

data, ensuring that the model generalizes effectively to new data. From a technical perspective, it involves estimating the evidence for simpler, reduced priors over latent causes, and selecting the model with the highest evidence [92]. Removing certain states or parameters from the model has a clear biological parallelism in terms of synaptic decay and switching off certain synaptic connections: it is implemented by homeostatic synaptic adjustment processes during sleep and resting wakefulness.

4.2.2 Deep temporal models

A deep temporal model [99] is a generative model with many layers that are nested hierarchically and act at different timescales.

To easily grasp the idea behind, there is a useful metaphor: consider the the whole model as a clock, and each layer as a hand of it. In a two-layer hierarchical model, a ticking (resp. rotation) of the faster hand corresponds to a time step (resp. trial of observation epochs) at the lower level. At the end of each trial at the lower level, the slower hand ticks once, which corresponds to a time-step at the higher level, and the process unfolds again. One can concisely summarize this by saying that a state at the higher level corresponds to a trial of observation epochs at the lower level [92]. In principle, these models can be extended to an arbitrary number of levels, accounting for a deeply structured world with dynamics that unfold at different time scales.

In these models, posterior state representations at the lower level are treated as observations at the higher level. State representations at the higher level, in turn, provide prior expectations over subsequent states at the lower level. This means that higher level state representations evolve more slowly, as they must accumulate evidence from sequences of state inferences at the lower level [7].

4.3 VFE and EFE

In Active Inference, all the heavy lifting is done by minimizing free energy with respect to expectations about hidden states, policies, and parameters. This is because different policies, through their impact on hidden states in the generative process, make certain observations more likely than others.

This implies that both the approximate posterior $Q(s|\pi)$, and the generative model $P(s|\pi)$, are conditioned on policies.

At this point we can frame the VFE introduced in 3.2.1, in the Active Inference framework, considering a particular given time point $\tau \in \{0, ..., T\}$, whence the agent has observed a sequence of outcomes $o_{0:\tau}$:

$$
\begin{aligned}
F_\pi &= \mathbb{E}_{Q(s|\pi)}\left[\ln \frac{Q(s|\pi)}{P(o,s|\pi)}\right] \\
&= \mathbb{E}_{Q(s|\pi)}[\ln Q(s|\pi) - \ln P(o,s|\pi)] \\
&= \mathbb{E}_{Q(s|\pi)}[\ln Q(s|\pi) - \ln P(s|\pi)] - \mathbb{E}_{Q(s|\pi)}[\ln P(o|s,\pi)] \\
&= \underbrace{D_{KL}[Q(s|\pi)||P(s|\pi)]}_{\text{Complexity}} - \underbrace{\mathbb{E}_{Q(s|\pi)}[\ln P(o|\pi)]}_{\text{Accuracy}}
\end{aligned} \tag{10}
$$

Observe that we obtain (Eq. 10) the same *complexity* vs *accuracy* decomposition as before. However, Active Inference is not solely concerned with minimizing prediction error in perception. It is also a model of action selection.

Variational free energy has a retrospective aspect, as it is a function of past and present, but not future, observations. Although it facilitates inferences about the future based on past data, it does not directly facilitate prospective forms of inference based on anticipated future data [1], which is fundamental in planning and decision-making.

To infer optimal actions, an agent predicts sequences of future states and expected observations for each possible policy, and then EFE associated with those different sequences of future states and observations.

As a model of decision-making, EFE serves as a criterion for evaluating different policies, allowing the agent to select the best one. In Active Inference, this is formally accomplished by introducing prior expectations over observations, denoted as $P(o|C)$, which represent the agent's preferences. These prior expectations encode the outcomes the agent considers desirable. The policy that is expected to generate observations most aligned with these preferences will maximize the model's accuracy, thereby minimizing the EFE. Consequently, the probability of selecting a particular policy is inferred based on how well the expected sequence of observations under that policy match the agent's preferred outcomes, thus maximizing model accuracy.

The expected free energy of a policy follows from the second line of equation 10, where (expected) observations, which have not yet occurred, enter the expectation operator as a random variable:

$$
\begin{aligned}
G_\pi &= \mathbb{E}_{Q(o,s|\pi)}[\ln Q(s|\pi) - \ln P(o,s|\pi)] \\
&= \mathbb{E}_{Q(o,s|\pi)}[\ln Q(s|\pi) - \ln P(s|o,\pi)] - \mathbb{E}_{Q(o|\pi)}[\ln P(o|\pi)] \\
&\approx \mathbb{E}_{Q(0,s|\pi)}[\ln Q(s|\pi) - \ln Q(s|o,\pi)] - \mathbb{E}_{Q(o|\pi)}[\ln P(o|C)] \\
&= -\underbrace{\mathbb{E}_{Q(o,s|\pi)}[\ln Q(s|o,\pi) - \ln Q(s|\pi)]}_{\text{Epistemic value}} - \underbrace{\mathbb{E}_{Q(o|\pi)}[\ln P(o|C)]}_{\text{Pragmatic value}}
\end{aligned} \tag{11}
$$

The key point of the EFE derivation, is that it eliminates the conditionalization on π in the second term and instead it conditions on the preferences C: the agent's preferences can be independent of the policy being followed, which allows us to drop the conditionalization.

4.3.1 More on the Expected free energy

EFE can be unpacked in different fashions, each of which highlights its importance in relation to many existing theories in neuroscience and engineering.

The first of this is perhaps the most intuitive (Eq.11), as it expresses the classic exploit-explore dilemma in behavioral psychology (or machine learning). The epistemic value, which ensures that free energy is an upper bound on surprise, represents the information gain. Note that here, the divergence assumes negative values, thus, we want to select those policies that maximize the expected divergence-hence, information gain. This way, EFE penalizes those observations for which there is a many-to-one mapping from observations to states as it precludes gathering other information [1]. In other words, in most uncertain situations, one must first perform epistemic actions to resolve uncertainty before confidently selecting a pragmatic action.

Another very common expression of EFE in the Active Inference literature is:

$$G_\pi = \underbrace{D_{KL}[Q(o|\pi)||P(o|C)]}_{\text{Risk (outcomes)}} + \underbrace{\mathbb{E}_Q(s|\pi)[H[P(o|s)]]}_{\text{Ambiguity}} \tag{12}$$

The first term on the right-hand side of this equation quantifies the expected divergence (KL divergence) between the beliefs about the probability of a sequence of outcomes under a given policy and the preferred outcomes, which are specified a priori in the model. This term is often referred to as *risk* (or expected complexity), but can be more intuitively understood as representing the reward the agent will get with respect to each choice. In other words, the smaller the expected divergence between the preferred outcomes and the outcomes anticipated under a policy, the higher the likelihood of achieving rewarding outcomes by following that policy. The second term is the expected value of the entropy of the likelihood. This term is referred to as *ambiguity*, because it scores how much the given observation reflects the beliefs of the agent about the hidden states of the world. Policies that minimize ambiguity are expected to bring forth the most informative observations.

$$\begin{aligned} G_\pi &= D_{KL}[Q(o|\pi)||P(o|C)] + \mathbb{E}_Q(s|\pi)[H[P(o|s)]] \\ &\leq \underbrace{D_{KL}[Q(s|\pi)||P(s|C)]}_{\text{Risk (states)}} + \underbrace{\mathbb{E}_Q(s|\pi)[H[P(o|s)]]}_{\text{Ambiguity}} \end{aligned} \tag{13}$$

An alternative formulation can be expressed if the model is equipped instead with a prior preferences about states. This way, policies will be deemed more likely if they bring about states that conform to prior preferences.

In summary, expected free energy is defined with respect to prior beliefs about future outcomes, which establish the expected cost or complexity and complete the generative model. These priors provide inference and action with a purposeful or goal-directed nature, as they embody the agent's preferences or goals. These preferences characterize agents by the specific states they prefer occupying, and through action, they tend to realize and revisit these preferred states [4].

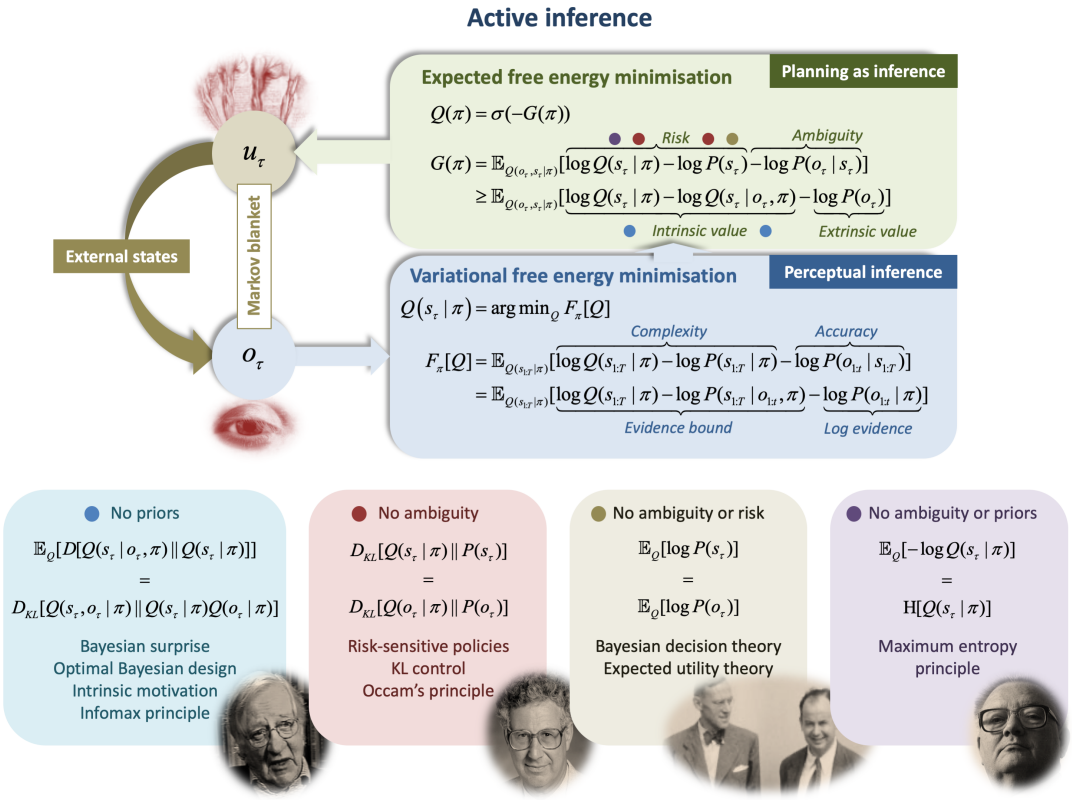

Figure 4.4: This figure illustrates the various ways in which minimizing EFE can be unpacked. The upper panel shows how Active Inference cast the minimization of VFE and EFE on perception and action, respectively. The lower panel shows that removing certain terms from the formulation of EFE, several special cases that predominate in the psychological, machine learning and economics literature emerge naturally. Adapted from [92]

4.4 Belief updating

At the core of the Active Inference framework, enabling agents to continuously adjust their internal models in response to incoming sensory information, lies belief updating. Belief updating mediates inference, where the agent refines its beliefs about hidden states and causes of observations, while simultaneously selecting actions that minimize surprise, and learning, which involves updating the model's parameters over time, allowing the agent to optimize its predictions based on past observations. This optimisation entails finding the sufficient statistics of posterior beliefs that minimise variational free energy [100].

