UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

DIPARTIMENTO DI MATEMATICA 'FEDERIGO ENRIQUES'
CORSO DI DOTTORATO IN SCIENZE MATEMATICHE
XXX CICLO

TESI DI DOTTORATO DI RICERCA

A PROBABILISTIC APPROACH TO THE CONSTRUCTION OF A MULTIMODAL AFFECT SPACE

SSD ING-INF/05

Autore
Vittorio Cuculo

Tutor
Prof. Giuseppe Boccignone

Coordinatore del Dottorato
Prof. Vieri Mastropietro

A.A. 2016-2017

This thesis is dedicated to my son,
who is not born yet and gave me the time to write it.

Acknowledgments

"A Bayesian is one who, vaguely expecting a horse,
and catching a glimpse of a donkey,
strongly believes he has seen a mule."

attributed to Karl Pearson

This thesis would not have been possible without the invaluable guidance of my supervisor, Professor Giuseppe 'Beppe' Boccignone. He made me know, appreciate and passionate about a research field that I did not even know the existence. He believed in me and backed me up during all these three years, even when my presence was more latent. I feel fortunate to have shared the beginning of my experience with the birth of the PHuSe Lab, where the work and the presence of Dr. Raffaella 'Lella' Lanzarotti and Dr. Giuliano Grossi have been crucial in these years. They have been always and widely available anytime I needed it. I would thank Prof. Paola Campadelli because, even though I have never had the fortune to work with her, we had many stimulating discussions and shared many pleasant lunches together. I also want to thank my lab mates, Claudio 'Ceru' Ceruti and the new acquisition, Alessandro D'Amelio. The former, because all the 'little' I know about DNN is thanks to him, and the latter because he has always been ready to share all my 'many' perplexities about Bayesian statistics.

I must thank all the members of the interaction studio 'dotdotdot', and in particular Alessandro 'Cemma' Masserdotti. They all believed in me and always supported me whatever I did, as does the parents with their son, even when it went against their direct interests. They allowed me to work and study during the first two years, to follow my choices and freely define my path.

A big thank to all my friends. Those who, although time and distance make it difficult to meet often, when it happens is as if we were never separated. The list would be long, but they already know who I am referring to.

Last but certainly not least, thanks to my family. The first teachers of my life, Antonella and Domenico. They have always followed me with their love, pride and support even from a great distance, leaving me free to make my choices. Lucia and Carmine, who have welcomed me as if they were their son and Luca, the younger brother I've never had. Finally, thanks to my wife Giulia for always be by my side and to bring that little life in her womb, that although I have never met yet, has already made my life better.

Abstract

UNDERSTANDING affective signals from others is crucial for both human-human and human-agent interaction. The automatic analysis of emotion is by and large addressed as a pattern recognition problem which grounds in early psychological theories of emotion. Suitable features are first extracted and then used as input to classification (discrete emotion recognition) or regression (continuous affect detection). In this thesis, differently from many computational models in the literature, we draw on a simulationist approach to the analysis of facially displayed emotions - e.g., in the course of a face-to-face interaction between an expresser and an observer. At the heart of such perspective lies the enactment of the perceived emotion in the observer. We propose a probabilistic framework based on a deep latent representation of a continuous affect space, which can be exploited for both the estimation and the enactment of affective states in a multimodal space.

Namely, we consider the observed facial expression together with physiological activations driven by internal autonomic activity. The rationale behind the approach lies in the large body of evidence from affective neuroscience showing that when we observe emotional facial expressions, we react with congruent facial mimicry.

Further, in more complex situations, affect understanding is likely to rely on a comprehensive representation grounding the reconstruction of the state of the body associated with the displayed emotion. We show that our approach can address such problems in a unified and principled perspective, thus avoiding *ad hoc* heuristics while minimising learning efforts. Moreover, our model improves the inferred belief through the adoption of an inner loop of measurements and predictions within the central affect state-space, that realise the dynamics of the affect enactment.

Results so far achieved have been obtained by adopting two publicly available multimodal corpora.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

SEAMLESSLY, in the course of our entanglements and conflicts, dealings and struggles, we "perceive" the social signals brought on by others, and we recognise and understand their meaning. Yet, gazing at a gesture, glimpsing a smile or hearing a laugh involves a kind of perception which is different from the appraisal of the lifeless world. A large body of evidence (Rizzolatti and Sinigaglia, 2016; Gallese et al., 2004) shows that alongside the sensory information concerning others' social stimuli - actions, in a wide sense -, one's own motor and visceromotor representations of those stimuli are enacted. Humans mirror gestures, postures, emotions, speech of other perceived humans, at least neurally, and sometimes bodily and behaviourally. These mirroring processes ground the capability of own reproduction of the action in question "as if" a similar action were performed or a similar emotion experienced. Such simulation-based mechanism is likely to play a crucial role in individual cognition and social interaction (Rizzolatti and Sinigaglia, 2016; Gallese et al., 2004).

The rationale behind this study is thus straightforward and stems from the attempt at answering a deceptively simple, albeit overlooked question: can we exploit such primitive and fundamental simulation-based mechanism for designing artificial agents?

Answering this question is important as witnessed by the growing interest for the computational modelling of emotion in the realms of natural interaction, cognitive robotics (Ziemke and Lowe, 2009), gaming industry (but see (Breazeal, 2003a; Trovato et al., 2013; Lim and Okuno, 2015) for specific discussion and examples). Precisely, it is of twofold value. First, at the individual level ("internal") emotions can influence reasoning and planning as happens in humans (Damasio, 1999); second, at the social level ("external", involving expression and detection) participating in interaction (e.g., human-robot or with animated virtual characters as in computer games) requires affective interplay.

If we go along with these assumptions that emotions are important for artificial cognitive systems, then the issue arises as to how such mechanisms can be suitably modelled. In a subject such as this, it is perhaps best to start by establishing models of an apt generality, so to avoid *ad hoc* heuristics, while considering relevant and yet well defined case studies, in order not to complicate an already difficult problem.

Hence, in this thesis we shall focus on the case of emotional facial expressions, but in a simulation-based context also involving the observer's physiological reactions (see (Adolphs, 2002a) and (Wood et al., 2016) for a review). In particular we address the issue of facial expression mirroring and mimicry, which is at the heart of simulationist accounts.

Face perception is likely to be the most developed visual perceptual skill in humans and, cogently, most face viewing occurs in the context of social interactions (Wood et al., 2016). Not surprisingly, automatic analysis of facial affect has fostered a wealth of approaches (for an in-depth review, see (Sariyanidi et al., 2015)). By and large, these approaches - grounding in early inferential theories of emotion (Goldman and Sripada, 2005) - share the assumption that understanding affective states from facial expressions can be accomplished through a computer vision and pattern recognition "pipeline" (Sariyanidi et al., 2015): namely, visual feature extraction/reduction followed by classification (discrete emotion recognition) or regression (continuous affect detection).

The (often concealed) rationale behind such attitude grounds in early psychological theories of emotion: the visual perception of emotional stimuli is followed by cognitive or appraisal/interpretative processing of the stimulus, which in turn triggers affective responses and feelings. Spelling out such account for faces, when an expresser is observed displaying a particular facial expression, the observer utilises the large amount of visual features/properties to infer and attribute an emotion state to the expresser.

However, facial expressions are facial actions and their perception is likely to draw on simulation mechanisms underlying action perception in general (Wood et al., 2016; Adolphs, 2002b). Clearly, part of the ability to extract affective information from faces can be attributed to visual expertise, and current computer vision techniques can provide flexibility and adaptivity to robotics systems (Fanello et al., 2017).

But people do indeed recognise emotions from other people's faces by experiencing changes in their own physiological state. Beyond visual expertise, visuomotor and visceromotor simulations play a key role. Indeed, when we observe emotional facial expressions, we react with congruent facial muscle mimicry, and this mirroring process is likely to reflect internal embodied simulation of the perceived facial expression (see Figure 1.1).

Visuomotor simulation concurrently supports cross-modal influences on visual processing and the activation of autonomic activities. Jointly, mirroring mechanisms concerning autonomic, visceromotor actions take place. Altogether, they participate in building a deep understanding of the perceived affective expression.

The lesson taken from affective neuroscience is that the development of social cognition is closely related to the development of emotional and affective communication between an infant and his or her mother, which grounds our remarkable capacity to share others' affective states and empathise with them. Actions, emotions and sensations experienced by others become meaningful to us because we can share with them:

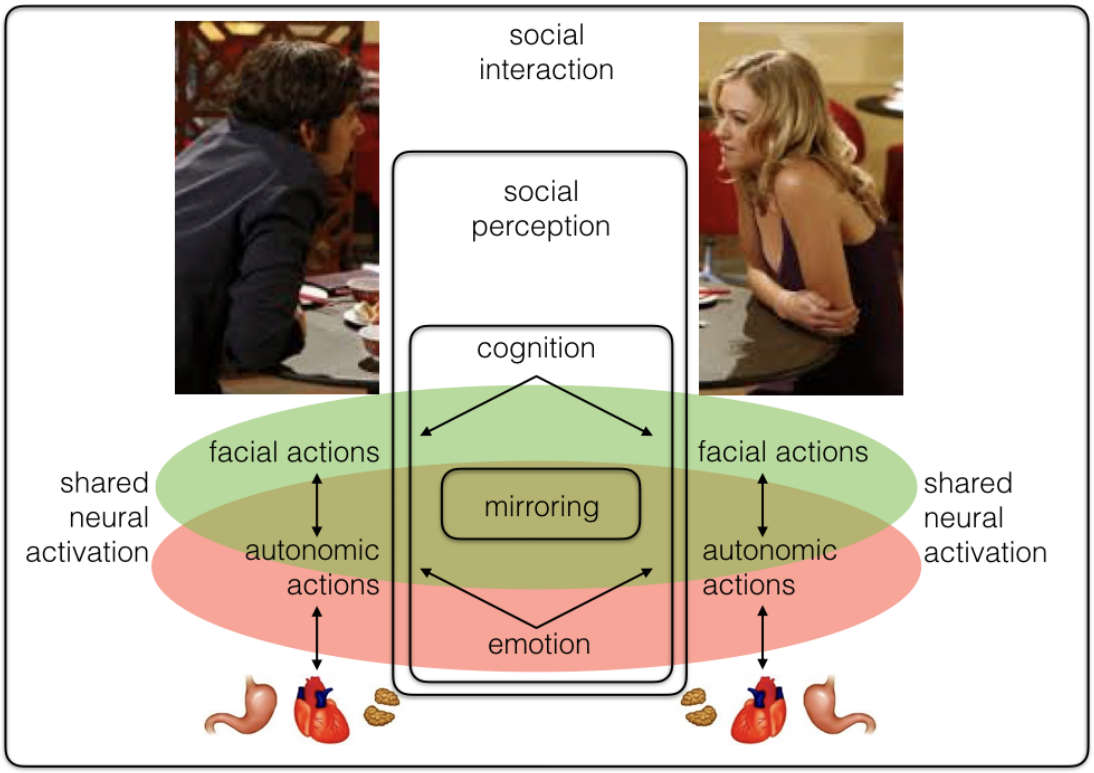


Figure 1.1: *Emotion processing during social dyadic interaction. Emotions that are expressed by one person, influence the emotions and expressions of the other person. The two individuals not only mimic each other's facial expression, but they also link and synchronise with one another at the physiological level. This basic mirroring processing grounds in the common shared neural architecture and body-proper, which is the basis for social perception, understanding and empathy.*

a capability that roots in the very fact that evolution has endowed us with common neural structures and body-proper.

In this dissertation we propose a novel, probabilistic scheme for dynamic affective facial expression processing relying on a mirroring mechanism, which involves both facial gesture and autonomic, physiological simulation. In brief, in the course of a dyadic engagement, the observer's visual system, while perceiving expresser's facial display (Figure 1.2), interacts with an extended system, which involves the emotion system. Interaction is regulated by the mediation of a visuomotor component for somatic action perception. The latter transforms the sensory information of observed facial actions into the observer's own motor representation, which, in turn, via interaction with a central valence/arousal affect space, triggers visceromotor processes. The simulation-based dynamics of the visuomotor and visceromotor routes can generate observer's actual responses, namely facial mimicry (observable) and physiological responses (hidden).

The overall goal of the approach is to allow the modelled observer to reach a core affect state (in terms of valence and arousal) similar to that of the expresser. Indeed, meeting such condition is preliminary, in the embodied perspective, to ground subsequent processing for affect understanding, e.g. the retrieval of conceptual knowledge

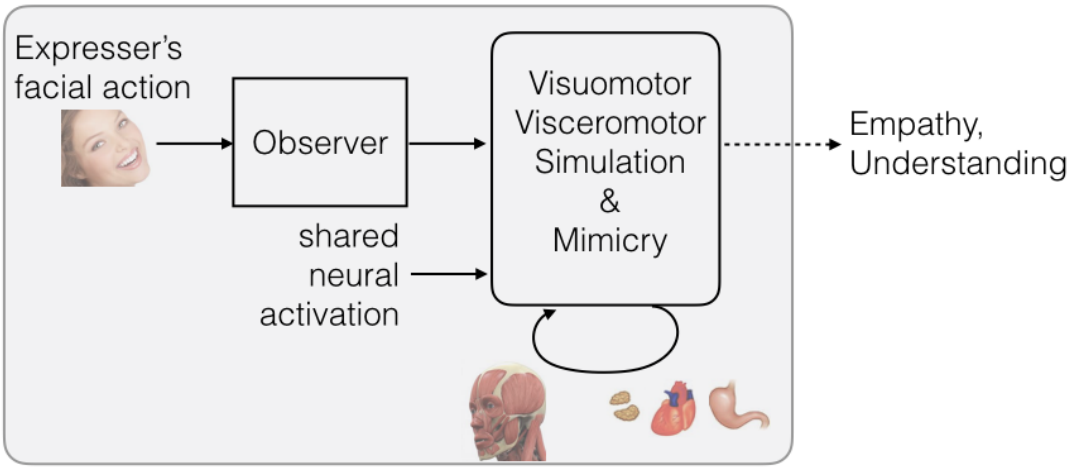


Figure 1.2: *The shaded grey box highlights the process that we are modelling in this thesis. The expresser's emotional state is reflected in nonverbal motor movements, namely, a facial action. Observer's perception activates neural representations (shared with the expresser) that in turn triggers somatic and autonomic responses resulting in somatomotor mimicry and autonomic, visceromotor reactions. A simulation-based loop facilitates physiological and motor feedback enacting emotion in the observer. These mechanisms are fundamental for further grounding empathy and other's understanding, issues that will not be addressed here.*

about the emotion signalled by the expresser (Adolphs, 2002a; Wood et al., 2016).

The dissertation unfolds as follows (see Figure 1.3 for a graphical illustration):

Chapter 2 broadly reviews the different contributions to the conundrum of affective modelling. The attempt here is also to identify some general perspectives to better frame our subsequent work.

Chapter 3 discusses the rationales behind our approach and presents the necessary neurobiological background for further modelling steps. The chapter has two objectives: firstly, to make clear what is the methodology followed in this thesis, namely, a multilevel analysis of the problem, which is instantiated at the theoretical model and at the implementation model levels. Secondly, to devise a suitable distributed neural architecture to be used as a blueprint, at a more abstract level, for a functional architecture. The latter will be used to constrain the structure of the probabilistic graph structure, which scaffolds the theoretical model. The somatomotor and visceromotor routes together with the core affect mediating between them are introduced and defined.

Chapter 4 formalises the proposed model. At the theoretical level, the model is instantiated in the form of a dynamic Probabilistic Graphical Model. The graph structure is shaped on the basis of the functional architecture previously devised, which constrains conditional independence assumptions between the graph nodes. A simulation-based internal loop is defined, hierarchically involving the somatomotor and visceromotor components with central affect mediation. The simu-

lation loop unfolds the dynamics of the observer's emotional perception of the expresser's facial actions. Subsequently, the implementation model is discussed. The aim here is to devise - relying on the compositional representation provided by the probabilistic graph -, an effective realisation of each sub-graph corresponding to the main functional components. As to the central affect level, we resort to a Deep Gaussian Process model. The somatomotor component assumes at its core an internal somatomotor state-space, which serves the purpose of both enacting observer's actual facial mimicry and predicting the dynamics of expresser's visual cues. A visuomotor mapping is defined to support the transformation from expresser's facial visible cues to observer's internal motor representation. A prediction-correction approach is also devised for the visceromotor state-space, in terms of a variational Kalman filter, to learn and represent internal stochastic dynamics of physiological signals.

Chapter 5 illustrates the experimental work to validate the model and discusses results achieved so far.

Chapter 6 discusses some theoretical implications of the model;

Chapter 7 summarises the key contributions of the thesis and presents some concluding remarks.

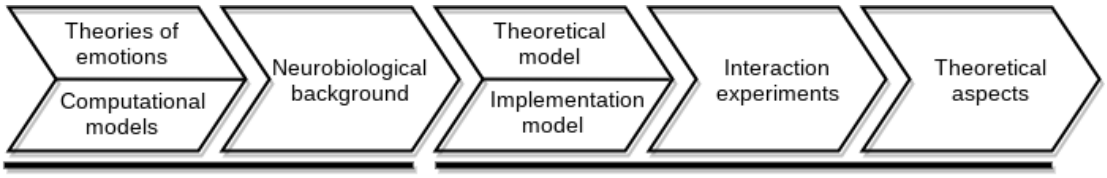


Figure 1.3: *Diagram of the presentation flow adopted in this thesis. Each step corresponds to a specific chapter, starting from Chapter 2 and ending in Chapter 6. The black lines below the modules ideally divide the thesis into two main parts.*

CHAPTER 2

What is an emotion? State-of-the-art of different accounts

TRYING to review the contributions to the field of computational models of affect, and thus of affective facial expressions, in order to systematically compare the different approaches is a mind-blowing endeavour. That being so, we arm the reader with a preliminary map outlined in Figure 2.1.

Such map shows at a glance that there are different research programs that contend for such field, which can be roughly identified as machine learning-based models, robotic models, artificial intelligence (AI) based models. Several research areas have contributed, in different vein and substance, to such programs, psychological and neurobiological theories providing the necessary underpinning, at least in our view. Henceforth, from such theories we will make a start on. Other areas that have fostered the flourish of methods are machine-learning and artificial intelligence (AI), either in their classic knowledge-based or the probabilistic, uncertainty-based accounts.

Although it is not possible to exhaustively cover a century of scientific research on emotion, in this chapter we will not only make an effort to get the gist of the field, but to outline some general trends/dimensions to frame our work and to better compare it with other proposals.

Early theories

The "theories of emotion" ground their roots in the great classical philosophers as Plato, Aristotle and Descartes. The dominant theory, advanced by the Stoics in the 3rd century AD, was to consider the emotions as a threat to reason, a danger that alters the opinion of humans and therefore an obstacle to virtue. This dualistic vision that clearly

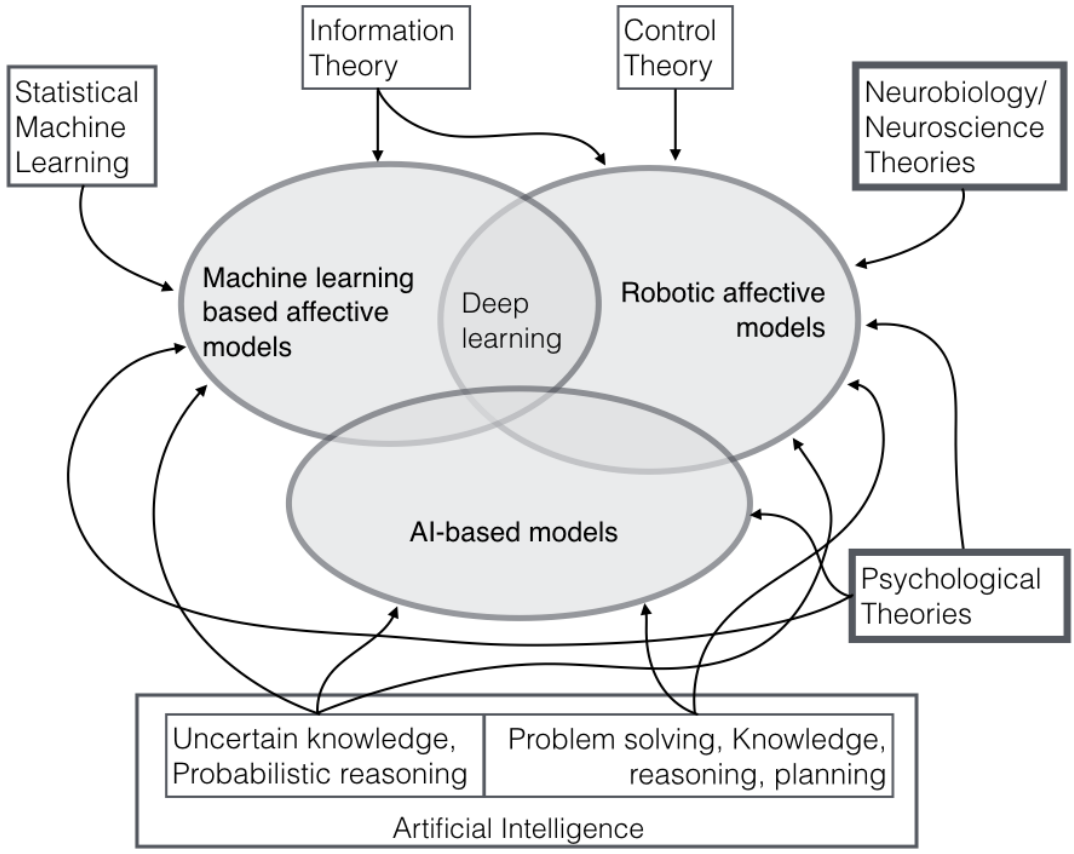


Figure 2.1: *A map of the main research programs in which the field of computational modelling of affect has been developed: machine learning-based models, robotic models, artificial intelligence (AI) based models. Boxes and related arrows represent independent research areas and theories that have mainly contributed to such field.*

separates 'emotion' and 'reason' is also carried out by Socrates and his student Plato, and represents the basis of a secular debate among supporters of Platonic theses and those who claim for emotions a different role. An alternative view, proposed by Aristotle in *Rhetoric*, establishes a relationship between reason and emotion as it argues that some physical reactions are caused by our convictions and our way of interpreting the world and the people around us. Emotions (or *pathe*), therefore, are considered for the first time as the result of a rational process.

This dichotomous conception remains in force throughout the Middle Ages, until the *Les passions de l'âme* (1649) of Descartes and the theories of Spinoza that enshrine the beginning of scientific studies on emotions. A century later, during the fervour of Scottish Enlightenment, David Hume in the second, undervalued, part of *A Treatise of Human Nature* (1739) defined the emotion as a 'kind of impression', pleasant or unpleasant, which (as in Descartes) is physically elicited by the movement of the 'animal spirits' in the blood.

Going forward to the nineteenth century, Charles Darwin studies mark the beginning of modern evolutionary theories. In his 1872 work, *The Expression of the Emotions in Man and Animals*, Darwin proposed the 'principle of serviceable habits', claiming that

the expressions of emotions are inherited from habitual patterns of behaviour developed during the evolution of primate.

At this point we may assert that an emotion constitutes an internal state with a related behaviour which is triggered by a specific stimuli, endogenous or exogenous to the organism. A crucial point in the theories of emotions has been the relationship and the direction of causality between the internal state and the physical action or behaviour. Some of these theories are listed in the next sections, differentiating between purely psychological ones, in Section 2.2, and those with neurobiological bases, in Section 2.3.

Around 1884, the American psychologist William James (James, 1884) and the Danish Carl Lange, published, independently of one another, a similar theory of emotion. The purpose was to challenge what they called the theory of 'common sense', according to which when someone is asked why he shakes, he usually answers: 'Because I'm afraid', or, as he cries, he answers: 'Because I'm sad'. These responses imply that the feelings come first, which, in turn, produce the physiological and behavioural aspects of emotions. According to James and Lange, the theory of common sense must be reversed, since we do not cry because we are sad, but we feel sad because we cry; we do not shake because we are scared, but we are scared because we are shaking. In brief, James-Lange's theory argues that emotion is the sensation of physiological modifications.

James-Lange's theory has dominated for several years and has stimulated numerous research into the physiological processes involved in emotional states. In 1927, a physiologist, Walter Cannon (Cannon, 1927), published a criticism that met a remarkable success. His writings raised many and important objections to James-Lange's theory and convinced many psychologists that it was an unsustainable theory.

Cannon pointed out that visceral organs are relatively insensitive structures, scarcely provided with nerves. For this reason, visceral modifications are rather slower than the changes we feel in emotional states. So how can visceral changes produce our rapid mood changes? In addition, emotional reactions are also present when the visceral organs have been surgically isolated by the central nervous system (CNS). When the spinal cord and vagus nerve were cut in a dog, so that the bowels had no connection to the brain anymore, the animal was still acting as if it were experiencing emotions. When it was threatened or hit, it barked, snarled, just as it did before the intervention.

Cannon (1927) advanced his hypothesis on the origin of emotions, hypothesis that was subsequently elaborated by Bard (1929), according to which the thalamus plays a critical role in emotional experience. For Cannon and Bard, nerve impulses that pass sensory information are then retransmitted through the thalamus, that receives this input upward from the cortex (causing a subjective emotional experience) and downward to the muscles, glands, and visceral organs (producing physiological modifications). Cannon and Bard argued that the subjective and physiological components of emotion are simultaneous, in disagreement with James, who argued that physiological modifications precede and trigger subjective states.

Psychological theories

Towards the end of the 1950s, cognitivist psychologists expanded this view, suggesting that the most important factor of the emotion we feel is the way we evaluate and interpret situations. In other words, it is not the environment itself that affects us, but the way we represent the environment itself. One of the first cognitive theories of emotion was formulated by Arnold (1960). Her theory of cognitive assessment suggests that when we come for the first time in a situation we evaluate it spontaneously as good or bad, useful or damaging. According to Arnold, these first evaluations are mediated by the limbic system. The evaluations, in turn, introduce 'tendencies to act'. Emotions emerge from both our assessment of the situation and our actions. For example, joy appears when we evaluate something as good and we are pushed in its direction. There is, instead, a state of anger when we judge an event bad and we feel its aversion.

In 1962, Stanley Schachter and Jerome Singer proposed a two-factor theory of emotion (Schachter and Singer, 1962). The theory assumed that when an emotion is felt there are two components: a physiological arousal and an interpretation (appraisal) of the state of arousal, which determines which emotion will be experienced. The specific emotion depends largely on the situation and the immediate environment where the person search for emotional cues to label the physiological arousal (Fig. 2.2).

The idea of appraisal since then has been dominant in the cognitive approach to emotions. This been extensively described in the *Handbook of Cognition and Emotion* (Dalgleish and Power, 2005) or by Ellsworth and Scherer (2003) and encompasses the work of psychologists such as Bower, Mandler, Lazarus, Roseman, Ortony, Scherer, and Frijda, among the most influential.

In this perspective, the cognitive-motivational-relational theory (Lazarus, 1991; Roseman, 1984) states that in order to predict how a person will react to a situation, the person's expectations and goals in relation to the situation must be known. The theory describes how specific emotions arise out of personal conceptions of a situation. Roseman (1984) have developed structural theories where a set of discrete emotions is modelled as direct outcomes of a multidimensional appraisal process. The cognitive theory by Ortony, Clore, and Collins (OCC) views emotions as reactions to situational appraisals of events, actors, and objects (Ortony et al., 1990). The emotions can be positive or negative depending on the desirability of the situation. They identified four sources of evidence that can be used to test emotion theories: language, self-report, behaviour, and physiology. The goal of much of their work was to create a computationally tractable model of emotion.

Roseman's and Ortony's theories are similar in many ways. They both converge upon the "universality" of the appraisal process. These models can be used to automatically predict a user's emotional state by taking a point in the multidimensional appraisal space (think of each contextual feature or appraisal variable as a dimension) and returning the most probable emotion. Despite its success, the appraisal theories are not able to explain a number of questions. Briefly, these include issues such as:

1. the minimal number of appraisal criteria needed to explain emotion differentiation;

2. should appraisal be considered a process, an ongoing series of evaluations, or separate emotion episodes;

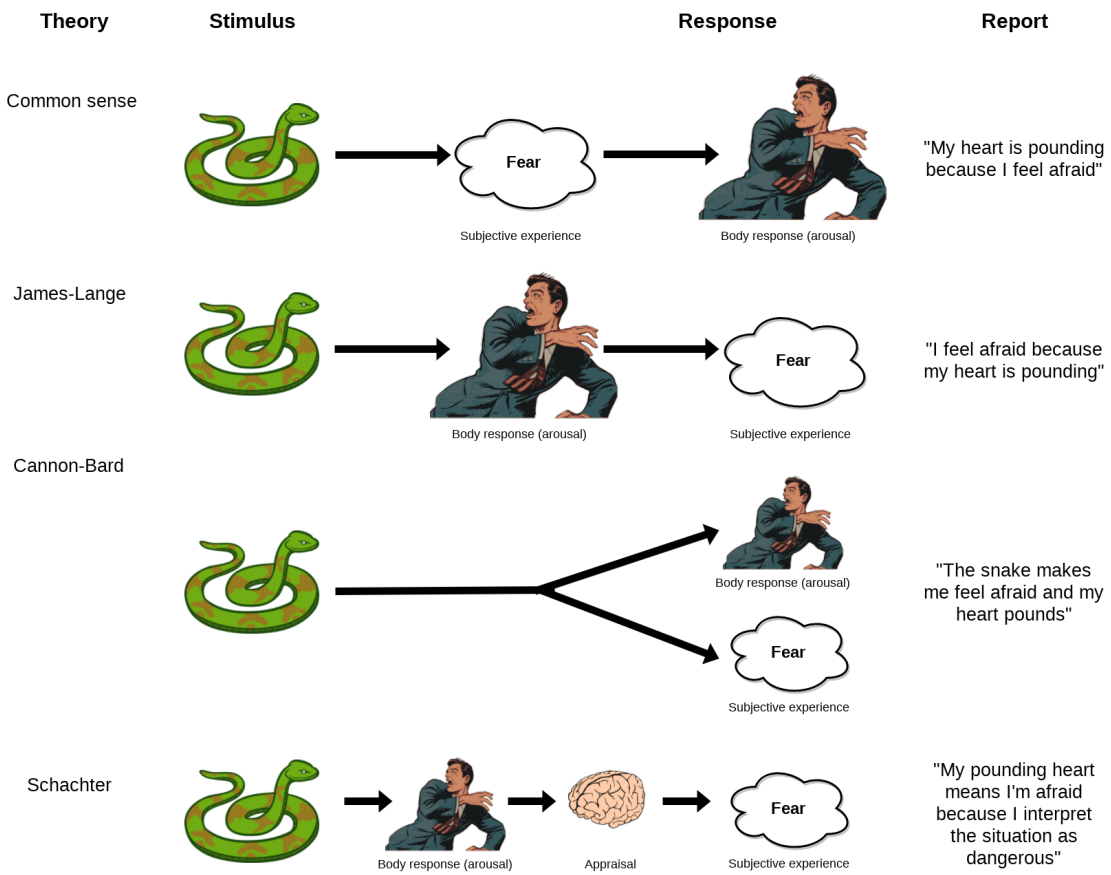


Figure 2.2: *Schematic representation of the three main psychological theories of emotions, compared to common sense.*

3. how can the social functions of emotion be considered within the cognitive approach;

4. how do appraisal theories explain evidence that preferences (simple emotional reactions) do not need any conscious registration, or that emotions can exist without cognition.

Despite these questions and limitations, as we will see in brief, the computational models derived from appraisal theories have proven to be useful to both Artificial Intelligence (AI) researchers and psychologists by providing both groups with a way to evaluate, modify, and improve their theories.

A different view stems from Darwin's work, which mostly addresses the communicative function of emotions (*emotions as expressions*). Researchers have expanded Darwin's evolutionary framework toward other forms of emotional expression. The most notable ones are from Tomkins (1962), proposing that there is a limited number of pan-cultural basic emotions, such as: surprise, interest, joy, rage, fear, disgust, shame, and anguish, and by Ekman and Friesen (1971). In particular, Ekman's facial action coding system (FACS, Ekman and Rosenberg, 1997) influenced considerable research that tackles the affect detection problem developing systems that identify the basic emo-

tions though facial expressions (and in particular extracting facial action units).

Neurobiological theories and embodiment

Despite the efforts of the various currents of psychology, it was only with the work of the neuroanatomist Papez that emotion started to have a coherent theory of flows involved in emotional and memory functions. In confirming the importance of the structural component, and not just the psychological one, in the analysis of behaviour and discomfort situations, Papez was the first to imagine the existence of a real intracerebral pathway (Circuit of Papez) involved in emotional responses. A pathway that for the neuroanatomist follows the route cortex-hypothalamus-thalamus-cortex.

Finally, Paul MacLean's studies (MacLean, 1949) gave a clear and exhaustive picture of the existing ties between cognitive and emotional processes. According to MacLean, emotional stimuli of the outside world produce "in the bowels" reactions, the messages of these reactions return to the brain (neocortex) where they are integrated with the perceptions that come from the outside world through the thalamus. The integration of these perceptions (visceral and cortical) is the mechanism that generates the emotional experience. If Papez is recognised with the intuition that emotions do not originate in a precise point in the brain but in an action involving multiple nuclei (circuit), MacLean deserves to have guessed that the amygdala plays a central role in communication between the upper cortical areas and the hypothalamus, namely that part of our brain where the central nervous system connects to the endocrine system which is then entrusted to the hormonal responses of our body.