4.4.1 Perception

In Active Inference perception is equated with estimating the states, from which the observation are generated. To infer the states of the environment, an agent must minimize the VFE with respect to the approximate posterior $Q(s|\pi)$ for each policy. Since the only quantity in free energy that depends on $Q(s|\pi)$ is F_π, the agent needs only to minimize it. This can be accomplished by performing a gradient descent on VFE. Gradient descent is an optimization technique that begins by selecting an initial value for s and then computing the VFE for that value. Subsequently, it evaluates the VFE for neighboring values of s, identifying the direction in which VFE decreases the most. The algorithm then updates s by sampling from the neighboring values that exhibit the largest decrease in VFE and repeats this process iteratively. This continues until a minimum VFE value is reached, meaning that VFE no longer decreases for any neighboring values. At this point, an approximation to the optimal beliefs about s has been obtained, given a set of observations o.

In Active Inference [92], this scheme is implemented by substituting $Q(s|\pi)$ with its sufficient statistics, so F_π becomes a function of those parameters. This enables us to write the gradient with state-estimation expressed as a softmax function of accumulated negative free energy gradients, denoted as $v_{\pi\tau}$. Hence, the VFE gradient can be defined as follows:

$$\begin{aligned} \dot{v}_{\pi\tau}(s_{\pi 1}, \ldots, s_{\pi T}) &= -\nabla_{s_{\pi\tau}} F_\pi(s_{\pi 1}, \ldots, s_{\pi T}) \\ s_{\pi\tau} &= \sigma(v_{\pi\tau}) \end{aligned} \tag{14}$$

However this solution is more robust and biologically plausible, another way to perform state estimation is to iterate until convergence the posterior distribution (here illustrated in matrix notation):

$$s_\tau^\pi = \sigma(\hat{A} \cdot o_\tau + \hat{B}_{\tau-1}^\pi s_{\tau-1}^\pi + \hat{B}_\tau^\pi \cdot s_{\tau+1}^\pi) \tag{15}$$

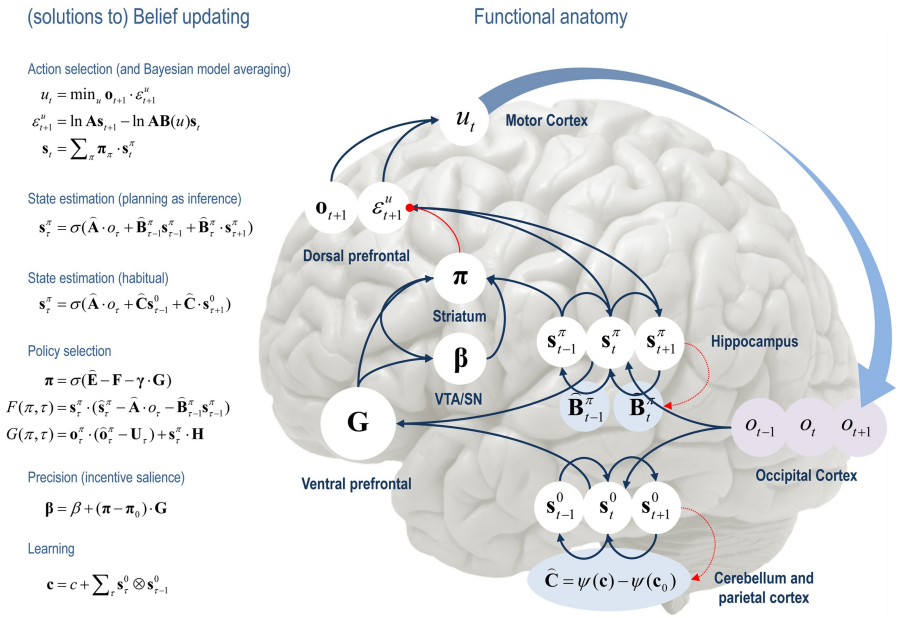

Figure 4.5: The left panel lists all the update solution. The image on the right illustrates a link between the sufficient statistics to various brain areas. The blue arrows denote message passing, while the solid red line indicates a modulatory weighting that implements Bayesian model averaging. The broken red lines indicate the updates for parameters or connectivity (in blue circles) that depend on expectations about hidden states. The large blue arrow completes the action perception cycle, rendering outcomes dependent upon action Adapted from [100]

where $\hat{A} = \mathbb{E}_Q[\ln A]$ and $\hat{B}^{\pi}_{\tau} = \ln \hat{B}^{\pi}_{\tau}$. Here, all future and past states are encoded explicitly. In other words, representations always refer to the same hidden state at the same time in relation to the start of the trial – not in relation to the current time [100].

4.4.2 Action

The other face of inference in Active Inference accounts for planning and decision making. To reach preferred states (or observations) an agent selects policies such that their predicted states, at some point in the future, reach the preferred ones.

The approximate posterior over policies is a softmax function of the negative free energy:

$$Q(\pi) = \sigma(-\ln E - F - \gamma G) \tag{16}$$

The function has three main components: a prior based on previous experience E, which can be used to model habits formation, the (posterior) free energy

F based on past outcomes and the expected (prior) free energy G based on preferences about future outcomes. The value of γ encodes the agent's confidence in policy selection and adjusts the contribution of G to the posterior distribution over policies. This term is updated according to the difference between the expected free energy under prior beliefs. In other words, when an observation is inconsistent with the model, the agent assigns lower precision to G entailing a stronger influence of the habits encoded in E on policy selection.

The most plausible action under all policies can be obtained through the approximation posterior beliefs about policies as a Bayesian model average:

$$u_t = \arg\max_{u \in U} \left(\sum_{\pi \in \Pi} Q(\pi) \right) \tag{17}$$

From a computational point of view, planning (i.e., computing G) for each possible policy can be very costly, due do the combinatorial explosion in the number of sequences of actions when looking deep into the future. Some researches suggest a simple answer, based on the brain functioning: humans stop evaluating a course of action as soon as they encounter a large loss. In Active Inference this can be replicated using an *Occam window*; that is, we stop evaluating when the expected free energy of a policy gets bigger than the best policy. Other researches [98, 101], try to tackle this problem computationally, by simulating competing course of action in the future, and selecting only the best one.

4.4.3 Learning

Finally we turn our attention to learning, which in Active Inference means updating prior beliefs over model parameters, such as the likelihood function or the transition beliefs, within a class of distributions called Dirichlet distributions (see Appendix A for more details). The updates for the parameters resemble classical Hebbian plasticity in the sense that, essentially, it just involves adding counts to a vector or matrix based on posterior beliefs, where larger numbers of counts indicate higher confidence.

Importantly, when learning is incorporated in the model, both the formulations of VFE and EFE change, depending on which parameter is being learned, because learning is also based on minimizing the free energy. For simplicity, assume that only the only variable being learned is A. This means that approximate posterior beliefs about A follow a gradient descent on variational free energy. The VFE, as a function of a (the sufficient statistic of $Q(A)$) is:

$$F(a) = D_{KL}[Q(A)||P(A)] - \sum_{\tau=1}^{t} \mathbb{E}_{Q(\pi)Q(s_\tau|\pi)Q(A)}[o_\tau \cdot \log(A) s_\tau] + \cdots$$
$$= D_{KL}[Q(A)||P(A)] - \sum_{\tau=1}^{t} o_\tau \cdot \log A s_\tau + \ldots \tag{18}$$

Here we only include the terms that depend on $Q(A)$ while ignoring the others. The KL-divergence between Dirichlet distributions is:

$$D_{KL}[Q(A)||P(A)] = \sum_{i=1}^{m} D_{KL}\big(Q(A_{\cdot i})||P(A_{\cdot i})\big)$$
$$= \sum_{i=1}^{m} \left(\log \Gamma(a_{0i}) - \sum_{k=1}^{n} \log \Gamma(a_{ki}) - \log \Gamma(a_{0i}) + \sum_{k=1}^{n} \log \Gamma(a_{ki}) \right) + (a - a) \cdot \log A \tag{19}$$

Incorporating Eq.19 in Eq.18, we can take the gradient of the VFE with respect to $\log A$:

$$\nabla_{\log A} F(a) = a - a - \sum_{\tau=1}^{t} o_\tau \otimes s_\tau \tag{20}$$

where $\otimes$ is the outer product. In computational terms, these are the dynamics for evidence accumulation of Dirichlet parameters at time t. Explicitly, setting the free energy gradient to zero at the end of the trial gives the following update for Dirichlet parameters:

$$a = a + \sum_{\tau=1}^{T} o_\tau \otimes s_\tau \tag{21}$$

Note that, in particular, the update formula counts the number of times a pair of states-observations have been observed. Following the same approach we obtain:

$$\begin{aligned} &\hat{A} = \psi(a) - \psi(a_0) \quad a = a + \sum_{\tau=1}^{T} o_\tau \otimes s_\tau \\ &\hat{B} = \psi(b) - \psi(b_0) \quad b(u) = b(u) + \sum_{\pi(\tau)=u} \pi_\pi \cdot s_\tau^\pi \otimes s_{\tau-1}^\pi \\ &\hat{D} = \psi(d) - \psi(d_0) \quad d = d + s_1 \end{aligned} \tag{22}$$

The vector encoding beliefs about initial states accumulate evidence by simply adding the nu,ber of times an initial state occurs.