MacLean's conjectures have been experimentally confirmed by several neuroscientists including Joseph LeDoux. His studies (LeDoux, 1998) have made it clear that some of the fundamental structures, including hippocampus, are specialized in recording and understanding perceptual patterns while others, such as the amygdala and the circuits connecting it to other parts of the brain, are more involved in emotional reactions.

More recently, in 1991, the studies on the body implications in emotion have been carried on by Damasio and colleagues, with the outline of the somatic marker hypothesis (Damasio, 1997) that considers, instead, the role of the prefrontal cortex (PFC).

The somatic marker hypothesis reflects the original ideas of James and Lange, and finds its roots on the earlier work of Nauta (1971), who coined the term 'interoceptive markers', and Pribram (1970), who used the phrase 'feelings as monitors'.

The decision process, to make a choice between two or more alternatives, according to Damasio, is often far from an analysis that takes into account the pros and cons of each choice. Most often, especially when dealing with complex problems, we are led to use a different strategy that refers to the outcomes of past experiences, in which we recognize some analogy with the present situation. Those experiences have left traces, not necessarily conscious, that call into us emotions and feelings, with negative or positive connotations.

Damasio calls these somatic codes, processed in the ventromedial PFC, somatic markers: somatic because they relate to bodily experiences, both visceral and non-visceral; the mark term comes from the notion that the particular body state invoked is a sort of "marker" or label.

Current knowledge, however, does not allow the 'cognitive' brain to be so clearly separated from the 'emotional' brain. In Pessoa (2008) the brain conceptualization as composed by 'affective' and 'cognitive' regions is questioned. Pessoa, indeed, states that the separation in 'emotional brain' and 'cognitive brain' is problematic for a number of reasons. In fact, brain regions considered as 'affective' are involved in cognitive processes and vice versa; moreover, the cognitive and emotional processes are clearly integrated with each other so that they jointly contribute to behaviour. Such state of affairs can be summarised as in Figure 2.3. It is worth remarking that such a multiscale

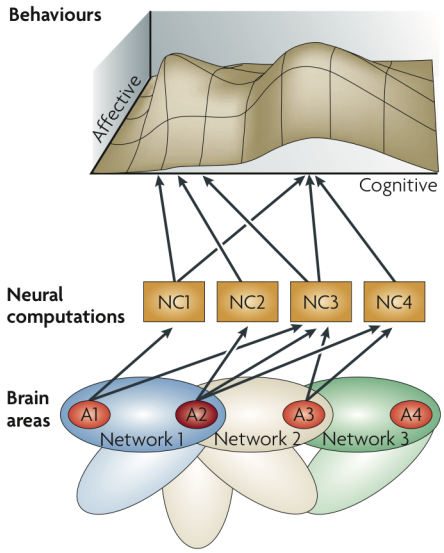


Figure 2.3: *Pessoa's conceptual proposal for the relationship between anatomical sites, neural computations and behaviours. Brain areas, which are connected to form networks (ellipses), are involved in multiple neural computations and specific computations are carried out by several areas. Therefore, the structure/function mapping is both one-to-many and many-to-one; in other words, many-to-many. Multiple neural computations underlie behaviour. Each behaviour has both affective and cognitive components, indicated by the affective and cognitive axes. Axes are not orthogonal, indicating that the dimensions are not independent from each other. (Figure from Pessoa (2008)).*

nature of cognitive and behavioural processes calls for a multi-level analysis of these processes, an issue we will confront with in Section 4.

We will not go further in the review of neurobiological facets of emotion, because these aspects are cogent to the work presented in this thesis and they will be discussed in some depth in Section 4.

A broader view on the theories of emotion

In the two former sections we presented a brief account of the developed theories of emotions, both psychological and neurobiological. At a first sight it is somehow difficult to identify some "essential variables" suitable to provide a clear road map in order to exploit them for the construction of a model, namely a computational model. In this section, before reviewing the current state of affairs in computational models we thus discuss some broad dimensions in order to more firmly ground our efforts. These

dimensions can be defined in terms of the fundamental rationales behind the different approaches.

The generative problem: assumptions on the arrow of causality

The many theories previously introduced reflect a strong disagreement over the primary question of the direction of causality between emotional states and their associated behaviours. A common intuition, shared by Charles Darwin, is that the emotional state causes a related action. However, the psychological view of 'I cry because I am sad' is not the predominant one, which typically makes the behaviour a cause of the emotion. William James, indeed, argued that the direction of causality is the opposite of the one mentioned above: 'I feel afraid because I run from the bear'. Even if this version of the relationship between emotion and action may seem counterintuitive and others have argued against it (Cannon, 1927), it remains a defended view.

The constructionist view, for example, suggests that emotions emerge from the interaction of several factors, among which the core affect is the prominent one. The *core affect*, as described in Russell (2003), is an internal state that can be experienced as free floating, resulting in mood, or can be attributed to some cause, bringing to the begin of a perceived emotional episode. In the latter case, there are involved processes responsible of the perception of the altering properties of stimuli as well as physiological and expressive changes.

As presented in Barrett and Russell (2014), this approach can be formalised adopting the *emergent variable model*, which claims that emotions do not cause, but are caused by their measured indicators. The model does not require any covariation among those indicators, nor any central neural circuit. Referring to Figure 2.4, in particular, the categorical emotion of 'fear' is the result of co-occurrence of multiple indicators weighted by the b values assigned to each of them. Such indicators are: amygdala activation, autonomic nervous system (ANS), freezing behaviour (action) and subjective experience, as well as other unmeasured indicators. The model error is represented by a disturbance term (D_{fear}), which includes all the measurement errors.

Its counterpart model is the so called *latent variable model* of emotion (Fig. 2.5), which assumes a central emotion circuit whose variation precedes activations in its measured indicators (E). This corresponds to a strong covariation among indicators, resulting in a relatively little need to measure multiple indicators.

How many emotions are there? Discrete, dimensional and componential approaches

In general, psychological theories mostly disagree on how we attribute emotions to people, and also at the neurobiological level there is some debate about this. The main answers to these issues can be summarised as follows.

Discrete theories These theories consider the emotions as a small set of discrete states. The assumption of this approach is that these fundamental, discrete emotions are biologically determined, and associated with a distinct patterns of behavioural expression that is shared across people and cultures. Indeed, since Darwin (1872), researchers have attempted to bring this subjective experience, how can be an emotion, to universal patterns. Tomkins (1962) proposed that there is a limited number of pan-cultural basic

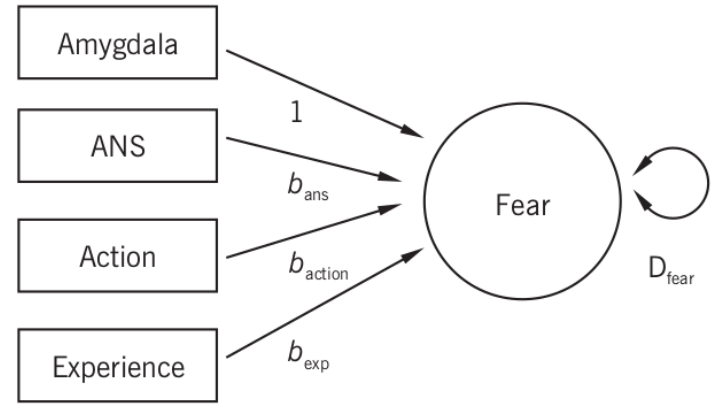


Figure 2.4: *The emergent variable model of fear expressed adopting structural equation model. In this representation, latent variables are drawn as circles. Manifest or measured variables are shown as squares. Residuals and variances are drawn as double headed arrows into an object. (Figure from Barrett and Russell (2014)).*

emotions, such as: surprise, interest, joy, rage, fear, disgust, shame, and anguish. Years later, Ekman and Friesen (1971), after performing a series of cross-cultural studies, came up with a shorter list, consisting of six basic emotions with universal facial expressions: anger, disgust, fear, happiness, sadness and surprise. Because of its emphasis on discrete categories of states, this approach is also named 'categorical approach' (Panksepp, 2004).

Dimensional theories Dimensional theories arose from the biological critics to discrete emotion categories. Indeed, they claim that there is no brain region or circuit that is unique to a single emotion category. These theories, instead, argue that an affective state should be rather expressed in terms of a point coordinate in a continuous two (or three) dimensional space. The axes of such space, when declined in two dimensions, typically are valence and arousal (Mehrabian and Russell, 1974; Russell, 2003). The former corresponds to a positive or negative affective state (e.g. happiness/sadness), while the latter is proportional to the intensity of the emotional state. A low arousal value is associated with low intensity (e.g. sadness), while high arousal with more energy (e.g. anger). Anyway, even this approach hides some drawbacks. Emotional states as anger and fear would share the same position in a 2D space (high arousal, low valence). For this reason, a third dimension, named 'dominance', is often added. The resulting space is then referred to as the PAD space, consisting of: Pleasure (valence), Arousal and Dominance (Mehrabian, 1980). The representational semantic primitives within this theoretical perspective are thus these 2 or 3 dimensions.

Componential theories The third view is that related to "appraisal theory" previously introduced (Lazarus, 1991; Scherer, 2001). All such theories involve a large amount of appraisal variables (Ortony et al., 1988), checks (Scherer, 2001), or dimensions (Roseman, 2001) (depending on the considered theory) from which a large set of emotions

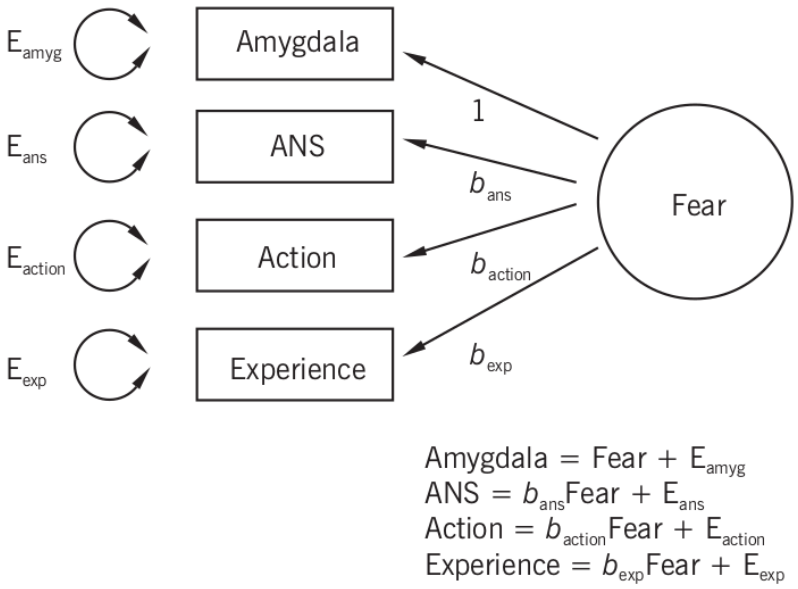


Figure 2.5: *The latent variable model of fear. (Figure from Barrett and Russell (2014)).*

can be derived. These variables consider the relationship between the individual and the environment in terms of meaning and consequences for the former. Scherer (2001), for example, proposes four categories of variables: relevance, implications, coping potential and norms; going from stimuli-specific properties to increasingly complex cognitive processes of the agent. Their evaluation consists in assigning a definite value to each of the appraisal variables, resulting in a vector of assessments that corresponds to a point in the n-dimensional space defined by the n appraisal variables. Each point is related to a specific emotion or affective state, resulting in a wide range of emotions, including states of varying intensity.

A summary of the presented theoretical perspectives is formalised in Table 2.1, where for each theory it is shown the mapping of the considered input and output variables, declined in the application of emotion detection.

How do we attribute emotions? The Theory-Theory vs. Simulation Theory debate

Until this point we considered the emotions as the result of an internal processing of (internal or external) emotional stimuli. In the course of our life, however, it is far from unlikely that the considered external stimulus is represented by the emotional output coming from another human being.

This scenario results in the interaction of two or more independent "emotional models", where each part tries to understand and decode the visible cues coming from the counterpart. This ability, known as "mindreading" (Goldman and Sripada, 2005), represents the capacity to identify the mental states of others, including emotional states. Also in this case, two are the main schools of thought. Briefly:

1. the Theory-Theory (TT), posits that a mental-state attributer deploys a naive psychological theory (innate or acquired) to infer mental states in others from their behaviour;

Table 2.1: *Formalisation of three theoretical perspective, declined in the emotion detection application.* $\mathbf{E}$*: emotional event,* $\mathbf{X}$*: affective abstraction,* $\mathbf{A}$*: affective state.*

Theoretical perspective	**Mapping**	**Example**
Discrete	$\mathbf{E} \rightarrow \mathbf{X}$	Dangerous event $\rightarrow$ fear
Dimensional	1. $\mathbf{E} \rightarrow \mathbf{X}$ $\mathbf{X} \in \{V, A\}$ or $\mathbf{X} \in \{P, A, D\}$ 2. $\mathbf{X} \rightarrow A$	Dangerous event $\rightarrow$ P = low, A = high, D = low $\rightarrow$ fear
Componential	1. $\mathbf{E} \rightarrow \mathbf{X}$ $\mathbf{X}$: n-tuples of appraisal variable values 2. $\mathbf{X} \rightarrow \mathbf{A}$	Dangerous event $\rightarrow$ novelty = high, valence = low, goal rel. = high, ... coping = low $\rightarrow$ fear

2. the Simulation-Theory (ST), where the same mental-state attributer arrives at a mental attribution by simulating, replicating, or reproducing in his own mind the same state as the targets, or by attempting to do so.

For instance, as a whole, Damasio's theory of emotions (Damasio, 2005, 1999, 2012) takes an ST stance.

In our specific case, we are considering the task of attributing emotion states to others based on their facial expressions. This task is different from those usually studied in the mindreading literature, in part because the attributed mental states differ from the usual ones. The vast majority of the literature is devoted to propositional attitudes such as desires and beliefs, almost entirely ignoring emotion states like fear, anger, disgust, or happiness. There is no good reason to exclude these mental states, which are routinely attributed to others in daily life. At the same time, it cannot be assumed that the style of mindreading in this subdomain is the same as the style that characterises other subdomains.

Let us further clarify and expound the two basic theoretical positions towards mindreading, which have loomed large in the literature (but see Goldman and Sripada, 2005 for an in-depth review and discussion). There are numerous ways of developing the TT approach, but the main idea is that the mindreader selects a mental state for attribution to a target based on inference from other information about the target. According to one popular version of TT, such an inference is guided by folk psychological generalizations concerning relationships or transitions between psychological states and/or behaviour of the target (Goldman and Sripada, 2005). The fundamental feature of TT is that it is an information-based approach. It says that attributors engage in mindreading by deploying folk psychological information. What they don't do, as a means to reading a target's mental state, is (try to) model or instantiate the very mental process that the target herself undergoes. The core idea of ST is that the attributor selects a mental state for attribution after reproducing or "enacting" within himself the very state in

question, or a relevantly similar one. In other words, the attributor tries to replicate a target's mental state by undergoing the same or a similar mental process to one the target undergoes. If, in his own case, the process yields mental state M as an output, she attributes M to the target. For example, if the attributor wants to attribute a future decision to a target, he might try to replicate the target's decision-making process in his own mind and use the output of this process as the decision to assign to the target. Alternatively, the attributor may test a hypothesised state by simulating it in his own mind and seeing whether its upshots match those of the target. In either scenario, the attributor must recognise his own state as being of type M in order to select M as the state type occupied by the target. This presumably requires some sort of 'information' about states of type M, so simulation isn't entirely information-free (Goldman and Sripada, 2005). However, in contrast to TT, ST says that the relevant information about M is applied to something like a token or facsimile of a mental state in attributor's own mind, not simply to information about the target from which he infers that the target instantiates M. There is, of course, much more to be said about the TT/ST contrast, but these points should suffice for present purposes and will be further discussed in the next chapter.

Computational models of affect

Computational modelling of emotion refers to attempts to develop and validate computational models of human emotion mechanisms (Reisenzein et al., 2013).

The early studies of emotions have always been a prerogative of philosophers, psychologists and neurobiologist. Indeed, although modern research on emotion dates back to the nineteenth century with Charles Darwin, the adoption of computer technologies in such field is much more recent. In 1984, the American sociologist, psychologist and technologist, Sherry Turkle, was the point of contact between the two research areas. In her book *The Second Self: Computers and the Human Spirit* (Turkle, 1984), she started asking whether computers may have emotions and how our emotional response to robots could lead to metaphysical and ethical issues. This has led the way to a new field of research, formalised by Rosalind Picard (1997), which takes the name of *Affective Computing* (AC).

Picard's definition of the field was quite broad:

> I call "affective computing," computing that relates to, arises from, or influences emotions.

In such perspective artificial agents might have the ability to

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

the latter point being the hard stuff.

Here, we present a summary of the main computational models adopted in Affective Computing. These will mostly deal with Picard's first two points. Nevertheless, the

issue of modelling robots "having emotion" is starting to gain currency in the robotic research programs.

It goes without saying, that the last two decades have seen a proliferation of interdisciplinary works of competing and complementary computational models contributing to a situation of potential confusion, worsened by the lack of a common lexicon between the psychological and computational worlds they derive from. As result, the grouping of such models is not a simple task.

One straightforward possibility could be to categorise them with respect to the considered *modalities*, such as: face, voice, text and physiological data (Calvo and D'Mello, 2010). Such engineering-oriented view has some practical merit. However, in a broader view it is more convenient to discuss such approaches in the framework of the theoretical dimensions previously introduced (Hudlicka, 2011; Dubois and Adolphs, 2015).

In the following, we will distinguish between three slightly different field-dependent perspectives or research programs, namely, machine learning-based approaches, robotic approaches, and AI-based approaches (cfr. Figure 2.1).

Machine learning models

Most current research in affective computing deals with the recognition and the expression of emotions, with machine learning-based affect detection playing a prominent role (the reader can refer to Calvo and D'Mello, 2010, for a broad survey). Somehow, the field of affective computing has by and large narrowed its original broader view within this perspective. As such, it can be considered as a sub-field of the recent and growing field of social signal processing (SSP, see Vinciarelli et al., 2012 for a review).

SSP, in its most general form is an automatic approach for the analysis of social signals which includes several steps. The first is data capture, performed in various settings and using different equipments (from simple laptop webcams to smart meeting rooms and wearable devices). The process of data capture results in signals (audio, video, etc.) that portray more than one person. This makes it necessary to perform person detection, that involves technologies like face detection, speaker segmentation, tracking, etc. The data segments isolated during person detection carry information about the behaviour of each interactant and it is from them that nonverbal behavioural cues are extracted. This is the third step of the process (behavioural cues extraction) and requires technologies like facial expression analysis, prosody extraction, gesture and posture recognition. At the end of the process, the automatically extracted behavioural cues are used in the last step (social interaction interpretation) to infer social signals.

Engineers and computer scientists typically adopt machine learning approaches for automatic emotion classification. These rely on the analysis of video, audio, text and physiological data bringing the problem back to a classical pattern recognition "pipeline": namely, feature extraction/reduction followed by classification (discrete emotion recognition) or regression (continuous affect detection). Indeed, as stated by D'Mello and Kory (2015), "Affect detection is an important pattern recognition problem". This approach is even more evident in the specific case of facial expression analysis as clearly stated in the recent review by Sariyanidi et al. (2015) (see Figure 2.6).

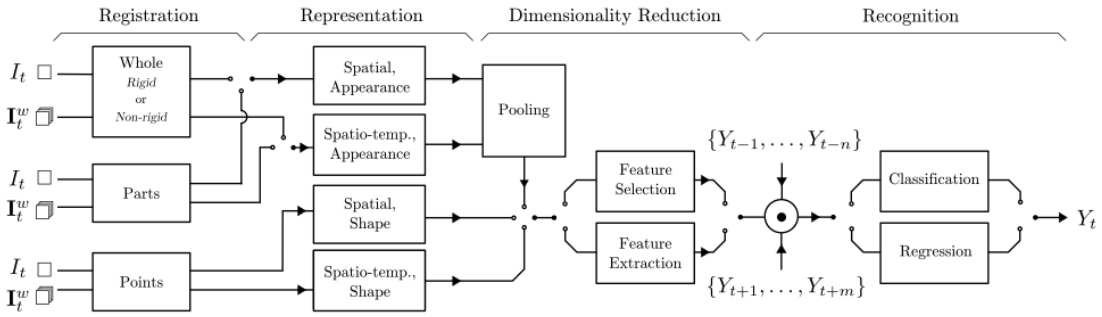


Figure 2.6: *The facial affect detection pipeline. The input is a single image (I_t) for spatial representations or a set of frames (I_t^w) within a temporal window w for spatio-temporal representations. The system output Y_t is discrete if it is obtained through classification or continuous if obtained through regression. The recognition process can incorporate previous ($Y_{t-1}, \cdots Y_{t-n}$) and/or subsequent ($Y_{t+1}, \cdots Y_{t+m}$) system output(s). (Scheme from Sariyanidi et al. (2015)).*

Overall, this approach assumes that an internal affective latent state "causes" external (facial) behaviour. Internal affect state can be either discrete or continuous. Discrete representations have been mostly adopted, especially for facial expression analysis, though continuous approaches are gaining currency as reviewed by Gunes and Schuller (2013). Also it basically relies on a Theory-Theory perspective and internal simulation is always neglected.

Adopting the discrete/dimensional/componential taxonomy previously introduced, Figure 2.7 shows at a glance how classic ML algorithms have been exploited in computational models of emotions.

By and large, machine learning approaches cover a wide range of different algorithms, and include the processing of single or multiple modalities combined together.

To provide some examples, in the *discrete emotion* perspective, McDaniel et al. (2007) adopted discriminant analysis to classify, in terms of five affective states, the facial expressions of people recorded during a tutoring session in a learning environment. Devillers and Vidrascu (2006), instead, focused on the paralinguistic features of spoken dialogues from a medical emergency call centre. In particular, they used a Support Vector Machine to classify prosodic and lexical cues against four discrete categories. More recently, AlZoubi et al. (2009) presented a performance comparison of Nearest Neighbour (NN), Support Vector Machines (SVM) and Naive Bayes (NB) on classification of EEG signals for affect recognition. Deriu et al. (2016), instead, proposed a classifier for predicting message-level sentiments of English micro-blog messages from Twitter recurring to a 2-layer convolutional neural networks whose predictions are combined using a random forest (RF) classifier.

Many other works focus, instead, on the *dimensional* perspective. These include Haag et al. (2004), who used a neural network to estimate the valence and arousal values from multiple physiological signals, while Heraz and Frasson (2007) adopted the PAD model to predict from EEG data via a multiple regression (MR) algorithm. A novelty in this field is given by Mollahosseini et al. (2017) who propose a database of 1M facial images annotated in continuous dimensional space, and used two different

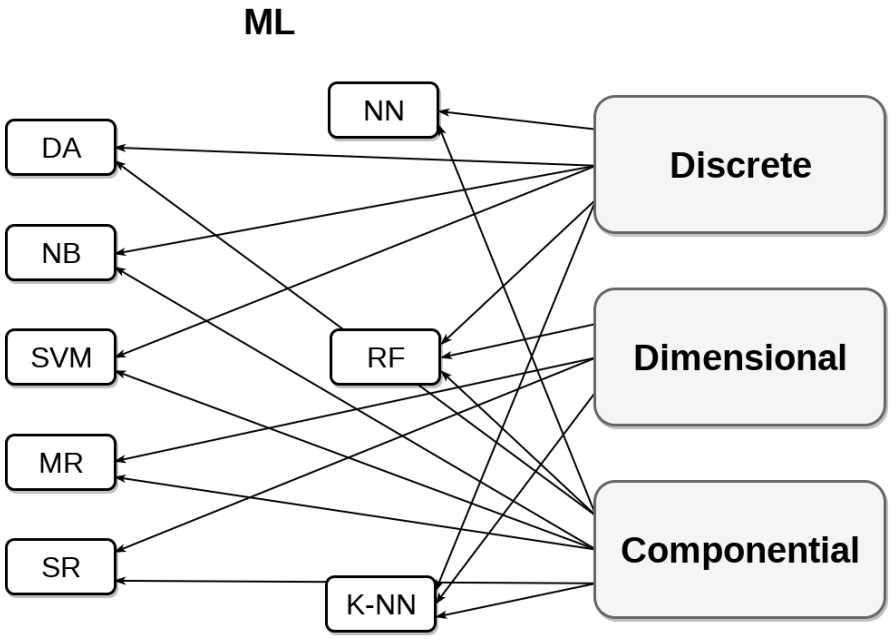


Figure 2.7: *Machine Learning (ML) approaches adopted in computational models of emotions. DA: Discriminant Analysis, k-/NN: k-/Nearest Neighbour, NB: Naive Bayes, SVM: Support Vector Machine, RF: Random Forest, SR/MR: Simple/Multiple Regression.*

DNNs to extract highly discriminative features from the data samples, followed by a simple regressor (SR) for each of the two dimensional values.

Finally, the componential perspective, as shown in Figure 2.7, is usually accounted for via classification or regression techniques. Meuleman and Scherer (2013) proposed a comprehensive comparison of machine learning approaches applied to the appraisal theoretical perspective.

D'Mello and Kory (2015), in their comprehensive and in-depth meta-analysis of 90 multimodal (MM) affect detection systems, reported that the state of the art mainly consists of person-dependent models (62.2% of systems) that fuse audio and visual (55.6%) information to detect acted (52.2%) expressions of basic emotions and simple dimensions of arousal and valence (64.5%) with feature- (38.9%) and decision-level (35.6%) fusion techniques. However, there were also person-independent systems that considered additional modalities to detect nonbasic emotions and complex dimensions using model-level fusion techniques. The meta-analysis revealed that MM systems were consistently (85% of systems) more accurate than their best unimodal counterparts, with an average improvement of 9.83% (median of 6.60%). However, improvements were three times lower when systems were trained on natural (4.59%) versus acted data (12.7%). Importantly, MM accuracy could be accurately predicted (cross-validated R^2 of 0.803) from unimodal accuracies and two system-level factors.

But coming to the specific case study of this thesis, automatic analysis of facial affect has indeed fostered a wealth of techniques. For an in-depth and recent review, the reader should refer to Sariyanidi et al. (2015).

By and large, these techniques share the assumption that understanding affective states from facial expressions can be accomplished through the computer vision and

pattern recognition "pipeline" (Sariyanidi et al., 2015).

The (often concealed) rationale behind such attitude grounds in early inferential/appraisal theories of emotion (Goldman and Sripada, 2005): visual perception of emotional stimuli is followed by cognitive or appraisal/interpretative processing of the stimulus, which in turn triggers affective responses and feelings (Chakrabarti et al., 2006). Spelling out such account for faces, when an expresser is observed displaying a particular facial expression, the observer utilises the rich body of visual features/properties to infer and attribute an emotion state to the expresser (Goldman and Sripada, 2005).

Clearly, one must be aware that the performance of the data-driven pipeline heavily depends on the amount of annotated facial behaviour (Zafeiriou et al., 2016). So far most datasets have been collected in controlled recording conditions and typically under restrictive scenarios; current efforts are thus devoted to collect and annotate spontaneous facial behaviour "in-the-wild", much like it has happened for face detection and facial landmark localisation (Kossaifi et al., 2017). However, in unrestricted environments not only the sampling of the data (Torralba and Efros, 2011), but annotation techniques are far from being evident, specially if continuous dimensional affect analysis is the goal (Kossaifi et al., 2017). Arguably, the introduction of deep learning techniques in the pipeline, in particular deep Convolutional Neural Network, could further improve performance on benchmarks, at least in principle, although this achievement is likely to be attained at the price of addressing even larger, web-scale datasets (Masi et al., 2016).

Clearly, the machine learning research trend does not make justice of the full problem we are addressing here: human facial behaviour processing for affective analysis is a complex problem and, most important a fairly different game from face detection or landmark localisation. Facial behaviour conveyance of affect depends on the context and on the person. Hence, pedantic following of the same research path that led to top performance results on previous problems (e.g, object or face recognition), might be questionable, and surely it is so when one moves to the field of social robotics.

Robotic models

Rapid progress in robotics calls for naturalistic interaction between humans and machines, where the emphasis is on collaboration, learning via imitation and socialising (Billard et al., 2016; Natale et al., 2013). It goes without saying, these are quite different scenarios with respect to the off-line learning / classification over billions of facial pictures. In a sense, problems posed in such realms are better conceived in terms of learning using few labelled examples (Yang et al., 2013; Lake et al., 2015).

For instance, in social robotics, an important challenge is to determine how to design robots that can perceive the user's needs, feelings, and intentions, and adapt to users over a broad range of cognitive abilities. It is conceivable that if robots were able to adequately demonstrate these skills, humans would eventually accept them as social companions. This approach requires understanding how humans interact with each other, how they perform tasks together and how they develop feelings of social connection over time, and using these insights to formulate design principles that make social robots attuned to the workings of the human brain.

Wiese et al. (2017) argue that the likelihood of humanoid robots being perceived as social companions can be increased by designing them in a way that they are perceived as intentional agents that activate areas in the human brain involved in social-cognitive

processing, and provide an in-depth and wide review of how neuroscientific methods can contribute to make robots appear more social. Robots that are supposed to act as social interaction partners in the future need to fit in human-attuned environments by emulating human form and cognition. Indeed, psychological research has shown that anthropomorphism, and specifically mind perception, are highly automatic processes that activate social areas in the human brain (Wiese et al., 2017).

There is a tradition of robotic research that utilised neuroscience studies as a starting point (Kawato, 1999; Scassellati, 2002; Demiris et al., 2014). These approaches led to accurate models of muscular-skeletal systems, and in particular of facial features appearance (Oh et al., 2006; Becker-Asano et al., 2010)

However, when talking about humanoid robot interaction, their social appearance is concerned with both the "bodyware" or hardware of the machine, and the behaviour concerning the observable results of the workings of its "mindware" or software (Wiese et al., 2017). As to mindware, an important distinction needs to be made between neurally accurate models, often proof of principles, and actual working implementations on real hardware, with profound differences between computers and human brains, impeding accurate real-time neural simulations of large brain systems, such as those of the social brain. Given the technological limitations associated with trying to reproduce large brain networks on actual bodyware, the goal needs to be the identification of a minimal set of features that can reliably trigger mind perception in non-human agents.

First attempts to build socially competent robots can be traced back to the work done at MIT Brook's robotics lab, e.g., Cog (Brooks et al., 1999) and Kismet (Breazeal and Scassellati, 1999; Breazeal, 2003b). With Kismet, Breazeal and Scassellati (1999) studied how an expressive robot elicited appropriate social responses in humans by displaying attention and turn-taking mechanisms. They also identify some of the requirements of the visual system of such robots (Breazeal et al., 2001) as for example the advantages of foveated vision, eye contact, and a number of sensorimotor control loops (e.g., avoid and seek objects and people).

Interestingly, cognitive/social visual behaviour grounds in a motivation system which consists of drives and emotions. The robot's drives represent the basic needs of the robot: to interact with people (the social drive); to be stimulated by toys and other objects (the stimulation drive); to rest (the fatigue drive). For each drive, there is a desired operation point, and an acceptable bound of operations around that point (the homeostatic regime). Unattended, drives drift toward an under-stimulated regime. Excessive stimulation (too many stimuli or stimuli moving too quickly) push a drive toward an over-stimulated regime. When the intensity level of the drive leaves the homeostatic regime, the robot becomes motivated to act in ways that will restore the drives to the homeostatic regime.

The robot's emotions, in turn, are a result of its affective state. The affective state of the robot is represented as a point along three dimensions: arousal (i.e. high, neutral, or low), valence (i.e. positive, neutral, or negative), and stance (i.e. open, neutral, or closed). This core affective state is thus continuous and is based on Jim Russel's core affect conjecture (Russell, 2003). Operatively, the affective state is computed by summing contributions from the drives and behaviours. Percepts may also indirectly contribute to the affective state through the releasing mechanisms. Each releasing mechanism has an associated somatic marker processes, which assigns arousal, valence and

stance tags to each releasing mechanism (a technique inspired by Bechara et al., 2000; Bechara and Damasio, 2005). At the same time, this continuous state is partitioned into a discrete set of emotion regions, which roughly correspond to Ekman's discrete emotions (Ekman, 1993).

In subsequent work Breazeal et al. (2006) have presented Leonardo, a robot that can imitate humans' facial expressions. They have shown how it is possible for the robot to bootstrap from this imitative ability to infer the affective reaction of the human with whom it interacts and then use this affective assessment to guide its subsequent behaviour. Their approach is heavily influenced by the ways human infants learn to communicate with their caregivers and come to understand the actions and expressive behaviour of others in intentional and motivational terms. Specifically, the approach is guided by the hypothesis that imitative interactions between infant and caregiver, starting with facial mimicry, are a significant stepping-stone to develop appropriate social behaviour, to predict other's actions, and ultimately to understand people as social beings. Leonardo learns the direct mapping between a person's facial expression and its expression by using a neural network. Scassellati (2002) went further and took some first steps, in a Theory-Theory vein, toward implementing a theory of mind for the robot Cog based on an established psychological model for mentalising developed by Baron-Cohen (1997).

Since then, several empathic robots that consider the internal state of others for their own expressions have been proposed. Hegel et al. (2006) presented an anthropomorphic robot called BARTHOC capable of recognising humans' emotion from speech and producing facial expressions corresponding to the six basic emotion. Though, the issue of mimicry was addressed, emotion recognition from speech was classically performed in the pattern recognition pipeline, i.e. feature extraction and classification. This was performed with simple methods, and directly mapped to facial expressions, in order to meet real-time recognition.