In practice, the learning updates are performed at the end of each trial or sequence of observations. This ensures that learning benefits from inferred (post-dicted) states, after ambiguity has been resolved through epistemic behaviour [100].

Chapter 5

Deep affect inference

In the light of previous considerations, we now turn our attention to the practical realization of a (temporal) deep Active Inference model, constituted of two levels: a perceptive level whose aim is to perform inference on the hidden states of the world, grounded in the theory depicted in Chapter 4, and a metacognitive level, that integrates valence and arousal representation, as tools to improve the agent's overall behavior.

This way, not only we can incorporate in the Active Inference framework an affective component that helps tuning the agent's parameters, to better face the environment, but we even provide a tool that shows a plausible emotional representation of the agent mental state.

In what follows, firstly we define what this metacognitive entails and the purpose of it. Subsequently, we embed it in a suitable simulation environment, the T-maze environment.

5.1 Implicit metacognition

Starting from the generative model defined in Eq. 4.2, recall that the expected precision term γ informs the agent of the success of its own model, which can be interpreted as a minimal form of (implicit, non-reportable) metacognition. Furthermore, from the approximate posterior over policies (Eq. 16), note that higher values of γ account for a greater influence of the expected free energy, while, lower values entails that the agent falls back to its policy prior E_π. Formulated this way, we can think of γ as an internal estimate of model fitness, because it represents an estimate of confidence in a phenotype-congruent model of actions, given inferred hidden states [7]. Formally, when the model accounts for new evidence, that emphasizes the correctness of the model, thus expected model evidence is greater under posterior beliefs compared to prior beliefs (i.e., $(\pi - \bar{\pi}) \cdot G_\pi > 0$), γ

increases. In the opposite case ($(\pi - \bar{\pi}) \cdot G_\pi < 0$)), γ decreases. Put differently, γ changes when inferred policies differ from expected policies.

Moreover, the γ updates are termed affective charge (AC) [7]:

$$AC = -\Delta\bar{\beta} = (\pi - \bar{\pi}) \cdot G_\pi \tag{23}$$

AC can be different from zero when inferred policies differ from expected policies. It is positive when inferred evidence favors the agent's model, and negative otherwise. Finally, going forward, AC helps linking valence representation to the action model.

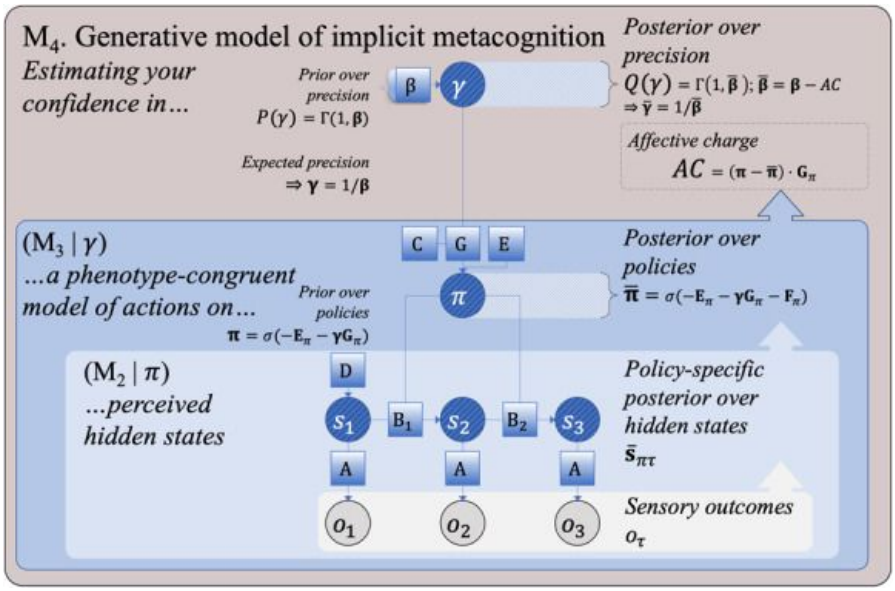

Figure 5.1: The figure highlights the recurrent formulation of the Active Inference generative model with respect to confidence formulation. Adapted from [7]

5.2 Temporal deep affect model

Here, we extend the classical Active Inference model to incorporate state representations of higher metacognitive states, such as valence and arousal. In doing this, we build a hierarchical model that comprises a lower level POMDP for context-specific Active Inference (perceptive inference) and a higher level MDP for affective inference.

As briefly presented in 4.2.2, in a deep temporal model, the posterior state representations at the lower level provide observations at the higher level, and in turn higher levels provide empirical priors to the lower level. Here we extend this

idea so that even lower level model parameter estimates are included in informing higher level posterior states. This allows the agent to form beliefs about its affective state regarding the outcomes gathered during the trial. In other terms:

> *the agent is equipped with a model that explicitly represents its emotional state, which, in turn, helps the agent to better tune its parameters to the subsequent trial.*

The lower level, as described in the preceding chapter, is defined by its main distribution: $A^{(1)}, B^{(1)}, D^{(1)}$, where the superscript (1) denotes the level for which they are responsible. Respectively, they account for the likelihood function, the state transition matrix, and the initial state beliefs. The perceptive level is context-specific in the sense that all its distributions need to be tailored according to the generative process it tries to model. Here, distributions are intended computationally, thus, the shape of the tensors embedding the categorical distributions.

The higher level is a Markov decision process whose aim is to infer from the lower level state representations and parameter estimates the affective states of the agent. Accordingly, this level includes hidden states to be inferred and the necessary quantities. Importantly, these states are general and context-free in the sense that they track the affective state of the agent, independently from the environment. The main difference from the lower level model-wise, is that the higher level does not include the notion of planning and decision making, thus, all the linked quantities, such as the EFE or the observation priors are not needed. Being context-free states, it makes sense to introduce beforehand. The second level has three hidden state factors: the context $s^{(C)}$, valence $s^{(V)}$ and arousal $s^{(A)}$. The first entails beliefs about the location of the reward [5] in the environment, the second state refers to the valence state of the agent, expressed as *positive* or *negative*, and it is intended to simulate the psychological construct of valence, and the latter state is the arousal, categorically defined as *high* or *low*, which mimics the arousal affective construct.

As such, the higher level includes accordingly a likelihood function $A^{(2)}$, a state transition matrix $B^{(2)}$ and the initial state belief $D^{(2)}$.

The main idea is that the hidden states at the higher level provide empirical priors over any variable at the lower level that does not change over the timescale associated with that level [7]. Here, we consider valence and arousal to influence the rate parameter (β) of the priors over expected precision and the context belief to shape the initial beliefs of the agent.

[5] Even though the higher level is stated to be context-free and this factor resemble a context-dependent hidden factor, in our model can be turned off easily, without impacting the overall functionality of the agent.

5.2.1 Hidden state affective factors

There are various researches that try to tie together and pack the two psychological constructs of valence and arousal into the framework of Active Inference. The literature mainly consider the valence aspect, while the arousal is usually left aside, as it is understood to be derived from the former. Within various paradigms, valence has been recognized to be linked subjective fitness: [102] proposed an interpretation of emotional valence in terms of rates of change in variational free energy, in [103] valence was understood as the match and mismatch processes in neural networks, monitoring the fit between a neural architecture and its input. However, little researches include formal connection to action. In this work, valence can is intended as the agent's confidence in its internal model, specifically in terms of how well this model predicts outcomes. The concept of affective charge, as defined in 5.1, helps bridge changes in free energy with an explicit model of action selection. Here, valence operates as a metric of how well the agent's action model G_π corresponds to perceptual evidence derived from actual outcomes F_π.

This dynamic relationship reflects a key principle of Active Inference: the agent seeks to minimize free energy by reducing discrepancies between its predictions and sensory input. Affective charge measures changes in expected precision (or confidence) about how much the action model matches reality. In this context, valence represents the alignment or mismatch between these expectations and experiences.

In this framework, valence ties directly to the expected precision of an agent's beliefs about its own behavior. Positively valenced states indicate high confidence in one's internal model, reinforcing reliance on prior expectations about outcomes. Conversely, negatively valenced states signal lower confidence in the model, prompting a greater reliance on incoming sensory evidence and a shift away from prior expectations.

From the perspective of Active Inference, when an agent selects actions, it weighs the expected free energy of different policies (potential actions). If the environment is stable and predictable, the agent's model is likely accurate, and high confidence (or precision) is granted to the expected free energy, which guides action selection based on risk-minimizing strategies. In unpredictable environments, however, expected precision is reduced, causing the agent to rely more on sensory evidence than on prior beliefs.

Valence, then, acts as a computational mechanism for adjusting this balance between prior expectations and sensory evidence. High valence states signal confidence in the agent's internal model and lead to risk-averse behavior, while low valence states correspond to uncertainty and a more exploratory stance toward incoming information.

On the other hand, arousal can be understood as the agent's response to the

uncertainty or complexity of incoming sensory stimuli, which reflects the intensity of the sensory input and its potential to disrupt or enhance the agent's internal model [10, 74, 75]. Drawing from Berlyne's idea of arousal potential [73], we align novelty and complexity as collative properties that impact the level of arousal. These collative properties challenge the internal model by introducing uncertainty, prompting the agent to update its beliefs. The greater the mismatch (prediction error), the more the agent's model is pushed to adapt, leading to higher arousal. Arousal can be interpreted as a reflection of this prediction error. When an agent encounters stimuli with high novelty or complexity, it experiences greater uncertainty, which increases free energy. The larger the discrepancy between the model and the incoming data, the higher the arousal potential. The level of arousal, in turn, influences how an agent selects actions. When the environment is predictable and free energy is minimized, arousal is neither low nor high, but sits in a sweet spot, and the agent can rely more confidently on its prior expectations. However, when free energy (and thus arousal) is high (or extremely low), the agent is prompted to explore and update its model to reduce the prediction error. This dynamic reflects how arousal functions as a mechanism that signals the need for learning or adjustment in response to novel or complex stimuli.

In what follows, we tie together the idea of arousal and valence as factors influencing action selection. We consider the belief updating in terms of messages that descend from the affective level to the lower level and ascend from the lower level to the affective level. Descending messages provide empirical priors that optimize policy selection. Ascending messages can be interpreted as mediating belief updates about the current context and affective states.