Adams and Robinson (2011) developed an android head robot that mimicked the facial expressions of humans with the aim of social-emotional intervention for autistic children. Their robot tracked facial feature points of subjects who expressed emotional states and directly converted them into corresponding control points to modify its own facial expression (direct mapping). Trovato et al. (2013) developed an emotional model for a humanoid robot, KOBIAN, based on psychological studies. Their model represented KOBIAN's internal state, which is modulated by external stimuli. It also had prototypes of facial expressions grounded on specific emotional states and, via a form of direct mapping, expressed facial patterns as combinations of these prototypes according to Plutchik (2003).

Following the discovery of mirror neurons in non-human primates and their involvement in action understanding (see, Rizzolatti and Sinigaglia, 2016 for a general introduction), neuroscientifically inspired approaches to robotics mainly focused on developing models for action recognition and imitation (Metta et al., 2006; Oztop et al., 2013). As we will discuss in the next chapter, the mirror neuron system activates both during the execution of their own actions and while observing the same actions performed by others. In the context of emotional communication, this mechanism is assumed to enable people to imagine the emotional state of others based on their own experiences of expressing the corresponding emotion.

The key concept of shared sensorimotor representations, dating back to Liberman and Mattingly (1985), guided a variety of implementations utilising, for example, recurrent neural networks (Tani et al., 2004) or various other machine-learning methods that learn direct-inverse models from examples (but see, for a review, Oztop et al., 2006, 2013). Among these attempts to implement a mirror neuron system into artificial agents, some models were more neuroscientifically accurate than others.

Inspired by the mirror neuron systems, Lim and Okuno (2014) proposed multimodal emotional intelligence (MEI), which utilises an integrated architecture to recognise the emotional states of others and generate its own emotional facial expressions. The MEI is composed of Gaussian mixture models (GMMs) to realize both recognition and generation in the same architecture. Recognition of the emotional states of others is represented as classification of input features by GMMs, whereas expressions of one's own emotional states are achieved by sampling features from selected Gaussian distributions corresponding to the specific state. An important characteristic of their model is that it computes four features (Speed, Intensity, irRegulation, and Extent, SIRE) assumed as very common among modalities. Therefore, following training from speech using the SIRE, the MEI was able to estimate categories of emotion from not only audio signals but also gait signals. It can also generate the SIRE for the robot's voice, gait, and gesture. However, there are two limitations in their system: first, the SIRE is a heuristic feature defined by a designer; each modality may include specific features to represent emotion. Second, although they considered MNS (Lim and Okuno, 2015), they did not examine the role of mental simulation for estimation of the other's emotion.

Brain-inspired models have dominated the field for several years, but are being replaced by the modern "brute force" data-driven approach of using deep networks, in particular, convolutional neural networks (Goodfellow et al., 2016) and managing the increased computational cost through specialised processors (e.g., GPUs), resulting in an improvement in performance of orders of magnitude.

Yet, there are still efforts to reconcile these two perspectives. Barros and Wermter (2016) have recently proposed a model that simulates the innate perception of audio-visual emotion expressions with deep neural networks, that learns new expressions (via a convolutional neural network) by categorising them into emotional clusters with a self-organizing layer. This process implements the emotion perception stage where the agent -the robot NICO, Neuro-Inspired COmpanion - observes the environment consisting of human and other artificial agents expressing a particular emotion, for example, a human smiling at the agent (Churamani et al., 2017). Then, in the emotion synthesis step, the robot factors in its own goals and beliefs to estimate an emotional state for itself; this is based on the inference engine of the agent so as to react to the perceived input from the environment. Eventually, once the agent has received an input from the environment, it then expresses its emotional state in the form of facial gestures, synthetic speech etc. evoking another response from the environment (Churamani et al., 2017). Clearly, here there is a sort of direct mapping (no internal simulation) from the latent space of categorising in discrete form, affective expressions learnt in a bottom-up, feed-forward sweep.

Kim et al. (2013) have rectified their MEI/SIR original proposal using deep neural networks that learn to extract features for emotional categorisation from audio-visual signals. In their system, deep belief networks (DBNs) comprising restricted Boltzmann

machines (RBMs, see Goodfellow et al., 2016) were used as unsupervised learning mechanisms. The RBM can abstract input signals and reconstruct the signals therefrom. In experiments, their model extracted emotion specific features from general ones, which are not always important for the classification of emotion.

More recently, Horii et al. (2016) proposes a model that can estimate human emotion and generate its own emotional expressions to imitate the human expressions based on the estimation of his/her emotion in human-robot interaction. The model overcomes two issues confronting the previous emotional model: constructing an emotional representation of multimodal signals for estimation and generation for emotion instead of using heuristic features, and actualising mental simulation to infer the emotion of others from their ambiguous multimodal signals. In the same vein of Kim et al. (2013), they employed RBMs to address these two issues as they are able to abstract input signals and recall the signals there from. The abstraction capability of RBMs was exploited to overcome the first limitation by reducing the dimensions of multimodal signals and associating the multimodal signals. The model also carries out mental simulation by exploiting the ability to generate sensorimotor signals. The mental simulation mechanism enables the model to estimate the emotional states of others from partial multimodal expressions based on its own experiences. Indeed, their proposal, as we will discuss further on, is the most related to the work we present in this thesis.

We can conclude the robotic overview by noting that the construction of empathy in robots is still a long way to go. The main problems and challenges have been lucidly dissected by Asada (2015).

AI-oriented models

The aim of AI-oriented models of emotion is to the enhance the human-computer interaction, by controlling the behaviour of virtual agents to motivate and establish empathy and bonding.

Appraisal and componential theories have been largely influential in this research program. Computational models of emotion had early success exemplified by Scherer's GENESE expert system. GENESE was built on a knowledge base that mapped appraisals to different emotions as predicted by Scherer's heuristics and theory (Scherer, 1993). However, despite its value as a tool to test emotion theories, the system was too limited for most real-world applications. It is too difficult, if not impossible, to construct the significant knowledge base needed for complex realistic situations. More recently, the OCC model by Ortony et al. (1990) has been widely adopted. For instance, it has been used by Conati (2002) in her remarkable work on emotions during learning. The model is used to combine evidence from situational appraisals that produce emotions (top-down) with bodily expressions associated with emotion expression (bottom-up). In this fashion, both a predictive (from the OCC model) and a confirmatory (from bodily measures) analysis of emotional experience is feasible.

In general, the emotional displays can be divided in two categories: the visible, behavioural and expressive manifestations, and the less visible, effects on the internal perceptual and cognitive processes. Therefore, a comprehensive model of emotion should account for multiple modalities including the effects on non-observable physiological components, visible physical manifestations and internal cognition. Figure 2.8 summarises a few well-known AI-oriented models with respect to the discrete/dimen-

sional/componential taxonomy.

In Bui (2004) it is presented a computational implementation of emotions for an embodied agent, named ParleE. Such model maps an emotion state vector expressed in terms of six basic emotion labels to a contraction of 19 different facial muscles. The specular implementation is represented by WASABI (Becker-Asano, 2008) which maps regions of dimensional core affect into one of seven possible facial expressions, and is implemented into the physical social robot EMYS. An hybrid approach resides in ALMA (Gebhard, 2005) which represents, via an agent named 'VirtualHuman', both discrete emotional states and a general mood expressed in a three-dimensional (PAD) space. Other works adopt the componential derivation which identifies emotion labels as the combination of several appraisal variables. This is the case of EMA (EMotion and Adaption) (Marsella and Gratch, 2009) that implements checks posited by appraisal theories based on a representation of the relationship between the agent and the environment. Under the same category relies the computational model behind the 'Prime Climb' educational game. Conati (2002) uses a probabilistic model, namely a Dynamic Decision Network, in a dualistic way: top-down, as a predictive assessment of the student traits, and bottom-up, as diagnostic assessment of bodily expressions associated with emotional states.

A lucid and in-depth review of the relationships between computational modelling of emotion and the general goals of AI , when these are restricted to the domain of emotions, has been provided by Reisenzein et al. (2013)

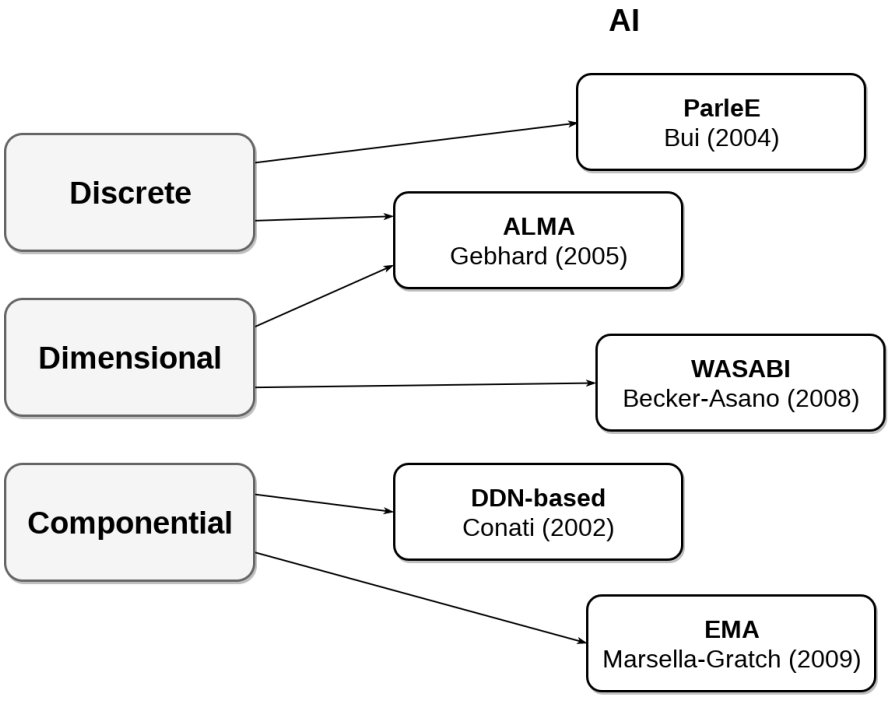


Figure 2.8: *Artificial Intelligence-oriented (AI) approaches adopted in computational models of emotions.*

Summary

We have reviewed the main contributions to the field of computational models of affect, with some particular reference to affective facial expressions. Among the plethora of approaches and the different research programs, we also tried to identify some general trends or dimension to more systematically frame the discussion. However, we cannot elude the fact that on the one hand, yet, it is difficult to compare approaches that, apparently seem to be driven by different levels of explanation (neural, behavioural, etc.). On the other hand we have used the terms "theory" and "model" in a sort of a loose definition, at least in a strict epistemological standpoint. Clearly, this is a cogent issue when addressing the computational modelling of affect, namely via facial expression analysis; thus the next chapter will provide first, a thorough and in-depth analysis of this aspect.

CHAPTER 3

Rationales and working hypotheses

In the previous chapter we have reviewed different contributions to the conundrum of affective modelling. By and large we have concluded that, in many cases, the plethora of approaches are hardly comparable with one another. Indeed there have been cases in which comparisons have fostered academical disputes that, in the end, were epistemologically undermined by the fact that different levels of explanations had been actually confronted.

In a subject such as this, it is perhaps best to start by establishing models of an apt generality, so to avoid *ad hoc* heuristics, while considering relevant and yet well defined case studies, in order not to complicate an already difficult problem. As we stated at the beginning, in this study we are addressing the construction of a multimodal affective space but focussing on the case of the perception of emotional facial expressions along a dyadic interaction. Thus, in this chapter we first make clear the methodological framework we have adopted (Section 3.1). This can be characterised as a multilevel analysis framework, which aims at devising a theoretical model but informed and constrained by knowledge that we have available both at the psychological and at the neuroscience explanation levels.

Next, we provide the necessary neurobiological underpinning (Section 3.2) to the subsequent modelling steps, while accounting for most recent results concerning motor-based views of affect unfolding (Section 3.3). Eventually (Section 3.4), the distributed neural architecture devised at this stage will be used as a blueprint for outlining, at a more general and abstract level, a functional architecture to support the theoretical model that will be formalised in Chapter 4.

A methodological foreword

Computational models in the cognitive and behavioural sciences can be used either as analytical tools for analysing empirical data or as instantiations of cognitive hypotheses (Palminteri et al., 2017). The work described in this thesis falls in the second case. Then, it is important to note that, as instantiations of cognitive theories, computational models can target different levels of description.

A key distinction (Palminteri et al., 2017) is that between aggregate versus mechanistic models: *aggregate models* describe average behaviours using a synthetic mathematical model; *mechanistic models* explain how behaviours are generated.

Such distinction has been further developed by Marr (1982), who proposed three levels of description/explanation (see Fig. 3.1, left):

1. the *what/why* level (computational theory, i.e. the individuation of a computable function as a model of a given behavioural phenomenon),

2. the *how* level (algorithm),

3. the *physical realisation* level (implementation).

Marr's multilevel approach has become a sort of paradigm in research work on the theoretical foundations of cognitive science (Dennett, 1987), while nourishing a vast philosophical debate. But more importantly for us, Marr's account can be seen, from a broader perspective, as the claim that the behaviour of a complex system, such as a living organism, has to be explained at various levels of organization, including psychological, neurological, cellular and biochemical levels.

Anderson (1991) remarkably synthesised the advantage of the distinction between the computational and the algorithmic levels in particular:

> The search for scientific explanation is easier in this approach. In a mechanistic approach, we must consider any combination of mechanisms as basically equivalent to any other, and this creates an enormous search space of possible mechanisms with no heuristics for searching it for an explanation [...] There is a sense in which rational explanations are more satisfying than mechanistic explanations. A mechanistic explanation treats the configuration of mechanisms as arbitrary. The justification for the mechanisms is that they fit the facts at hand. There is no explanation for why they have the form they do rather than an alternative form. In contrast, a rational explanation tells why the mind does what it does (p. 410)

A critical point here concerns the constraints assumed by computational theory (Anderson's"rational explanation") when aiming to reduce the underdetermination and the arbitrariness or ad-hocness or mimicry, as Marr put it, of cognitive models at the algorithmic level (Anderson's "mechanistic explanation"). In particular, the "heuristics for searching an explanation" can be seen as a guide in the choice of a cognitively plausible mechanism, given that there are usually many mechanisms instantiating the same performance. Briefly, this is the model underdetermination problem, which although being a general problem in scientific explanation, has proven to be remarkably acute

for cognitive explanation. In classical Cognitive Science, this issue was deeply discussed in depth by Pylyshyn (1984) in terms of the specific constraints the cognitive scientist has to assume in order to guarantee the "psychological reality" or plausibility of computational models. This issue is also dealt with by embodied cognitive science too, this time introducing plausible constraints stemming from the environment and the body. Further, this seems to be an issue raised at the time of Cybernetics and early Artificial Intelligence (AI); all these topics have been pointed out by recent analyses: see Cordeschi (2002).

Indeed, a hallmark of the present state of research in Cognitive Science, is that one is generally ignorant of how exactly to cast the different levels into a grounded relationship, and any proposal has its limitations (Boccignone and Cordeschi, 2015). Marr himself contended with a persistent ambiguity in the role of the implementation level with respect to the algorithmic one. On the one hand, the implementation level was hypothesised as a rather independent level of explanation, never constraining the algorithmic level from the bottom up. On the other hand, it has occasionally been endowed with the role of arbitrating the selection of the most suitable algorithm, from among those that consistently embodied constraints imposed by the computational level (see Marr, 1982 , Chapter 3). For in such a case, an algorithm is preferred by virtue of its apparent greater biological or neurological (thus, implementation level) plausibility.

However, within the Bayesian approach issues about constraints can be settled in a way quite different from Marr's, particularly in relation to his three-fold hierarchy of levels of explanation.

In the light of the growing exploitation of Bayesian methods in the cognitive sciences, it has been argued (Chater et al., 2006; Knill et al., 1996; Boccignone and Cordeschi, 2007) that Marr's three-fold hierarchy could be reorganised into two levels: the *computational theory* level, which can be formalised precisely in terms of Bayesian theory, and the *implementation theory* level, embedding both Marr's algorithmic and physical realisation levels (see Fig. 3.1, right).