5.2.2 Descending messages

Descending messages refer to the flow of information from the higher level of the generative model down to the lower level. These messages convey top-down empirical prior beliefs to condition upon parameter estimates at the lower level. The descending signals are primarily influenced by prior beliefs encoded at higher levels of the generative model, such as contextual states (e.g., whether a reward is expected on the left or right) and affective states (e.g., positive or negative valence/high or low arousal).

The high level provides empirical priors about the context state in the shape of categorical distribution, in other words, when the lower level receives this signal, update it prior distribution $D^{(1)}$ accordingly:

$$P(s_1^{(1)}|s_T^{(C)}) = Cat(A^{(C)}) \tag{24}$$

Conversely, the affective states provide prior employed in the estimate of the β parameter. Importantly, while the rate parameter is continuous, the valence and

arousal states are not, so it is necessary to associate these states with two values of the parameter, that for simplicity we assume are the same. Effectively, these two values are the upper and lower bounds on the expected precision under the levels of the affective states.

$$P(\gamma) = \mathbb{E}_{Q(s_T^{(2)})}\left[P(\gamma|s_T^{(A)}, s_T^{(V)})\right] = \Gamma(1, \beta)$$
$$\beta = \left(\beta^{(+,-)} \cdot A^A s_T^{(A)} + \beta^{(+,-)} \cdot A^{(V)} s_T^{(V)}\right)/2 \tag{25}$$

Valence and arousal are combined together in the same fashion, because their effect on decision-making and action selection is similar, as discussed in the preceding section. Inspired by [7], we decided to add the arousal conditioning this way for two main reasons: to maintain a similarity between valence and arousal, since in many works they are intrinsically coupled, and because their influence is similar, but opposite. High values of valence entail high confidence in one's model, in accordance with non-extreme values of arousal. Conversely, low values of valence and extreme values of arousal show low confidence in the model, thus, the agent takes on an exploratory behavior.

5.2.3 Ascending messages

At the end of each trial, after the agent has navigated the world and gathered evidence, exogenous (context state) and endogenous (affective charge and information gain) signals induce belief updating at the higher level of hidden states. This process is designed to ensure that belief updates at this higher level (contextual and affective states) evolve more slowly than belief updates at the lower level.

These state updates define how the high states evolve:

$$\bar{s}_T^{(C)} = \sigma\left(\ln B^{(C)} \bar{s}_{T-1}^{(C)} + \ln A^{(C)} \cdot \bar{s}_1^{(1)}\right) \tag{26}$$

This second-level expectation about context comprises empirical priors from the previous trial and evidence based on the posterior expectation of the initial (context) state at the lower level.

The valence state is based on previous affective expectations and evidence for changes in affective state based on AC:

$$\bar{s}_T^{(V)} = \sigma\left(\ln B^{(V)} \bar{s}_{T-1}^{(V)} - \ln \frac{\beta^{(+,-)} - AC}{\beta^{(+,-)}} \frac{\beta}{\beta - AC}\right) \tag{27}$$

This contains empirical priors based on previous affective expectations and evidence for changes in affective state based on affective charge, evaluated at the end

of each trial time step Note that when the affective charge is zero, the affective expectations on the current trial are determined completely by the expectations at the previous trial (as the logarithm of one is zero).

On the other hand, the calculation of arousal is based on the divergence between the prior beliefs at the beginning of a trial and the posterior beliefs at then, which is termed *Information gain* (InG) [10]. This term, in turn, represents the Bayesian surprise of the model, that we use in a sigmoid function to obtain a categorical distribution over two states: *high arousal* and *low arousal*.

$$\begin{aligned} InG &= D_{KL}\left[P(s_T^{(1)})||P(s_1^{(1)})\right] \\ Ar &= \frac{1}{1+e^{-InG}} \\ \bar{s}_T^{(A)} &= (1-Ar, Ar) \end{aligned} \tag{28}$$

5.3 Simulating emotions

The aim of this thesis is not only to frame arousal in the Active Inference approach, but to even show that valence and arousal, computed this way, form a simplified emotional space, similar to Russell's circumplex, that simulate a plausible emotional model. Firstly it was necessary to map both valence and arousal in a new range of values. Formally, the range $[-1, 1]$ is commonly employed in this kind of task, therefore we developed a way to map, not only the values in this new range, but even the meaning of the values themselves. The last part refers to the fact that the mapping starts from probability values (i.e., these values, in simple terms, answer the question "What is the probability that the agent has high arousal\valence?") to a new range of values, without losing the semantic interpretation underneath.

Regarding the valence term, the adopted mapping is a linear mapping:

$$Valence = 2s_T^{(V)}(positive) - 1 \tag{29}$$

This linear mapping suffices our needs, because the valence interpretation is linear with respect to its values: the lower the probability of having a positive valence, the lower the valence value itself.

On the other hand, the arousal mapping is more complicated, because its representation is not linear: the arousal value assumes negative interpretation for extreme values, while for mild values of the probability of having a high arousal level, the arousal value reaches the maximum value. To account for these implications, we employed as a mapping function the sum of two sigmoids [6]: one

[6]Even tough this formulation is usually employed in portraying valence as a function arousal, the shape and the formulation of this function are what we want to achieve, as for the mapping.

refers to the aversion system and the other to the reward system, as defined by Berlyne [104]. The first system generates a positive response when the arousal potential increases (i.e., the InG increases), while the second system, counters the first, providing the negative response, as the arousal potential increases too much. Note that, the aversion curve has a higher absolute maximum value, than the reward does. Thus, the joint sigmoids create the *inverse-U* Wundt curve.

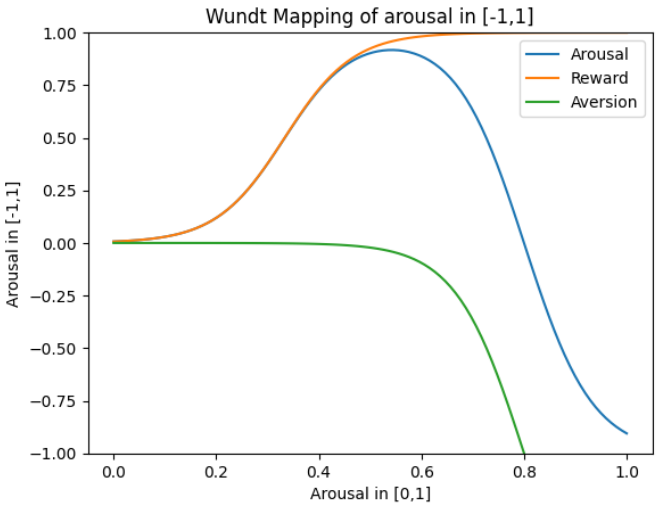

Figure 5.2: The figure illustrates the employed arousal mapping as the sum of two sigmoids functions representing reward and aversion.

This way, we account for negative values of arousal, when the arousal level is too high, thus detrimental for the model, and highlight how milder levels are optimal for the agent. Once the valence and arousal values, have been computed, they are coupled as to represents a point in the emotional circumplex (Figure 5.3).

5.4 A case study

To frame the model we described so far, we resort to a well known simulated environment: the T-maze environment, which is depicted in 5.4.

The environment is provided by the pyMDP [105] suite, which supplies even the Active Inference framework to solve it and on which we built our own model. A more comprehensive summary of this package is described in the Appendix B.

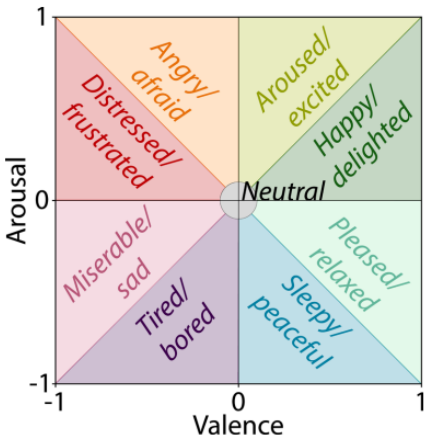

Figure 5.3: The figure illustrates the simplified circumplex of emotion used in our model, where valence and arousal account for the dimensions of the space.

5.4.1 Generative process: T-maze environment

The environment involves a T-maze experiment, where an agent (a mouse) navigates to locate a reward while avoiding punishment. The mouse starts at the central position of the maze and can choose to stay or move in one of three directions: left, right, or down. In each trial, the reward (food) is hidden in either the left or right arm of the maze, while the opposite arm contains a punishment (shock). The maze presents a challenge because the reward location changes at some point in the simulation, creating uncertainty for the agent.

Additionally, once the agent enters either the left or right arm, it cannot leave until the trial ends, making these absorbing states. The downward path, however, offers an informative cue, which reliably indicates the current location of the reward. By accessing this cue, the agent can gather valuable information, helping it to make better decisions.

The agent already has prior knowledge about the maze and the behavior of its elements. It knows that the left and right arms are one-way streets and that the cue provides critical information about the trial's context. Each trial consists of two decision points, meaning the agent must make two consecutive moves. Its first action may involve either moving toward the reward or gathering information from the cue, while the second move depends on what happens after the initial decision.

The mouse's decisions involve more than just maximizing reward. It must also consider the value of gaining information that might help in future actions. This setup creates a natural tension between exploratory behavior, where the mouse

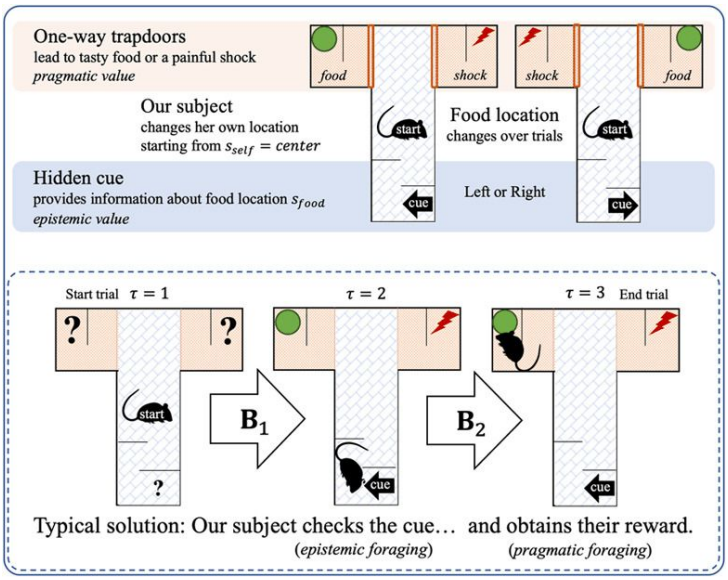

Figure 5.4: The figure illustrates the T-maze environment. Figure from [7]

seeks new information to reduce uncertainty, and exploitative behavior, where it acts to achieve immediate reward. The environment thus encourages the agent to balance the benefits of exploration and exploitation when making decisions, especially in situations of uncertainty.