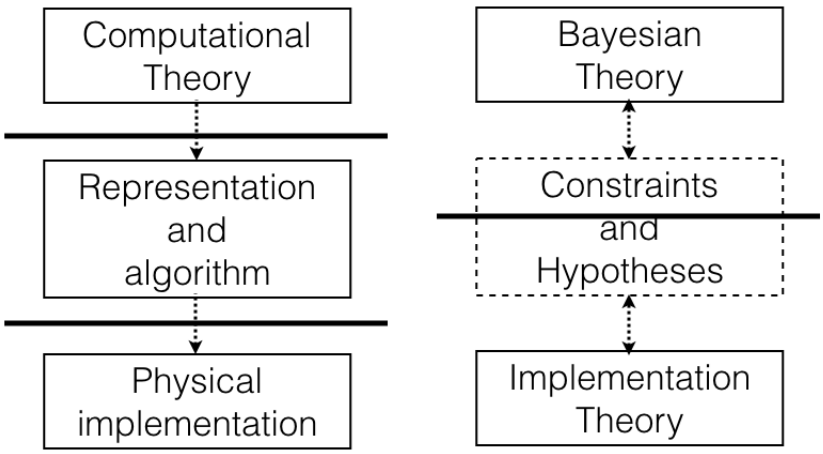


Figure 3.1: *The levels of explanation in cognitive/behavioural sciences. Left: Marr's original proposal (Marr, 1982). Right: Marr's revision according to Yuille and Kersten (adapted from Knill et al., 1996).*

Note that both levels are denoted "theories" here and, differently from Marr, a close interaction between the computational (here Bayesian) theory and the implementation theory level is assumed. Further, hypotheses and constraints are somehow shared between the two levels (see broken-line box in Fig. 3.1).

In this thesis we do endorse this two-level view. Also, we use the term "model" to qualify the two theory levels in Fig. 3.1: briefly, in what follows we will refer to such levels as the *theoretical model*[1] and the *implementation model*. This well reflects the fact that, at both levels, the cognitive scientist is devising models embodying constraints related to a number of physical or biological laws and theoretical hypotheses that are relevant to the explanation of a given phenomenon. This is a point regarding both models explaining behavioural regularities (at the Bayesian theory level) and models explaining neural regularities (at the implementation theory level).

The theoretical model

The computational theory level is the highest level of a cognitive theory. This is a functional specification of cognition as "a mapping from one kind of information to another" where "the abstract properties of this mapping are defined precisely (Marr, 1982). The details of how this mapping is implemented are left to lower levels. He gave one example of a high-level theory from mathematics. The field axioms specify the abstract properties of algebraic expressions, such as the commutativity of addition, but are silent on low-level matters of implementation, such as how numbers are represented (Roman numerals, base-10, base-2, etc.).

Marr's computational theory can be specified in the Bayesian framework in terms of the so called generative model: namely, the joint probability distribution $P(\{X_k\}_{k=1}^K)$ of the random variables (RVs) of interest $X_1, \cdots, X_K$ factorised according to given constraints. A representation of such generative model can be given in terms of a Probabilistic Graphical Model (PGM) (Lauritzen, 1996; Jordan, 1998; Koller and Friedman, 2009), say $\mathcal{G}$ (see the appendix A for an introduction to PGM).

The graph $\mathcal{G}$ can be viewed in two very different ways:

- as a compact representation for a set of conditional independence assumptions about a distribution;
- as a data structure that provides the skeleton for representing a joint distribution compactly in a factorized way.

Example 3.1. *Let the random vectors $\mathbf{X}_{\mathcal{I}}^{hidden}$ and $\mathbf{Y}_{\mathcal{I}}^{obs}$ denote the hidden affective state and the observable facial expression of an agent $\mathcal{I}$, respectively. Denote RVs $\mathbf{E}$ and $\mathbf{C}$ an emotional event and the context or circumstances in which the event occurs, respectively.*

Define $P(\mathbf{Y}_{\mathcal{I}}^{obs}, \mathbf{X}_{\mathcal{I}}^{hidden}, \mathbf{E}, \mathbf{C})$ the joint probability distribution over the RVs of interest. By applying the product rule, one possible factorisation of the joint distribution

[1] In a sense, our use of the term theoretical model is close to that of the philosopher Ronald Giere, who reserved the term "for a special class of abstract models, those constructed with the use of [...] theoretical principles": Newton's laws, Mendel's or Darwin's are different examples of such principles (Giere, 1999).

is

$$P(\mathbf{Y}_{\mathcal{I}}^{obs}, \mathbf{X}_{\mathcal{I}}^{hidden}, \mathbf{E}, \mathbf{C}) = P(\mathbf{Y}_{\mathcal{I}}^{obs} \mid \mathbf{X}_{\mathcal{I}}^{hidden}, \mathbf{E}, \mathbf{C}) P(\mathbf{X}_{\mathcal{I}}^{hidden} \mid \mathbf{E}, \mathbf{C}) P(\mathbf{E} \mid \mathbf{C}) P(\mathbf{C})$$

However, on the basis of psychological theories (e.g., Ekman and Rosenberg (1997)) we may assume that the affective state $\mathbf{X}_{\mathcal{I}}^{hidden}$ will be determined by both event $\mathbf{E}$ and context $\mathbf{C}$. In turn, the internal affective state $\mathbf{X}_{\mathcal{I}}^{hidden}$ will be responsible for the generation of the visible expression $\mathbf{Y}_{\mathcal{I}}^{obs}$. Such constraints can be encoded in the PGM shown in Figure 3.2. The PGM is a DAG, thus Eq. A.1 holds so that the joint pdf factorisation above simplifies to

$$P(\mathbf{Y}_{\mathcal{I}}^{obs}, \mathbf{X}_{\mathcal{I}}^{hidden}, \mathbf{E}, \mathbf{C}) = P(\mathbf{Y}_{\mathcal{I}}^{obs} \mid \mathbf{X}_{\mathcal{I}}^{hidden}) P(\mathbf{X}_{\mathcal{I}}^{hidden} \mid \mathbf{E}, \mathbf{C}) P(\mathbf{E}) P(\mathbf{C}).$$

From the graphical representation $\mathcal{G}$, it is straightforward to read the fundamental conditional independence assumption:

$$(\mathbf{Y}_{\mathcal{I}}^{obs} \perp \{\mathbf{E}, \mathbf{C}\} \mid \mathbf{X}_{\mathcal{I}}^{hidden})$$

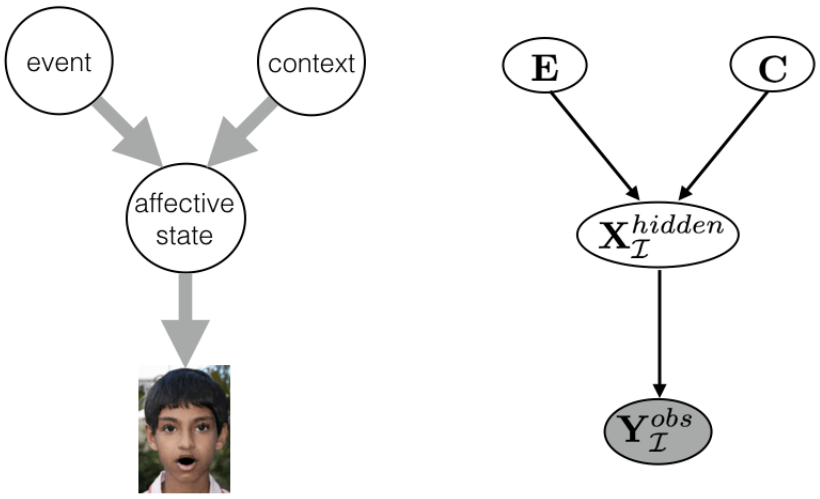


Figure 3.2: *Capturing psychological theories in a PGM. Left: a schematic representation of the assumption that event and context, together, trigger agent's internal emotional state, which is then externally displayed via facial expression. Right: the probabilistic directed acyclic graph (Bayesian network) where nodes represent the RVs of interest and arrows encode conditional dependencies between nodes.*

Note that, in the example provided above, the constraints to shape $\mathcal{G}$ architecture have been "top-down" derived by taking stock of common assumptions in the psychological literature. Nevertheless, a theoretical model related to the behavioural level, as far as it can be identified in terms of the underlying neural architecture, can be "bottom-up" constrained by the latter, thus mirroring the organization of groups of neurons or of functional brain areas (depending on the grain of the analyses).

Subtleties of the implementation model

The most straightforward implementation model to "put into work" the theoretical model can be obtained by specifying the probability distributions defining the conditional dependencies and by applying suitable PGM-based algorithms such as Belief Propagation, Variational Bayes learning, etc. (but see Koller and Friedman (2009) for an in-depth introduction). This can be seen as the coarsest-grain implementation model. But it might be the case that an implementation model at a finer grain is needed to be addressed.

Probabilistic models are also compositional in nature, a lower implementation level can be devised by designing inference as a collection of local inference problems, defined over sub-graphs of graph $\mathcal{G}$. This indeed is the route we will follow to devise our implementation model.

Clearly, there is a multiplicity of finer analysis levels downwards to the ultimate neural level. If a neural grain of analysis is pursued, then it has been shown that the PGM can be used as a blueprint for devising a neural implementation model (neural architecture) and simulation can be performed at that level.

This raises the fundamental issue of what should be then considered as the neural (implementation) level in cognitive science modelling. Clearly, paraphrasing Wiener and Rosenblueth, the best material model of a brain is another, or preferably the same, brain. Thus, in the end, if this ultimate level is addressed a rational / computational theory explanation should confront with experimental data at this level (neurophysiological, fMRI, etc.).

However, in the wild of the neural jungle, the gap between levels of explanations can turn to be huge and a variety of sub-levels can be derived downward the hierarchy. In a very elegant work (Abbott and Kepler, 1990), Abbott and Kepler have mathematically derived from the Hogkin-Huxley model, via subsequent reductions and approximations, the FitzHugh-Nagumo model, the integrate-and-fire model, and eventually the binary Hopfield-type model. If one assumes *tout court* any of this "neural implementations" as a proxy for the brain, one then must be aware of its explanatory limitations (which could be enough, depending on the goal of the researcher). Note that levels can be even explored further downward: Angela and Dayan (2005) have proposed that the neuro-modulators acetylcholine and norepinephrine play a major role in the brain's implementation of Bayesian priors at the cognitive/behavioural level.

Thus, dealing with the implementation theory level, if neural simulation is addressed we must be ready to deal with a multiplicity of (sub) levels. Computations can thus be carried out using classic artificial neurons, or at a lower level by using membrane potential as the crucial variable, or further down, at a chemical level, by taking into account concentrations of calcium or other substances governed by reaction-diffusion equations. As pointed out by Koch,

> [...] the principal differences are the relevant spatial and temporal scales dictated by the different physical parameters, as well as the dynamical range of the [...] sets of parameters (Koch (1999), p. 279).

Putting all together: multilevel analysis

Multilevel analysis is a consistent way of dealing with the multiscale nature of cognitive and behavioural processes. Behavioural and cognitive phenomena, and markedly emotions, exists at multiple temporal and spatial scales.

The kind of explanatory pluralism that is involved by a Bayesian account of Marr's multilevel analysis, affords the scientist a method for developing fuller explanations of relevant phenomena. To sum up the main features discussed above:

1. the notion of architecture becomes a central issue, since it embodies constraints assumed by the cognitive scientist for his own purpose at the chosen level of explanation;

2. the implementation level turns out to be a lower-level model, which is suitable to be used for instantiating the computational theory level at different sub-levels;

3. Marr's algorithmic level does not so far provide an autonomous level of explanation, rather one encompassing simulations of different grains: from a coarse-grained simulation of Bayesian inference and learning processes close to the behavioural/computational theory level, down to fine-grained simulation(s) at the neural level.

Top-down constraining affords the cognitive scientist a basis for unifying multiple levels of analysis by identifying longer-scaled levels as contextual constraints for the smaller-scaled levels. Bottom-up scaffolding provides a framework for identifying what can emerge from lower-level patterns (i.e., patterns existing at shorter time scales or smaller spatial scales), and the dynamics and processes by which these patterns are formed. It is the substrates of lower levels that allow higher-level phenomena to emerge.

As just mentioned, this is relevant in the study of affect, where underlying processes compose in a complex way, to produce behavioural dynamics where, as lucidly remarked by Pessoa (2008), no longer emotion and cognition can be eventually distinguished (see, Figure 2.3, Section 2).

As said, this is the route we are taking in the following where we will first exploit the constraints and the functional architecture outlined here to devise, in Chapter 4 a theoretical model in terms of a dynamic PGM and subsequently, at a lower level, a compositional implementation model.

Neurobiological background

To set out, we need to define explicitly the term "emotion", as it is used throughout this dissertation. This is a controversial issue at large (see Ziemke and Lowe (2009) for a thorough discussion). Here we assume Damasio's cleavage between emotions and feelings, the latter being the first person experience of the corresponding emotion (Damasio, 1999). The term emotion should be rightfully used to designate a collection of responses triggered from parts of the brain to the body, and from parts of the brain to other parts of the brain, using both neural and humoral routes. The end result of the collection of such responses is an emotional state, defined by changes within the *body-proper*, e.g., viscera, internal milieu, and within certain sectors of the brain, e.g., somatosensory cortices; neuro-transmitter nuclei in brain stem (Damasio, 1999).

In this perspective, an emotion is a neural object, that is a neural reaction to a certain stimulus, realised by a complex ensemble of neural activations in the brain (*internal emotional state*). These often are preparations for muscular and visceral actions (facial expressions, heart rate increase, etc.); as a consequence, the body will be modified into an *emotional body state*. For our purposes, we reckon the latter to be "observable", either explicitly - vision, hearing -, or implicitly, e.g. via physiological sensor measurements.

Such a disentanglement is a cautionary one: in these terms emotions are likely to be amenable to third person description (observability and modelling), whilst feelings would necessarily involve first person experience (opening to the conundrum of consciousness (Damasio, 1999; Harnad and Scherzer, 2008), which is out of the scope of this thesis). Also, we will use the term "affect" interchangeably with "emotion" in a broad sense (Ziemke and Lowe, 2009).

Under these circumstances, how does affective neuroscience spell out the understanding of facial expression of emotions?

In a nutshell, according to Adolph's model (Adolphs, 2002b), upon the onset of an emotionally meaningful stimulus, namely the expresser's facial action, observer's response undergoes the following stages:

1. fast early perceptual processing of highly salient stimuli (120 ms);
2. detailed perception and emotional reaction involving the body (170 ms);
3. retrieval of conceptual knowledge about the emotion signalled by the expresser's face ($>$ 300 ms).

This deceptively simple account hides a number of cumbersome issues that in order to be addressed call for a finer analysis of neurobiological underpinnings. In the following we summarise in some detail current knowledge concerning main brain areas involved in the process and their structural/functional relationships.

A classic view of the "information flow" from early vision to the affective system, with main areas involved is presented in Figure 3.3 Such complex network of brain areas of specific interest for our work is briefly reviewed in the following.

Visual areas. A broad delineation of two major functional pathways is believed to extend to the organization of the human brain.

The ventral pathway (the *where* pathway) projecting from V1 (striate cortex) through areas V2 and V4 (prestriate cortex) to the inferior temporal (IT) cortex and to the anterior section of superior temporal sulcus (STS) is primarily concerned with the recognition of objects.

The dorsal pathway (the *what* pathway) projecting from V1 through areas V2 and V3 to the middle temporal area (V5/MT) and thence to the superior temporal and parietal cortex is concerned with the perception of spatial information and with the visual guidance of actions towards objects. These two pathways are not completely separate; indeed, a polysensory area in the STS receives inputs both from the ventral and dorsal pathways where form and motion can interact.