5.4.2 Generative model: affective agent

The subject of the simulations is a deep temporal Active Inference agent, developed as described in the preceding chapters. For the sake of simplicity, the agent is equipped with (previously gathered) prior knowledge about the workings of the T-maze in its generative model. In other words, the agents is not equipped to learn any world dynamics.

The agent is equipped with a generative model whose dynamics are controlled by the following beliefs in form of matrices. For the lower level we define:

- A vector $D^{(1)}$ of prior beliefs about the starting position of the agent in the maze and beliefs about the context, thus, the reward location.
- A likelihood function $A^{(1)}$ per each observation modalities and hidden states.
- A state transition matrix $B_{\pi}^{(1)}$, which specifies how hidden states evolve over time accordingly to the chosen action.
- A vector C of preferred observation, in this case it is highlighted that the agent wants the reward.

- A baseline expectation on policies E_π, which represents the probability that a given policy will be selected a priori.

Furthermore, the context-specific hidden states factor the agents tries to infer from incoming observations are:

- The agent's position in the T-maze, namely the *location*, which consists of four hidden states: {center, cue, left, right}, where cue refers to the bottom position.
- The agent's belief about the location where to find the food and it consists of two hidden states: {left, right}.

After each action, the agent collects the following observations:

- Its own location in the maze: {center, cue, left, right}.
- Whether the agent found the food or took the shock: {Null, Reward, Loss}. The Null observation is observed when the agent is positioned outside the two arms.
- The information the agent can obtain from the maze, only from the bottom location: {No Cue, Cue Left, Cue Right}. Similarly, No cue is the default observation outside the cue location.

As for the affective level, we briefly reformulate what already has been extensively presented. There are three hidden state factors, each with two hidden states:

- $s^{(A)}$ = (high, low).
- $s^{(V)}$ = (positive, negative).
- $s^{(C)}$ = (left, right).

Then, we define the observations for the same level:

- β = (0.5, 2.0), the beta rate as describe before.
- The beliefs about the reward location: {

The following matrices describe the dynamics of the elements in the higher level of the model:

- A categorical vector $D^{(2)}$ of beliefs about the reward location.
- A likelihood function $A^{(V),(A)}$ which reflects some uncertainty degree in the affective predictions.

- The likelihood mapping $A^{(C)}$, which maps from the context states to the lower level.
- The state transition probabilities $B^{(2)}$, which specify how each hidden state evolve from one trial to the next.

5.4.3 The Action-perception-metacognition cycle

Once the agent has been defined with its generative model, it follows the step deligned in 1.

Algorithm 1 Action-perception-metacognition cycle

Input: agent($A^{(1)}$,$B^{(1)}$,$C^{(1)}$,$D^{(1)}$,$A^{(2)}$, $B^{(2)}$, $D^{(2)}$)
env ⟵ environment

for $T = 1$ in $Trials$ **do** ▷ Initialization
descending_messages()
observations ⟵ env.reset()
for $t = 1$ in *timesteps* **do** ▷ Action-perception cycle
states ⟵ agent.infer_states(observations)
π ⟵ agent.infer_policies(states)
action ⟵ agent.sample_action(π)
observation ⟵ env.step(observation)
end for ▷ Metacongitive step
compute_AC()
ascending_messages()
agent.update()
end for

The cycle adopts two different time scale: trials and time steps. The former refers to the number of simulations, and is the scale at which the affective level works. The latter refers to the number of actions the agent can take in the environment.

At the beginning of each trial, the agent initializes its generative model, in particular, the *descending_messages* provide the agent with prior beliefs about *reward position* and *rate parameter*. After updating its beliefs accordingly, the agent starts the action-perception cycle, in which, it infers first, the hidden states of the environment, from the observed quantities, and then from them, the policies and lastly the best action to perform. At last, it observes again the environment, and the cycle restarts.

After that, the agent can perform the metacognitive step: compute the *affective charge* using the accumulated evidence in the trial and then infers the affective states and beliefs about the context.

Chapter 6

Results

The generative model described in the preceding chapters was employed for both simulating the affective inference framework and showig that our model is apt to provide a plausible model of affect.

In particular, the task involves a synthetic agent that experiences 64 T-maze trials, with the food location switching after the first 32 trials. In each trial, the agent can either receive a reward or a painful shock, depending on whether it chooses the left or right arm. Initially, the agent experiences uncertainty about outcomes, simulating an anxious affective state. Over time, as the agent gathers contextual evidence and improves its beliefs about the reward location, it is expected to grow more confident and transition to a positively affective state.

In the second phase of trials, a reversal in reward location (from left to right) introduces a challenge, requiring the agent to adjust its beliefs again. The agent is expected to eventually regain confidence after sufficient experience. The task aims to illustrate how confidence influences affective states, particularly in transitioning from anxiety to a more positive state.

After each trial, the valence and arousal levels are combined together to form a point of the (simplified) emotional space to even show a representational validity of our model, in the sense that the agent emotional trend can be semantically represented, and the inner quantities of the statistical model are ontologically justified. Accordingly, the agent should initially show a calm emotional state, that is disrupted by the context switch after 32 trials, which causes the agent to lose confidence in its own model and therefore feel a negative affective state. This state of frustration is easily overcome thanks to the adaptability of the model, which regains confidence in itself once the new uncertainty is dissipated.

In the following, we describe the outcomes for the setup we have described. The quantities we analyze are what we would consider the most meaningful in this task: the AC term, the expected precision and the context starting beliefs.

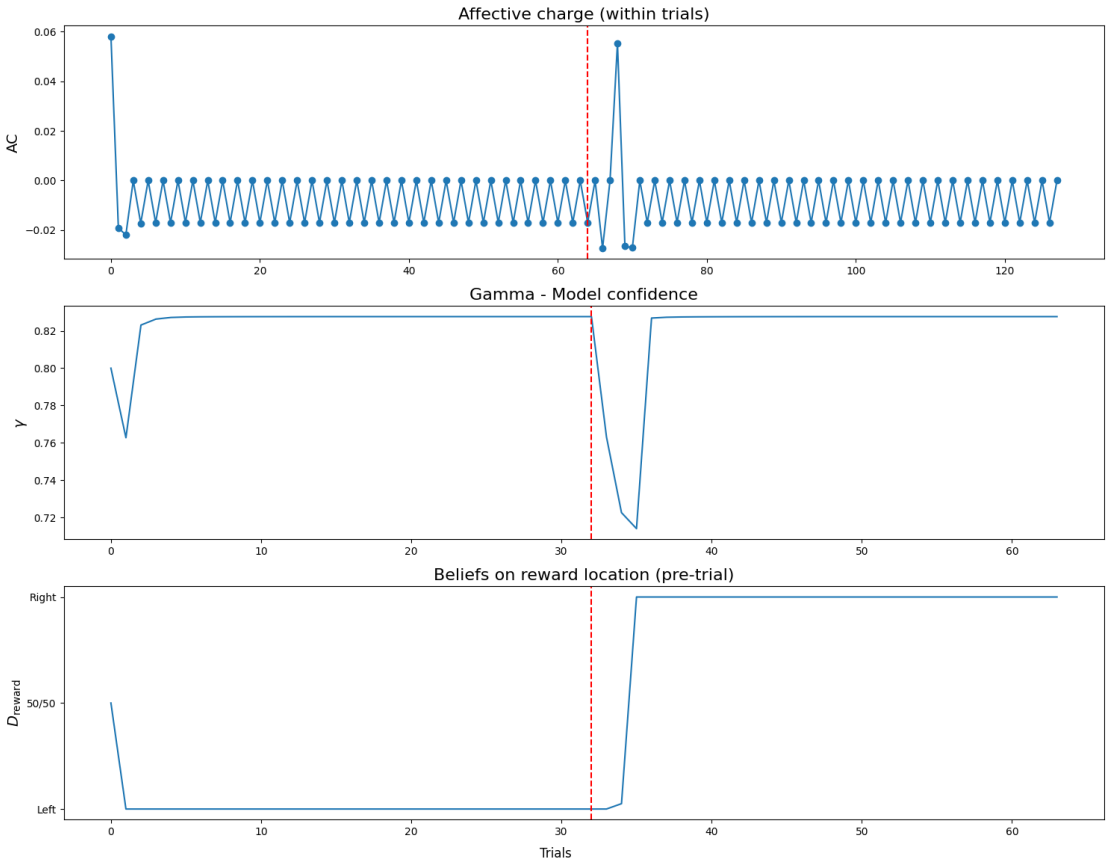

Figure 6.1: A summary of the main quantities updates in the affective agent model during 64 trials.

Figure 6.1 shows the simulation results for an agent that includes a metacognitive layer in his generative model. The dynamics of this simulation can be divided into four parts: two periods within each of the 32 trials before and after the context reversal. These periods show an initial phase of foraging, as the agent search for cues about the context (i.e., the agent immediately asks for the cue), followed by a phase of confidence exploitation. As defined by its own priors the agent starts the simulation in a negative anxious state (as for anyone that takes on a new challenge!). The subject confidence immediately gets boosted as it resolve the environment uncertainty by immediately asking for the context cue. This allows the agent to quickly adapt, and start assuming an epistemic behavior until the context reversal. On the 32nd trial, the food location changes and in response the agent's confidence drops (as the expected precision plot shows) for a few trials, in which it swiftly regain courage (i.e., confidence) to adapt to the sudden change.

To illustrate the importance of the metacognitive higher level, we repeated the

simulations in the absence of a hierarchical model, thus employing a classic Active Inference agent. After removing the higher level, the resulting (less sophisticated) agent, which could be thought of as an agent with a "lesion" to higher levels of neural processing, updated expectations about food location by simply accumulating evidence in terms of the number of times a particular outcome was encountered, as described in 4.4.3.

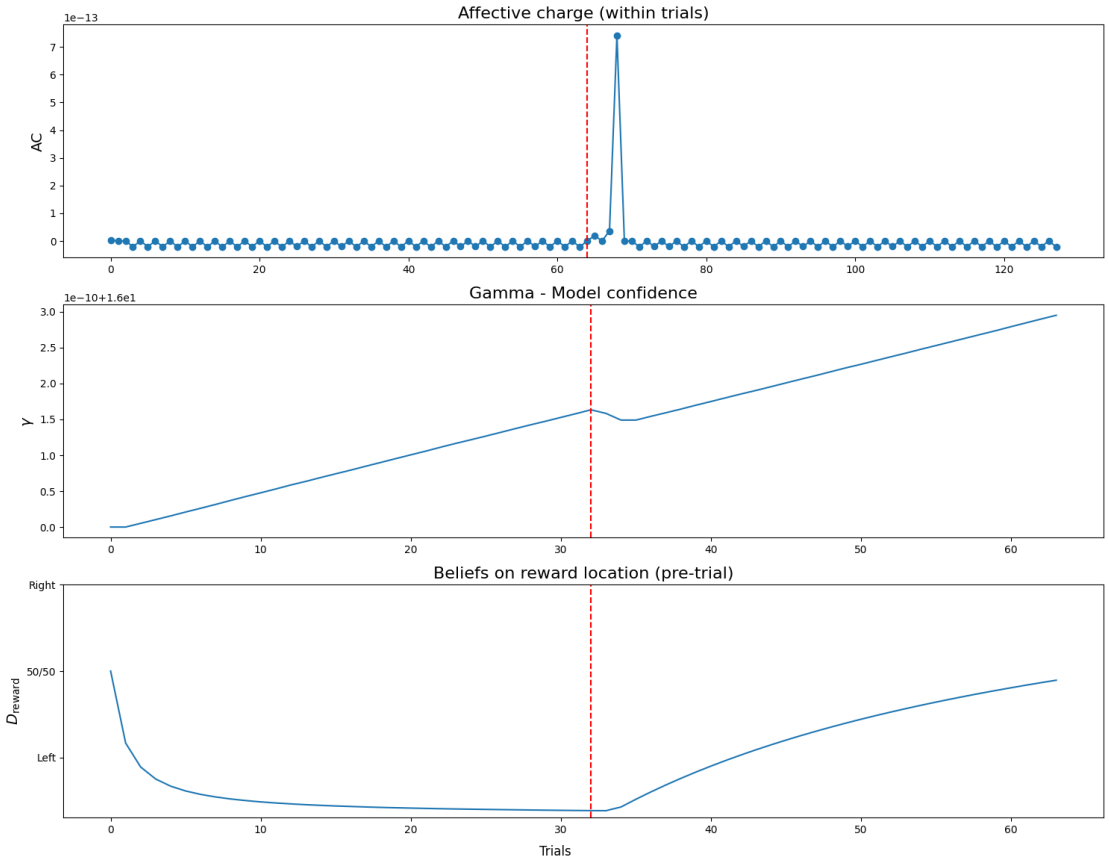

Figure 6.2: A summary of the same quantities updates in a classic agent during 64 trials. The comparison between the performance of the two models is evident: a classic approach fails when put against a sudden change, that disrupt the agent. The effects are shown here as the agent struggles to get to same level of confidence as before (bot panel) and it does not actually understand what happened, since its own confidence is not affected by the change (mid panel).

Figure 6.2 provides a summary about the results of this agent. Even in this case the same four parts can be distinguished, however, it is clear that are not as sharp before, rather, they are less distinguishable, thus showing that this kind of agent is not able to adapt as ours. In the first 16 simulations, the agent becomes

very confident very quickly about the context it is in, and around the 10^{th} it is certain about the food location. After the reversal, where the agent has to adapt again to the environment, this process becomes slow , so that at the end of the sixteen more trials, it is still unsure about the location of the food. As for the expected precision, which, we recall, represents the agent confidence in its own model, it shows an increasing trend. Note that around the very first trials and the ones immediately after the switch, this trend slows down, as the agent adopts a more epistemic behavior in search for the right cue. At last, the first plot shows the affective charge trend, that is, has the scale shows, meaningless, since it essentially is (computationally) zero.

Overall the classic agent performs very well with respect to the task, but its own internal model is very slow to adapt, and shows very little resilience against content switch. Moreover, the values gamma assumes are intuitively against what one would think, as they practically increase (the agent keeps getting more confident) as the trials advance.

Figure 6.3 presents the second part of this thesis: the mapping from statistical meaningful quantities into affective states. At the end of each trial, the level of the arousal and the valence was collected and mapped through the function described in 5.3 into commonly known emotions. This part allows us to appreciate even more the described affective model, because it shows from a new and more accessible point of view, the same results we obtained. Remember that the agent starts in a negative state, which is promptly changed into positive, as soon as the agent solves its uncertainty about the context, and starts observing the rewards coming. After that, the agent starts a period of excitement where it keeps getting the reward and nothing arousing happens, until a sudden change in the context catches him by surprise. Surprise, that is solved as before. This long periods of serenity match the parts of high confidence, where the agent does not need to update its model, and can keep exploit it, without any energy expense.

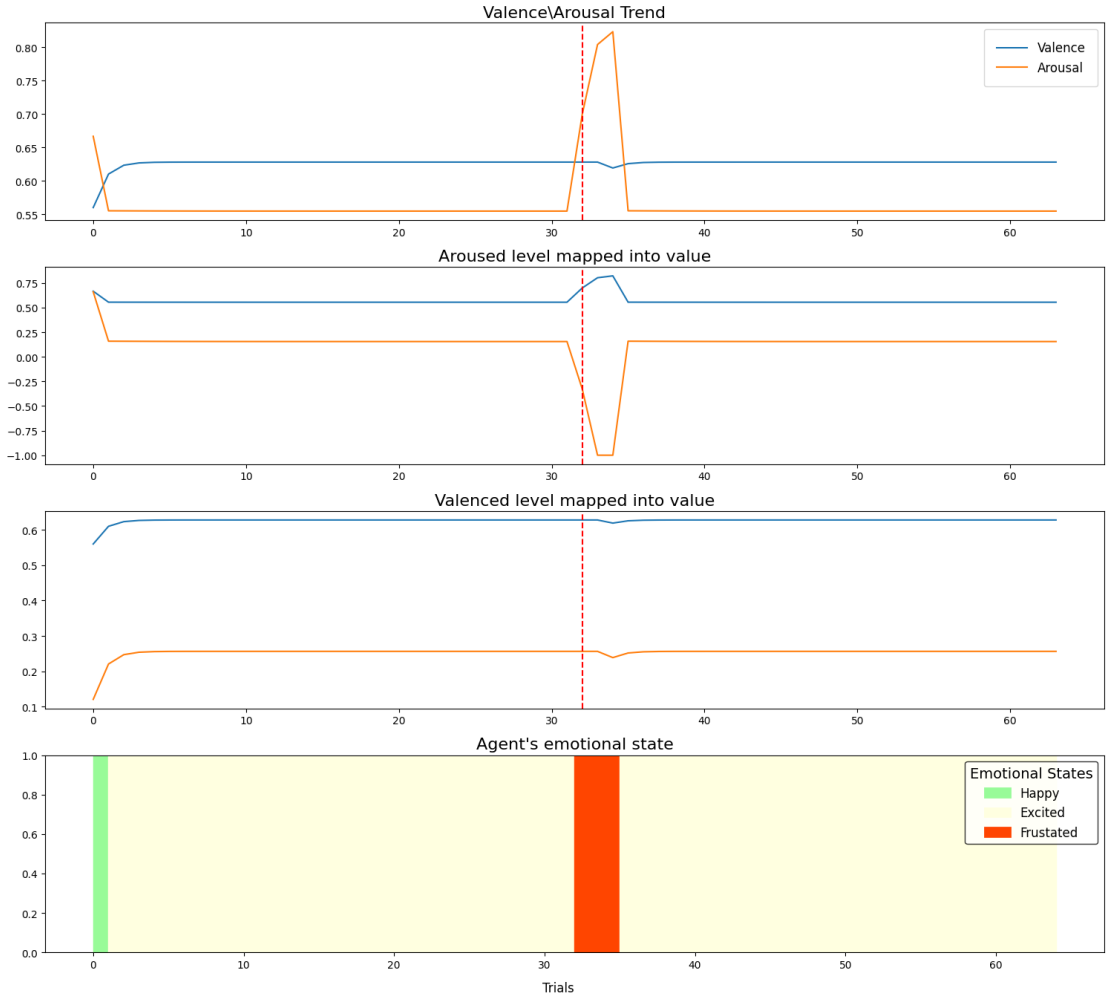

Figure 6.3: Emotion representation of the agent affective state during the 64 trials. The top panel shows the trend of both valence and arousal. The second and third plots show the mapping of the valence and arousal into the right range of values to be employed as as a point in the emotion space. The last panel shows the affective trend during the whole simulation, highlighting the three main parts: the beginning of the simulation, in which the subject starts happy, the long periods of quiet, in which the agent keeps getting the reward, thus it has complete confidence in its own model, and lastly, the context switch that altered the subject.

Chapter 7

Conclusions

This thesis has advanced the field of affective computing by developing a novel computational framework that integrates the Active Inference approach with the constructs of valence and arousal. By bridging agency and emotions, the work presented here provides a more comprehensive and biologically plausible model of emotional processes. The model not only captures how individuals infer states of the world through perception, action, and learning but also demonstrates how affective states influence, and are influenced by, these inferential processes.

The main contributions of this thesis can be summarised as follows:

- Theoretically we extended the Active Inference approach to accommodate a temporal deep higher level that would introduce both valence and arousal into the framework.
- Computationally we implemented a synthetic agent, starting from the pyMDP suite, to evaluate our hypothesis.

The primary contributions of this work lie in the formalization of affective states within the Active Inference framework. By arming the generative model with parametric and temporal depth via an higher level of inference, we have shown how emotions naturally arises as part of the agent's ongoing effort to minimize free energy. By embedding valence (a measure of the desirability of outcomes) and arousal (a driver of behavior) into the generative model, we have proven that the affective dimension can be used not only as a tool for communicating with others, but even as an estimate of our own action model, which help us engage better in unexpected situations. The simulations carried out with a synthetic rat navigating a T-maze illustrate how the agent's confidence in its actions influences its affective states, and vice-versa, providing insight into the relationship between decision-making and emotions.