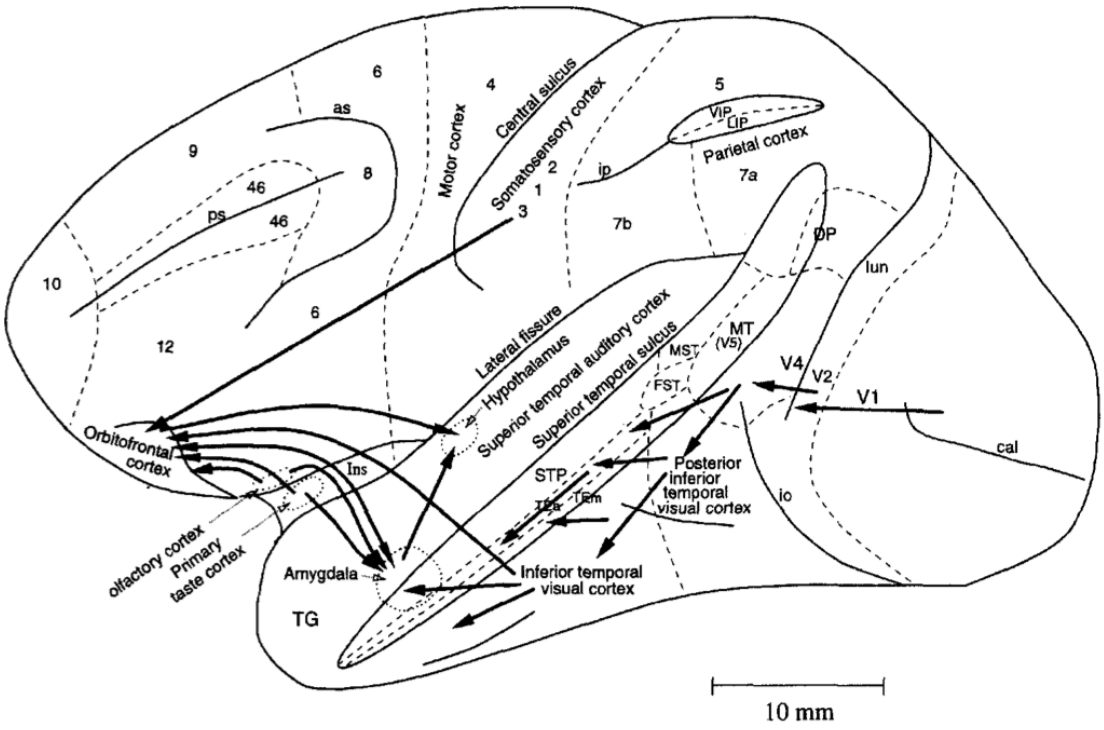


Figure 3.3: *A classic outline of some of the pathway from perception to emotion, in a lateral view of the brain of the macaque monkey. Connections are shown in the ventral visual system from V1 to V2, V4, the inferior temporal visual cortex, etc., with some connections reaching the amygdala and orbitofrontal cortex, as, arcuate sulcus; cal, calcarine sulcus; cs, central sulcus; lf, lateral (or Sylvian) fissure; lun, lunate sulcus; ps, principal sulcus; io, inferior occipital sulcus; ip, intraparietal sulcus (which has been opened to reveal some of the areas it contains); sts, superior temporal sulcus (which has been opened to reveal some of the areas it contains). AIT, anterior inferior temporal cortex; FST, visual motion processing area; LIP, lateral intraparietal area; MST, visual motion processing area; MT, visual motion processing area (also called V5); PIT, posterior inferior temporal cortex; STP, superior temporal plane; TA, architectonic area including auditory association cortex; TE, architectonic area including high-order visual association cortex, and some of its sub-areas TEa and TEm; TG, architectonic area in the temporal pole; V1-V4, visual areas 1-4; VIP, ventral intraparietal area; TEO, architectonic area including posterior visual association cortex. The numerals refer to architectonic areas, and have the following approximate functional equivalence: 1,2,3, somatosensory cortex (posterior to the central sulcus); 4, motor cortex; 5, superior parietal lobule; 7a, inferior parietal lobule, visual part; 7b, inferior parietal lobule, somatosensory part; 6, lateral premotor cortex; 8, frontal eye field; 12, part of orbitofrontal cortex; 46, dorsolateral prefrontal cortex*

Milner and Goodale (1993) substantially reinterpreted these functions and postulated that both streams process information about object features and their spatial localisation, but that the visual information is used differentially by each stream. The ventral

pathway is implicated in the recognition, categorisation and high-level significance of objects. In contrast, processes supported by the dorsal pathway concern on-line information about the spatial location of objects that is used for the programming and visual control of skilled movements. In this scheme, the primary role of the ventral stream is object recognition whereas the primary role of the dorsal stream is to locate stimuli relative to the observer for the purpose of on-line actions, thus it codes in viewer-centred coordinates. To summarise, the nature of the perception (or action) determines the nature of the processing engaged. This functional dissociation emphasises the output side of visual analysis rather than the input side.

Jeannerod (1994) has proposed a more general distinction between these two streams that relates to pragmatic and semantic representations of action. The former refers to rapid transformation of sensory input into motor commands, whereas the latter refers to the use of cognitive cues for generating actions. The proposed pragmatic representation might depend on cooperation across distributed areas in the parietal lobe and premotor cortex.

A key area, for what concerns the subject of this thesis, is the superior temporal sulcus. This latter region of caudal superior temporal cortex encompasses areas such as the medial superior temporal area (MST), the middle temporal area (MT) and the fundal superior temporal area (FST). Desimone and Ungerleider in a number of seminal studies (Desimone and Ungerleider, 1986; Ungerleider and Desimone, 1986a,b) explored the sequence of visual information processing along the pathway from V1, through MT, into the parietal lobe. They sought to establish the relationships among MT, the heavily myelinated zone of the STS, and the V1 and V2 projection fields in the STS. They concluded that MT in the macaque supplies inputs to a large cortical region surrounding it within the caudal superior temporal sulcus. MT may be the source of inputs to a hierarchical system for motion analysis that includes several different visual areas. More rostral parts of the STS cannot be assigned to either the dorsal or the ventral stream of visual processing, they are located at the transition between the two pathways. The properties of neurons in the rostral parts of the STS, namely those within a large portion of the upper bank and fundus of the rostral STS comprise large receptive fields, which always include the centre of gaze and usually extend well into both visual hemifields. These neurons typically show sensitivity to movement, have polymodal responsiveness to visual, somatosensory and/or auditory input, and are insensitive to stimulus form (Karnath, 2001).

Most important, research on the STS has shown that it responds more to facial expressions than to neutral faces. Said et al. (2010) using targeted high-resolution fMRI measurements of the lateral cortex and multivoxel pattern analysis, have provided evidence that the response to seven categories of dynamic facial expressions can be decoded in both the posterior STS (pSTS) and anterior STS (aSTS). They were also able to decode patterns corresponding to these expressions in the frontal operculum (FO), a structure that has also been shown to respond to facial expressions. The response in FO reflects activation of the mirror neuron system. Mirror neurons are especially concentrated in monkey area F5, which is believed to be homologous to area FO in humans (Said et al., 2010). It is likely that the activity in FO related to specific facial expressions is due to mirror neurons, which fire upon the perception of expressions and which might also drive microexpression production in response.

Amygdala. A great deal of sensory input reaches the amygdala, a key structure for many of the visceral and behavioral expressions of emotion. Its circuitry and function has been well-conserved across evolution although species differences do exist. Even non-mammalian species such as reptiles, birds and fish have an amygdala-like brain region with similar circuits and functions to the amygdala in mammals: conservation of amygdala circuitry allows findings from one species to inform our appreciation of amygdala functioning in others Janak and Tye (2015)

As to the perception and response to emotional signals, it has been classically linked to fear processing. However, more recent findings have extended its functions to recognition of other emotions or to multiple processes beyond emotion perception, including memory formation, reward processing or social cognition (Diano et al., 2017). Diano et al. (2017) investigated whether facial expressions of different basic emotions (anger, disgust, fear, happiness, sadness and emotional neutrality) and modulate the functional connectivity of the amygdala with the rest of the brain. They have shown that amygdala communications change dynamically depending on perception of various emotional expressions to recruit different brain networks, compared to the functional interactions it entertains during perception of neutral expressions. Besides these differences, all emotions enhanced amygdala functional integration with premotor cortices compared to neutral faces

The amygdala is a complex collection of about a dozen nuclei lying beneath the uncus of the limbic lobe, at the anterior end of the hippocampus and the inferior horn of the lateral ventricle. It merges with the periamygdaloid cortex, which forms part of the surface of the uncus, and with the parahippocampal gyrus. The nuclei are subdivided into a medial, central, and basolateral group, each of which plays a different role in amygdalar function.

Sensory input is of two general types. Much of it is the familiar type about sights, sounds, touches, smells, and tastes. Olfactory information arrives at the medial nuclei, abut the periamygdaloid cortex, both directly from the olfactory bulb and from olfactory cortex. Medial nuclei are relatively small in humans. The rest reaches the basolateral nuclei (by far the largest part of the human amygdala) from the thalamus and from unimodal visual, auditory, somato-sensory, and gustatory association areas. The basolateral nuclei, by far the largest part of the human amygdala, are in some ways like a cortex without layers; they contain pyramidal neurons, are continuous with parahippocampal cortex, and are extensively interconnected with other cortical areas.

A second kind of sensory input, dealing in a more general sense with levels of physical and emotional comfort and discomfort, also reaches the basolateral nuclei from orbital and anterior cingulate cortex and especially the insula.

Finally, visceral sensory inputs reach the central nuclei. The central nuclei are also small, but their interconnections with the hypothalamus and related brainstem nuclei (e.g., periaqueductal gray, PAG) underlie one aspect of the amygdala's involvement in emotional responses. Projections from the basolateral nuclei to the central nuclei provide a key link between the experience of emotions and their expression.

In particular, the quantitative analysis of global brain connectivity shows that the amygdala occupies a central position in the graphical representation (Fig. 3.4), suggesting its role as a central hub linking multiple peripheral hubs, each linking regions within separated functional clusters. In this way, the amygdala appears as a strong

candidate for the integration of cognitive and emotional information.

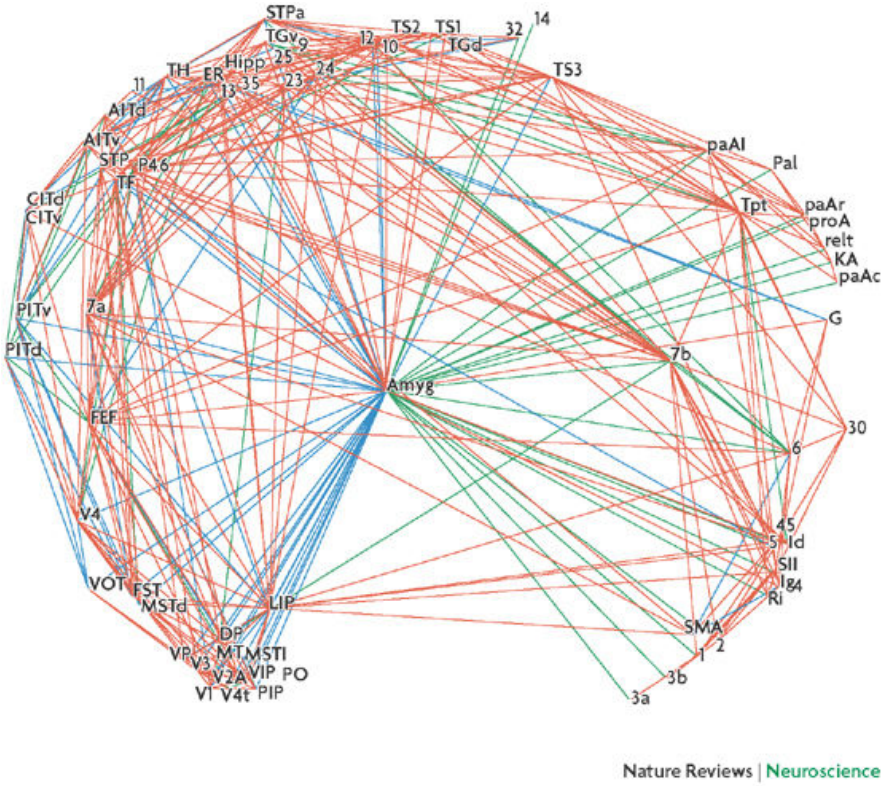


Figure 3.4: *Quantitative analysis of brain connectivity reveals several clusters of highly interconnected regions (represented by different colours). In this analysis, the amygdala (Amyg, centre of figure) was connected to all but 8 cortical areas. Figure from Sporns et al. (2004)*

Orbitofrontal cortex. An additional key structure in emotion processing, with which the amygdala is intimately connected, is the orbitofrontal cortex (OFC) (Adolphs, 2002a,b) Damage to the orbitofrontal cortex, especially on the right, can result in impaired recognition of emotions from the face and the voice. These findings are consistent with the activation found in right orbitofrontal cortex when comparing presentations of fearful and neutral faces. In contrast to the amygdala's activation in response to passive viewing of emotional faces or gender judgements, prefrontal regions may be activated when subjects are engaged in a cognitive task requiring explicit identification of the emotion (Adolphs, 2002a,b)

The OFC is the portion of prefrontal cortex that sits just above the orbits of the eyes and extends posteriorly several centimetres to form the frontal "base" of the brain. It is considered anatomically synonymous with the ventromedial prefrontal cortex (VMPFC). Therefore, the region is distinguished due to the distinct neural connections and the distinct functions it performs. It is defined as the part of the prefrontal cortex that receives projections from the magnocellular, medial nucleus of the mediodorsal thalamus, and is thought to represent emotion and reward in decision making (Bechara et al., 2000).

Descriptively, locations within anterior prefrontal cortex can be expressed in terms of medial and lateral orbitofrontal cortex (Öngür et al., 2003). A better way to view

this cortex is based on its connectivity pattern, which reveals orbital and medial networks. The orbital network receives wide-ranging sensory information and appears to integrate it, particularly in relation to the assessment of food and reward. The medial network is distinctively and heavily connected with regions of the medial wall of the brain, including those of cingulate cortex and surrounding areas. In contrast to the orbital network, the medial network receives few sensory inputs. Importantly, the medial OFC projects to the hypothalamus and other visceral-control areas suggesting that it is involved in the visceral modulation of emotion Öngür et al. (2003). Indeed, via the hypothalamus, descending medial orbitofrontal influence appears to extend as far as autonomic centers in the spinal cord. In contrast, there are relatively few projections to the hypothalamus from the orbital network. Therefore, at least through its medial network, the orbitofrontal cortex interfaces with autonomic brain regions. Notably, the same network is strongly interconnected with cingulate cortex.

Destruction of the OFC through acquired brain injury typically leads to a pattern of disinhibited behaviour. Examples include swearing excessively, hypersexuality, poor social interaction, compulsive gambling, drug use (including alcohol and tobacco), and poor empathising ability. Disinhibited behaviour by patients with some forms of frontotemporal dementia is thought to be caused by degeneration of the OFC(Bechara et al., 2000).

Bechara et al. (2000) posit the OFC as the basis of the "somatic marker hypothesis" (SMH). According to the latter, over time, emotions and their corresponding bodily changes (the "somatic markers"), become associated with particular situations and their past outcomes. When making subsequent decisions, these somatic markers and their evoked emotions are consciously or unconsciously associated with their past outcomes, and influence decision-making in favor of some behaviors instead of others. For instance, when a somatic marker associated with a positive outcome is perceived, the person may feel happy and thereby motivated to pursue that behavior. When a somatic marker associated with the negative outcome is perceived, the person may feel sad, which acts as an internal alarm to warn the individual to avoid that course of action. These situation-specific somatic states based on, and reinforced by, past experiences help to guide behaviour in favour of more advantageous choices, and therefore are adaptive. The amygdala and VMPFC are essential components of this hypothesised mechanism, and therefore damage to either structure will disrupt decision-making.

Insula. Besides the amygdala, OFC is closely associated with the anterior insula. The insular cortex lies buried in the depths of the lateral sulcus (also called the Sylvian fissure) between the frontal and temporal lobes; it is concealed from view by portions of the frontal, parietal, and temporal lobes. It overlies the site where the telencephalon and diencephalon fuse during embryological development and it can be revealed by prying open the lateral sulcus or by removing the overlying portions of other lobes. The portion of a given lobe overlying the insula is called an operculum (Latin for "lid"); there are frontal, parietal, and temporal opercula. The circular sulcus outlines the insula and marks its borders with the opercular areas of cortex.

The insular cortex is the primary interoceptive cortex and integrates visceral, pain, and temperature sensations (Craig, 2003). The dorsal insula has a viscerotropic organization and receives topographically organised inputs from gustatory, visceral, muscle,

and skin receptors via the thalamus. The dorsal insula projects to the right anterior insula, which conveys the conscious experience of bodily sensation by integrating these interoceptive inputs with inputs from cortical areas involved in perceptive, emotional, and cognitive processing (Craig, 2003). For example, functional magnetic resonance (fMRI) studies showed increased activity of the anterior insula during tasks in which subjects attended to the timing of their heartbeats (Craig, 2003). The insula is also a visceromotor area controlling both the sympathetic and parasympathetic outputs, primarily via a relay in the lateral hypothalamus.

The pattern of activation observed in many functional imaging studies suggests a posterior-to-anterior processing gradient in the human insular cortex (Craig and Craig, 2009), and it has been proposed that subjective feelings are based directly on representations of homeostatic sensory integration in the anterior insula, consistent with the James-Lange theory of emotion and Damasio's "somatic marker" hypothesis (Craig, 2003).

To sum up, insula cortex appears to be particularly important in the cortical representation of internal state, with right anterior insula subserving interoceptive awareness and its expression as emotional feeling states and the interaction of interoceptive information with the conscious appraisal of other information (Craig, 2003; Cauda et al., 2012)

Hypothalamus. The human hypothalamus is a pearl-sized structure containing a number of nuclei. As its name implies, it is located just below the thalamus and thus just above the brainstem. The importance of the hypothalamus in certain aspects of emotion is well known, as highlighted by the early work of Philip Bard and Walter Cannon, who showed in their decortication experiments that coordinated emotional expressions were abolished when the hypothalamus was excised, but not when only cortex was compromised

The hypothalamus and the adjacent preoptic area have a central role in integrated autonomic and endocrine responses necessary for homeostasis and adaptation. These include regulation of the circadian rhythms and sleep-wake cycle, thermoregulation, glucoregulation, osmoregulation, responses to stress, and immunomodu- lation (Saper, 2002). The hypothalamus acts as a visceromotor pattern generator that initiates specific patterns of autonomic responses according to the stimulus, such as hypoglycemia, changes in blood temperature or osmolarity, or external stressors.