In particular, to the best of our knowledge, this work presents a first attempt to include the arousal construct into the Active Inference framework as an informative quantity; meanwhile, arousal is formally intertwined with valence.

Even though there exist some proposals on this topic in the Active Inference field, they mostly focus on modeling arousal around the observations: the sensory input the agent observes are themselves arousing or not.

In a different vein, we modelled arousal as an affective state that the agent has to infer from the its own model. In other words, arousal emerges from within, in accordance with the predictive brain theory.

Moreover, we implemented a simplified but very plausible emotional mapping, starting from the affective levels measured in our subject during the simulations. The results so far achieved show that our approach is promising. The the emotional responses of the agent to the environment are in accordance with the underlying theories: the agent feels a strong negative discomfort when its arousal level is too high, which means, that the information gathered are conflicting with agent's beliefs. In other words, the agent has to update its internal model to make up for the new evidence, which, in turn, implies that the agent needs to actively make an effort to change accordingly. On the other hand, when the situation is very predictable, the agent sits calmly on its prior beliefs, and can profit without indulging in any major update.

Clearly the presented work is not innocent of limitations. First of all the model current bounded capability of dealing with real world behavioural data due to its burdening computational complexity. This is a limitation that generally concerns *per se* the computational architecture of Active Inference. In this perspective, some recent proposals are investigating the possibility of approximated computation that take advantage of the work done in generative deep neural networks.

To conclude, this thesis posits valence and arousal as linked emergent states of a deep temporal and parametric Active Inference agent. These states, in turn, drive the agent behavior, favoring exploitative behaviors, when the agent feels comfortable in its own model, while shifting to more defensive choices (i.e., exploratory) once the agent is put against evidence that collides with its past experiences. Thus, the agent loses confidence on the past and focus on the present.

Appendix A

Dirichlet distribution and learning

The Dirichlet distribution is used as a prior over the parameters of a categorical distribution because it is is the conjugate prior for the categorical distribution. This means that when computing the posterior distribution, thus multiplying the categorical distribution with its Dirichlet prior, we end up with another Dirichlet distribution.

Let's see what this entails mathematically in the context of learning a parameter in active inference.

The Dirichlet distribution is defined as follows:

$$P(\theta|\alpha) = Dir(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{30}$$

where $\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)}$ is a normalization constant that assures that the distribution sums to 1. Γ is the gamma function, the variable $\theta = (\theta_1, \ldots, \theta_K)$ is a vector of length K containing the factors of a categorical distributions, and $\alpha = (\alpha_1, \ldots, \alpha_K)$ is the set of concentrations parameters of the Dirichlet distribution. Importantly, each parameter satisfies $\alpha > 0$.

Similarly, the categorical distribution is defined as follows:

$$P(x|\theta) = Cat(x|\theta) = \frac{1}{x_1! x_2! \ldots x_K!} \prod_{k=1}^{K} \theta_k^{x_k} \tag{31}$$

where $x = (x_1, \ldots, x_K)$ is a categorical variable in the form of a one-hot vector (e.g., $x = [00100]$), $\theta = (\theta_1, \ldots, \theta_K)$ are the parameters of the distributions, such that $\theta_k > 0$. The first term is a normalization constant.

If we multiply the Dirichlet and the categorical to arrive at the posterior over θ of the categorical distribution we get (ignoring the normalization constant):

$$P(\theta|x,\alpha) = Dir(\theta|x,\alpha+x) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k+x_k-1} \tag{32}$$

which has the same form of the prior, except for x_k. This term can be seen as a "count" of 1 to the concentration parameters corresponding to the observed variable. It is the concentration parameters of the Dirichlet distributions in the POMDP structure that are updated during learning [94].

Appendix B

pyMDP

The open-source pyMDP suite provides tested and modular routines for simulating active inference agents with POMDP generative models.

pyMDP allows to:

- easily build a generative model using prior and likelihood distributions
- initialize an agent
- connect it to an external environment for running active inference processes

The library follows the standardized framework of OpenAIGym, commonly used in reinforcement learning, where the Agent and Environment classes exchange observations and actions over time. In order to enhance the user-friendliness of pyMDP without compromising flexibility, it is highly modular and customisable, allowing agents to be specified at different levels of abstraction with customised parameterisations.

B.1 The Agent Class

pyMDP offers the Agent class, an high-level API. Instantiating an Agent, the user can abstract away the various optimization routines and sub-operations that make up an active inference process, such as state estimation, action selection, and learning. These sub-routines are abstracted as methods of Agent.

B.1.1 Building the matrices

pyMDP package implements generative models of discrete states and observations that evolve in discrete time. These models use categorical distributions[7], which can be represented as multidimensional arrays NDarrays, to store their parameters. There are two types of categorical distributions: vector-valued marginal distributions (such as $P(x)$), which typically serve as priors, and conditional categorical distributions (such as $P(y|x)$), which are represented as NDarrays and act as likelihoods in the generative model. In pyMDP, marginal categorical distributions are represented as 1-D vectors that are instances of `numpy.ndarray`, while conditional categorical distributions are encoded as 2-D NDarrays and higher-order NDarrays.

The core distributions are the observation likelihood and the transition likelihood, represented by the A array and B array respectively. The initial prior over states is represented by the D vector and a goal distribution or prior preferences can be represented by the C array.

To build an active inference agent, the Agent() constructor is used, with mandatory inputs of the A and B arrays and optional inputs of the C and D vectors (setup as uniform distributions by default). Once the agent is created, its methods can be used to perform active inference.

B.2 The Environment Class

Usually, the agent need to interface with an environment or external world. To define an environment, it is sufficient a class or a function:

1. taking the agent's actions as input
2. updating the true hidden state of the environment
3. returning the observations generated by the updated hidden state

In the Bayesian modelling literature, the environment is also called generative process, which does not have to be identical to the generative model. It is important that the environment accepts the agent's actions and returns observations that are discrete and are compatible with the support of the likelihood $P(o_\tau|s_\tau)$ of the agent's generative model.

The library pyMDP includes pre-made environments, otherwise users can create their own custom environment class. This class typically includes a step() method, which takes an action from the agent as input and returns the resulting observations that the agent will process during the next time step.

[7] A categorical distribution, referred to as $Cat(\phi)$, assigns a probability value between 0 and 1 to each outcome level of the distribution's sample space.

B.3 The action-perception loop

The typical active inference loop consists of three main steps:

1. sample an observation from the environment
   In pyMDP, call the function env.step()

2. update the agent's beliefs about states and policies using the observation
   In pyMDP, call the functions agent.infer_states() and agent.infer_policies()

3. choose an action, based on the agent's posterior over policies
   In pyMDP, call the function agent.sample_action()

Wrapping these three steps into a loop over time entails the entire active inference process.

Bibliography

[1] Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press, 03 2022.

[2] James A. Russell. Emotion, core affect, and psychological construction. *Cognition & Emotion*, 23(7):1259–1283, 2009.

[3] Giovanni Pezzulo, Thomas Parr, and Karl J. Friston. Active inference as a theory of sentient behavior, 2024.

[4] K.J. Friston T. FitzGerald F. Rigoli P. Schwartenbeck G. and Pezzulo. Active inference: A process theory. *Neural Computation*, vol. 29(no. 1), 2017.

[5] Karl Friston. Am i self-conscious? (or does self-organization entail self-consciousness?). *Frontiers in Psychology*, 9, 2018.

[6] David Rudrauf, Daniel Bennequin, Isabela Granic, Gregory Landini, Karl Friston, and Kenneth Williford. A mathematical model of embodied consciousness. *Journal of Theoretical Biology*, 428:106–131, 2017.

[7] Casper Hesp, Ryan Smith, Thomas Parr, Micah Allen, Karl J Friston, and Maxwell JD Ramstead. Deeply felt affect: The emergence of valence in deep active inference. *Neural computation*, 33(2):398–446, 2021.

[8] Hideyoshi Yanagisawa and Shimon Honda. Modeling arousal potential of epistemic emotions using bayesian information gain: Inquiry cycle driven by free energy fluctuations. 2023.

[9] Hideyoshi Yanagisawa, Oto Kawamata, and Kazutaka Ueda. Modeling emotions associated with novelty at variable uncertainty levels: A bayesian approach. *Frontiers in Computational Neuroscience*, 13, 2019.

[10] Hideyoshi Yanagisawa. Free-energy model of emotion potential: Modeling arousal potential as information content induced by complexity and novelty. *Frontiers in Computational Neuroscience*, 15, 2021.

[11] Rafael A. Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[12] Rosalind W Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[13] Jonathan Gratch. The field of affective computing: An interdisciplinary perspective. *Transactions of the Japanese Society for Artificial Intelligence*, 36(1):13, 2021.

[14] Dagmar M Schuller and Björn W Schuller. A review on five recent and near-future developments in computational processing of emotion in the human voice. *Emotion Review*, 13(1):44–50, 2021.

[15] Yuanyuan Liu, Ke Wang, Lin Wei, Jingying Chen, Yibing Zhan, Dapeng Tao, and Zhe Chen. Affective computing for healthcare: Recent trends, applications, challenges, and beyond, 2024.

[16] Daniel McDuff, Rana El Kaliouby, Jeffrey F. Cohn, and Rosalind W. Picard. Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, 6(3):223–235, 2015.

[17] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111, 2014.

[18] Verma R. Nandwani P. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11, 2021.

[19] Sidney K D'Mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, 47(3):43, 2015.

[20] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[21] Hillary Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological bulletin*, 128:203–35, 03 2002.

[22] Sidney D'Mello, Arvid Kappas, and Jonathan Gratch. The affective computing approach to affect measurement. *Emotion Review*, 10(2):174–183, 2018.

[23] Marc Mehu and Klaus R. Scherer. A psycho-ethological approach to social signal processing. *Cognitive Processing*, 13(2):397–414, 2012.

[24] Yaacov Trope and Ofra Cohen. Perceptual and inferential determinants of behavior-correspondent attributions. *Journal of Experimental Social Psychology*, 25(2):142–158, 1989.

[25] Paul Ekman, Wallace Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William LeCompte, Tom Pitcairn, Pio Ricci Bitti, Klaus Scherer, Masatoshi Tomita, and Athanase Tzavaras. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53:712–7, 10 1987.

[26] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980.

[27] Philipp V. Rouast, Marc T. P. Adam, and Raymond Chiong. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing*, 12(2):524–543, April 2021.

[28] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, and Wenqiang Zhang. A systematic review on affective computing: emotion models, databases, and recent advances. *Information Fusion*, 83-84:19–52, 2022.

[29] William James. What is an emotion? *Mind*, pages 188–205, 1884.

[30] Barrett LF Adolphs R, Mlodinow L. What is an emotion? *Current biology*, 2019.

[31] Lisa Feldman Barrett. Navigating the science of emotion. In Herbert L. Meiselman, editor, *Emotion Measurement*, pages 31–63. Woodhead Publishing, 2016.

[32] James J Gross and Lisa Feldman Barrett. Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion review*, 3(1):8–16, 2011.

[33] Maria Gendron and Lisa Feldman Barrett. Reconstructing the past: A century of ideas about emotion in psychology. *Emotion Review*, 1(4):316–339, 2009. PMID: 20221412.

[34] Silvan S Tomkins. *Affect, imagery, consciousness*. Springer, New York, 1962.

[35] Paul Ekman. *Darwin and facial expression: A century of research in review*. Ishk, 2006.

[36] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[37] Wilhelm Wundt, J. E. Creighton, and E. B. Titchener. Lectures on human and animal psychology. *Philosophical Review*, 4(1):90–93, 1895.

[38] Lisa Feldman Barrett and Eliza Bliss-Moreau. Affect as a psychological primitive. *Advances in experimental social psychology*, 41:167–218, 2009.

[39] David Irons. The nature of emotion. *Philosophical Review*, 6(3):242, 1897.

[40] Magda B Arnold. *Emotion and personality.* Columbia University Press, 1960.

[41] Haiwei Ma and Svetlana Yarosh. A review of affective computing research based on function-component-representation framework. *IEEE Transactions on Affective Computing*, pages 1–1, 2021.

[42] Albert Mehrabian. Communication without words. 1968.

[43] Nusrat J. Shoumy, Li-Minn Ang, Kah Phooi Seng, D.M.Motiur Rahaman, and Tanveer Zia. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications*, 149:102447, 2020.

[44] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103–126, 2020.

[45] Robert Plutchik. Emotions and life: Perspectives from psychology, biology, and evolution. *American Psychological Association*, 2003.

[46] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

[47] Smith K. Khare, Victoria Blanes-Vidal, Esmaeil S. Nadimi, and U. Rajendra Acharya. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, 102:102019, 2024.

[48] Andreas Triantafyllopoulos, Lukas Christ, Alexander Gebhard, Xin Jing, Alexander Kathan, Manuel Milling, Iosif Tsangko, Shahin Amiriparian, and Björn W. Schuller. Beyond deep learning: Charting the next frontiers of affective computing. *Intelligent Computing*, 3:0089, 2024.

[49] Batja Mesquita, Michael Boiger, and Jozefien De Leersnyder. The cultural construction of emotions. *Current opinion in psychology*, 8:31–36, 2016.

[50] James R. Averill. A constructivist view of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, chapter 12, pages 305–339. Academic Press, 1980.

[51] Lisa Feldman Barrett, Christine D Wilson-Mendenhall, and Lawrence W Barsalou. The conceptual act theory: A roadmap. In Lisa Feldman Barrett and James A Russell, editors, *The psychological construction of emotion*, pages 83–110. The Guilford Press, 2015.

[52] Gerald L Clore and Andrew Ortony. Psychological construction in the occ model of emotion. *Emotion Review*, 5(4):335–343, 2013.

[53] Jeremy Pober. What emotions really are (in the theory of constructed emotion). *Philosophy of Science*, 85(4):640–59, 2018.

[54] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 10 2016.

[55] Jennifer K. MacCormack and Kristen A. Lindquist. Bodily contributions to emotion: Schachter's legacy for a psychological constructionist view on emotion. *Emotion Review*, 9(1):36–45, 2017.

[56] Kristen A. Lindquist and Lisa Feldman Barrett. Constructing emotion: The experience of fear as a conceptual act. *Psychological Science*, 19(9):898–903, 2008. PMID: 18947355.

[57] Linda A Wood and Rom Harré. *The Social Construction of Emotions*. Blackwell, 1986.

[58] Lisa Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc*, 10:20–46, 02 2006.

[59] Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, New York, NY, 2017.

[60] Douglas Ramsay and Stephen Woods. Clarifying the roles of homeostasis and allostasis in physiological regulation. *Psychological review*, 121:225–47, 04 2014.

[61] Walter Bradford Cannon. Organization for physiological homeostasis. *Physiological Reviews*, 9:399–431, 1929.

[62] Jay Schulkin and Peter Sterling. Allostasis: a brain-centered, predictive mode of physiological regulation. *Trends in neurosciences*, 42(10):740–752, 2019.

[63] Peter Sterling. Allostasis: a model of predictive regulation. *Physiology & behavior*, 106(1):5–15, 2012.

[64] ROGER C. CONANT and W. ROSS ASHBY. Every good regulator of a system must be a model of that system †. *International Journal of Systems Science*, 1(2):89–97, 1970.

[65] L. W. Barsalou. Ad hoc categories. *Memory and Cognition*, 11:211–277, 1983.

[66] L.F. Barrett, A. Scott, and Macmillan Press. *How Emotions Are Made*. Expert Thinking Series. Pan Macmillan, 2017.

[67] Moshe Bar. The proactive brain: memory for predictions. philos. trans. r. soc. lond. b. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364:1235–43, 05 2009.

[68] Lisa Barrett and Ajay Satpute. Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience Letters*, 693, 07 2017.

[69] Lisa Feldman Barrett. *Seven and a half lessons about the brain*. Pan Macmillan, London, UK, 2020.

[70] Joseph E. LeDoux. *The Deep History of Ourselves: The Four-billion-year Story of How We Got Conscious Brains*. Viking, 2019.

[71] James Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110:145–172, 01 2003.

[72] Hideyoshi Yanagisawa, Oto Kawamata, and Kazutaka Ueda. Modeling emotions associated with novelty at variable uncertainty levels: A bayesian approach. *Frontiers in Computational Neuroscience*, 13, 2019.

[73] Daniel E. Berlyne. Novelty, complexity, and hedonic value. *Attention Perception & Psychophysics*, 8(5):279–286, 1970.

[74] Hideyoshi Yanagisawa and Shimon Honda. Modeling arousal potential of epistemic emotions using bayesian information gain: Inquiry cycle driven by free energy fluctuations. 01 2024.

[75] Hideyoshi Yanagisawa. Free-energy model of emotion potential: Modeling arousal potential as information content induced by complexity and novelty. *Frontiers in Computational Neuroscience*, 15, 2021.

[76] G. Ettlinger. Conflict, arousal and curiosity. by d. e. berlyne new york: Mcgraw-hill publishing company ltd., 1960. pp. 350. *Journal of Mental Science*, 108(452):109–110, 1962.

[77] H. von Helmholtz. Concerning the perceptions in general. In J. P. C. Southall, editor, *Treatise on physiological optics*, volume 3. Dover, New York, 1866/1962.

[78] Beren Millidge, Anil Seth, and Christopher L Buckley. Predictive coding: a theoretical and experimental review, 2022.

[79] Karl Friston. Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352, 2003. Neuroinformatics.

[80] Andrew G. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[81] Andy Clark. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, 01 2016.

[82] Karl Friston. Friston, k.j.: The free-energy principle: a unified brain theory? nat. rev. neurosci. 11, 127-138. *Nature reviews. Neuroscience*, 11:127–38, 02 2010.

[83] Fynn Comerford. Predictive coding: A unified theory of cognition in health and disease?, 2022.

[84] Michał Piekarski. Incorporating (variational) free energy models into mechanisms: the case of predictive processing under the free energy principle. *Synthese*, 202, 08 2023.

[85] Anil Seth and Karl Friston. Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371, 11 2016.

[86] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

[87] Matthew J. Beal. Variational algorithms for approximate bayesian inference. 2003.

[88] Rajesh Rao and Dana Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2:79–87, 02 1999.

[89] David Mumford. On the computational architecture of the neocortex. ii. the role of cortico-cortical loops. *Biological cybernetics*, 66:241–51, 02 1992.

[90] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360:815–36, 04 2005.

[91] Andre M. Bastos, W. Martin Usrey, Rick A. Adams, George R. Mangun, Pascal Fries, and Karl J. Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

[92] Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, 99:102447, 12 2020.

[93] Giovanni Pezzulo, Thomas Parr, and Karl Friston. Active inference as a theory of sentient behavior. *Biological Psychology*, 186:108741, 02 2024.

[94] Ryan Smith, Karl Friston, and Christopher Whyte. A step-by-step tutorial on active inference and its application to empirical data. 01 2021.

[95] Théophile Champion, Marek Grzes, and Howard Bowman. Realising active inference in variational message passing: the outcome-blind certainty seeker théophile champion. *Neural Computation*, 05 2021.

[96] Michael Kirchhoff, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein. The markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, 15:20170792, 01 2018.

[97] Thomas H. B. FitzGerald, Raymond J. Dolan, and Karl Friston. Dopamine, reward learning, and active inference. *Frontiers in Computational Neuroscience*, 9, 2015.

[98] Théophile Champion, Lancelot Da Costa, Howard Bowman, and Marek Grześ. Branching time active inference: The theory and its generality. *Neural Networks*, 151:295–316, 2022.

[99] Karl J. Friston, Richard Rosch, Thomas Parr, Cathy Price, and Howard Bowman. Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, 90:486–501, 2018.

[100] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O'Doherty, and G. Pezzulo. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68, Sep 2016.

[101] Théophile Champion, Howard Bowman, and Marek Grześ. Branching time active inference: empirical study and complexity class analysis. 2022.

[102] Mateus Joffily and Giorgio Coricelli. Emotional valence and the free-energy principle. *PLoS Computational Biology*, 9(6):e1003094, 2013.

[103] R. Hans Phaf and Mark Rotteveel. Affective monitoring: A generic mechanism for affect elicitation. *Frontiers in Psychology*, 3:47, 03 2012.

[104] D. E. Berlyne. Arousal and reinforcement. In *Nebraska Symposium on Motivation*, volume 15, pages 1–110. University of Nebraska Press, 1967.

[105] Conor Heins, Beren Millidge, Daphne Demekas, Brennan Klein, Karl Friston, Iain D. Couzin, and Alexander Tschantz. pymdp: A python library for active inference in discrete state spaces. *Journal of Open Source Software*, 7(73):4098, 2022.