Our knowledge of hypothalamic connectivity has considerably expanded, too. The traditional view of the hypothalamus emphasised its descending functions. However important the hypothalamus may be for descending control, though, a significant recent insight is that mammalian cerebral cortex and the hypothalamus share massive bidirectional connections. Recent findings have made it clear that, through extensive ascending projections, the hypothalamus provides direct input to the entire cortical mantle with the potential to influence both perception and cognition (Pessoa, 2008) .

In sum, whereas the hypothalamus is involved in a host of basic control functions, it is part of an extensive bidirectional connective system with cortex and many other subcortical structures in a manner that allows for integration of wide-ranging signals. Critically, the hypothalamus is linked to other structures that have themselves broad connectivity, including the basal forebrain and the amygdala, further expanding its po-

tential for influencing information processing.

Brainstem. The spinal and brainstem levels of the hierarchical homeostatic network are present in all mammals, but in humans there is a high-resolution representation of the condition of the body in primary interoceptive cortex in the posterior insula, which generates an energy- efficient map of all brain activity in the anterior insula that underpins the subjective "material me". Thus, there is a fundamental relationship between the physiological condition of the body and subjective feelings of all kinds, just as William James hypothesised (Craig and Craig, 2009).

The brainstem contains several structures that are likely of critical importance in the generation and experience of emotion (see Venkatraman et al., 2017; Holstege, 2016 for a review). The brainstem areas controlling the sympathetic and parasympathetic outputs include the periaqueductal gray matter of the midbrain (PAG), the parabrachial nucleus (PBN) and adjacent areas of the pons, several medullary regions, including the nucleus of the solitary tract (NTS), ventrolateral reticular formation of the medulla, and medullary raphe.

The PAG is the interface between the forebrain and the lower brainstem and has a major role in integrated autonomic and somatic responses to stress, pain modulation, and other adaptive functions. The PAG participates in cardiovascular responses associated with pain modulation, thermoregulation, coordination of the micturition reflex and respiratory rhythm and airway resistance. Experimental studies also indicate that the PAG may have an important role in mechanisms of arousal and regulation of REM sleep. These multiple functions depend on extensive PAG interconnections with the prefrontal and anterior cingulate cortex, amygdala, hypothalamus, and brainstem autonomic, motor, and pain modulatory areas.

The parabrachial complex PBN is a major relay and coordinating center. It receives converging visceral, nociceptive, and thermo- receptive inputs from the spinal cord and conveys this information to the hypothalamus, amygdala, and thalamus. The PBN also contains separate sub-nuclei that participate in the control of respiratory, cardiovascular, and gastrointestinal reflexes.

The nucleus of the solitary tract (NTS), is the first relay station of taste and visceral afferent information and includes several subnuclei with a viscerotropic organization. The rostral portion of the NTS receives taste inputs; the intermediate portion receives gastrointestinal afferents; and the caudal portion receives afferents from baroreceptors, cardiac receptors, chemoreceptors, and pulmonary receptors. The different subnuclei of the NTS relay this information, either directly or via the PBN, to the PAG, hypothalamus, amygdala and to ventral thalamic nuclei projecting to the insular cortex . The second major function of the NTS is to serve as the first central relay for all afferents that trigger medullary reflexes controlling cardiovascular function, including the baroreflex and cardiac reflexes, respiration, including the carotid chemoreflex and pulmonary mechanoreflexes and gastrointestinal motility, particularly that of the esophagus and stomach.

The rostral ventrolateral medulla (RVLM), including the C1 group of epinephrine-containing neurons, is a key area for regulation of arterial blood pressure. Glutamatergic neurons of the RVLM project directly to sympathetic preganglionic neurons controlling cardiac output and blood flow to the skeletal muscle and visceral organs. The

sympathoexcitatory RVLM neurons receive and integrate a large variety of inputs from the brainstem and forebrain. Other medullary inputs to the RVLM contribute to cardiac, vestibular, exercise, and nociceptive sympathoexcitatory reflexes, as well as sympathoexcitatory responses to hypoxia The RVLM also receives several inputs from the hypothalamus, including the PVN, which trigger sympathoexcitatory responses to internal or external stimuli

The caudal ventrolateral medulla contains GABAergic neurons that maintain a tonic inhibitory control on the RVLM and relay the inhibitory inputs from barosensitive neurons of the NTS, thus mediating the sympathoinhibitory component of the arterial baroreflex. The medullary reticular formation also contains the ventral respiratory column, which consists of several groups of inspiratory or expiratory neurons organized into different rostrocaudal groups and involved in respiratory rhythmogenesis.

For what specifically concerns emotional facial expressions, an overall synthetic account of how such perception and affective systems interact in the case of basic, discrete emotion display, has been proposed by Haxby et al. (2000). At its simplest, the model proposes a core visual system for face perception. This constitutes the inferior occipital gyrus (IOG, for low- level facial feature analysis), the lateral fusiform gyrus (LFG, for higher-order invariant aspects of faces such as identity) and the STS (for variable aspects of faces such as lip movement and speech comprehension). This then interacts with an extended system, which involves different structures for different emotions (Figure 3.5).

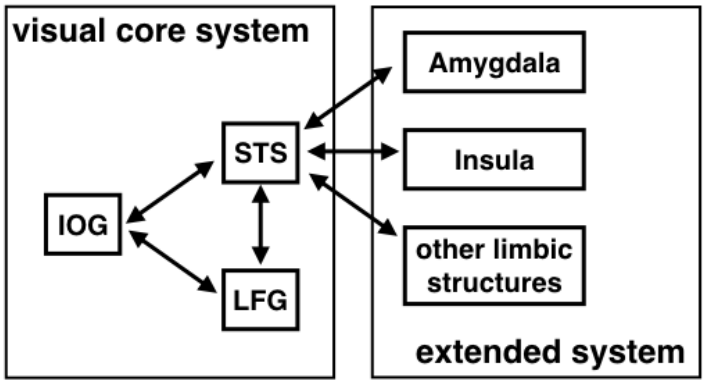


Figure 3.5: *Haxby et al. (2000) model of the distributed human neural system for face perception. The model is divided into a core system, consisting of three regions of occipitotemporal visual extrastriate cortex (IOG, LFG, STS) and an extended system, consisting of regions that are also parts of neural systems for other cognitive functions. Among them, there are the structures involved in the emotional response. Interactions between these representations in the core system and regions in the extended system mediate processing of the spatial focus of expresser's attention, speech-related mouth movements, facial expression and identity. Abbreviations: IOG, inferior occipital gyrus; LFG, lateral fusiform gyrus; STS, superior temporal sulcus. Adapted from (Haxby et al., 2000)*

Facial expressions in dyadic interactions: a fresh view

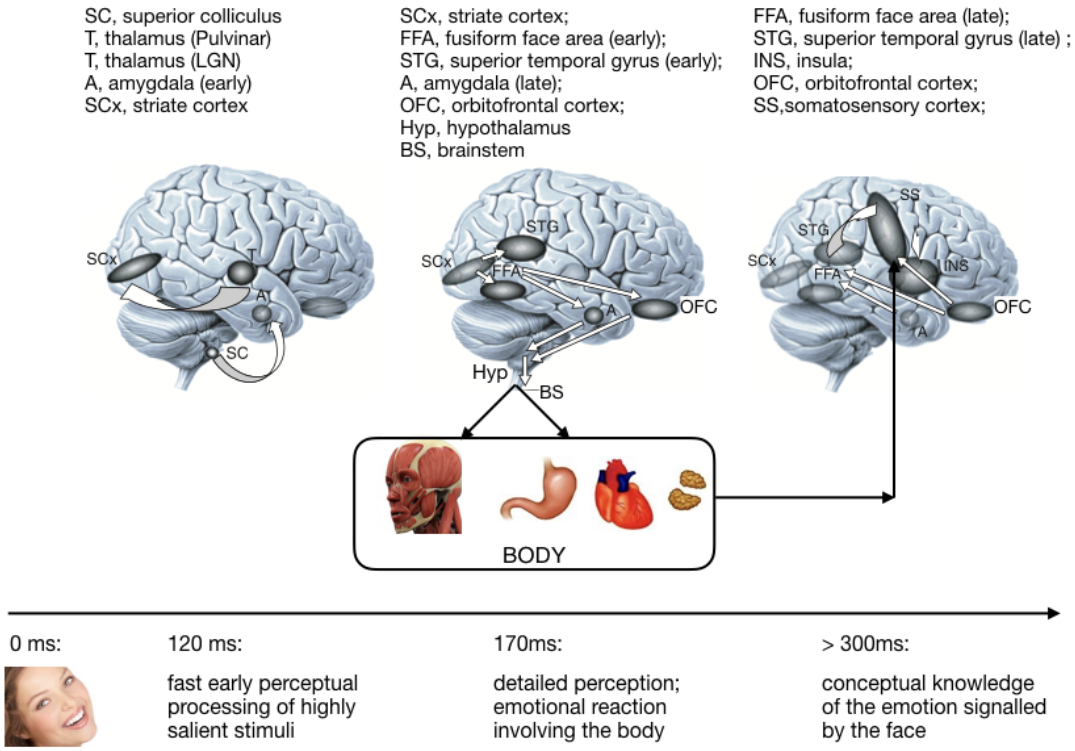


Figure 3.6: *Time course of facial expression perception and recognition according to Adolphs (2002a,b). The figure begins on the left with the onset of the stimulus, a facial expression of emotion, and progresses through perception to final recognition of the emotion on the right. Certain brain structures are preferentially engaged in processing structural information of the stimulus (early perception), whereas others participate more in retrieving conceptual knowledge or linking the per- ceptual representation to the modulation of other cognitive processes or to the elicitation of physiological states (e.g., an emotional somatic reaction to a stimulus). Attempts to localise the perception/recognition of the stimulus in space or in time have trade-offs: spatial localisation permits us to tie aspects of the processing to particular brain structures but suffers from the fact that the same brain structure participates in different components of processing at different points in time. Temporal localisation that treats the brain as a dynamical system has the benefit of describing the evolution of perception and recognition more accurately in time, although this will usually encompass a large set of different structures. A full account will need to describe both the spatial and temporal aspects of processing, a description that is becoming possible by combining techniques with high spatial resolution, such as functional magnetic resonance imaging, with techniques with high temporal resolution, such as event-related potentials. Note that the figure omits many structures and connections to provide a schematic overview. Adapted from Adolphs (2002a,b)*

Beyond the classic "visual pipeline view" on the perception and recognition of facial emotion expressions along a dyadic interaction, the simulation-based account, as discussed in Chapter 2, offers a perspective that is more deeply grounded in what we currently know about the neurobiological underpinnings of such processes. Adolph's

proposal, briefly summarised at the beginning of this section, is one such account (Adolphs, 2002a,b) and it is synthetically outlined in Figure 3.6.

Upon presentation of an emotionally meaningful stimulus deployed by an expresser, a feed-forward sweep of information processing proceeds along occipital and temporal neocortices and extracts perceptual information from expresser's faces so that after 100 ms in humans, would coarsely categorise the stimulus as displaying an emotion or not. Amygdala and orbitofrontal cortices are likely to participate in at least three distinct ways. First, they may modulate perceptual representations via feedback. This mechanism can contribute, in particular, to fine-tuning the categorisation of the facial expression and to the allocation of attention to certain of its features.

Second, the amygdala and orbitofrontal cortices may trigger associated knowledge, via projections to other regions of neocortex and to the hippocampal formation. This mechanism enables retrieval of conceptual knowledge about the emotion, necessary for final expression recognition.

Third, and most important here, they generate an emotional response in the subject, via connections to the motor structures, hypothalamus, and brainstem nuclei, where components of an emotional response to the facial expression can be activated. This mechanism contributes to the elicitation of knowledge about the expresser's emotional state, via the process of internal simulation, and would draw on the insula and somatosensory related cortices (right hemisphere) for representing the emotional changes in the observer. However, it may be probable that the observer's simulation of expresser's emotion proceeds via the generation of a somatosensory image of the associated body state, even in the absence of actual motor mimicry, namely the "as if" mechanism that has been conjectured by Bechara and Damasio (2005), though details are matter of debate (Adolphs, 2002b; Rizzolatti and Sinigaglia, 2016; Wood et al., 2016).

It is worth noticing, that Adolph's scheme, together with a large body of knowledge that we have tried to summarise above, point at the OFC, the amygdala and the insula, three highly interconnected areas, as the key "central" structures for the development of an emotional episode along affective interaction. Interestingly, it has been argued (Salzman and Fusi, 2010) that functional interactions between the amygdala and prefrontal cortex form a potential neural substrate for the encoding of the psychological dimensions of valence and arousal, the core affect postulated by Russell (1980)), at the psychological theory level. A similar role is also played by the insula (Craig and Craig, 2009; Craig, 2003) The valence and arousal dimensions, originally formulated at the psychological explanation level, can be thought of as "emotion primitives" supporting at the neurobiological level a central continuous emotion space. Cogently, it has been surmised that such evolutionary building blocks are shared across emotions and across phylogeny (Anderson and Adolphs, 2014).

However, Adolph's account, besides the merits of clearly providing a neurobiological framework for a simulation-based grounding of expresser/observer affective interaction, partially overlooks or, at least, does not provide further details about the explicit involvement of motor mechanisms in the process, albeit these being acknowledged (Adolphs, 2002b). This is not a minor issue. For instance, while investigating whether facial expressions of different basic emotions modulate the functional connectivity of the amygdala with the rest of the brain, Diano et al. (2017) have shown that all queried

emotions enhanced the amygdala functional integration with premotor cortices. The coupling of amygdala with motor areas outlines the influence of different emotions in fostering action preparation and planning, as well as motor resonance. Observing emotional stimuli increases motor excitability relative to neutral images, which may reflect approach and avoidance preparation, motor mimicry and emotional contagion. Such functional interaction is consistent with the anatomical evidence indicating white matter connectivity between the amygdala and the motor regions in nonhuman primates and humans.

Indeed, if one dismisses the traditional cognitive view that facial action understanding is grounded on serial inferential processing from early to higher order sensory elaboration, a viable path is explicitly taking into account the observer's "inner motor knowledge".

Motor aspects of mirroring other's emotion

Facial expressions are facial actions and it has been posited that a "resonance behavior" is one such mechanism where an individual repeats overtly a movement made by another individual. A striking example is provided by either human or rhesus monkey newborns imitating adult facial gestures, (Tramacere and Ferrari, 2016; Ferrari et al., 2006). At the neurobiological level the resonance or mirroring behaviour has been initially explained in terms of mirror neuron (MN) activity.

MNs have been first localised in monkeys' ventral premotor (PMv) area F5 and then in the inferior parietal lobule (IPL). These visuomotor neurons activate when the monkey executes a hand or mouth goal-directed motor act (e.g., grasping, biting or manipulating an object) and when the monkey observes the same, or a similar, act performed by the experimenter or by a conspecific. Subsequently, other areas have been endowed with mirroring capabilities, so that it is currently more appropriate to address the properties of a MN system (MNS) or network (Bonini, 2017; Tramacere and Ferrari, 2016).

Crucially, evidence has been given that also the human brain is provided with mirroring capabilities, and internal simulation has been related to coding the intentions of actions performed by others (Fogassi and Ferrari, 2011). In recent years, evidence has accumulated for the existence of a MNS. The initial studies on the neurobiology of imitation in humans suggested a core imitation circuitry composed of three major neural systems (Iacoboni, 2005, 2009): the posterior part of the superior temporal sulcus (pSTS), the rostral part of the inferior parietal lobule (rIPL), and the posterior part of the inferior frontal gyrus and adjacent ventral premotor cortex (pIFG/vPMC complex). The information processing flow between these neural systems that is relevant to imitation is likely to occur as follows (Carr et al., 2003):

1. the posterior part of the STS provides a higher order visual processing of the observed action by coding an early visual description of the action;

2. this information is sent to the other two neural systems (rIPL and pIFG/vPMC complex), which are thought to form a parieto-frontal mirroring (both motor and visual) system; this privileged flow of information is supported by the robust anatomical connections between superior temporal and posterior parietal cortex,

and becomes active during action observation, action execution, and also during imitation;

3. the posterior parietal cortex codes the precise kinesthetic aspect of the movement and sends this information to inferior frontal mirror neurons in the pIFG;

4. the inferior frontal cortex codes the goal of the action;

5. efferent copies of motor plans are sent from parietal and frontal mirror areas back to the STS; here there would be a matching process between the visual description of the observed action and the anticipated outcome of the planned imitative action;

6. if there is a good match, the action is executed; otherwise, a correction of the motor plan is implemented.

Clear evidence has been provided by Nishitani and Hari (2002) who asked subjects to imitate non-namable lip forms presented in still pictures. The brain activation, recorded by magnetoencephalography (MEG) , showed increasingly longer latencies starting from occipital visual areas, then STS, IPL, inferior frontal gyrus, and finally the primary motor cortex; the whole sequence took 250 ms. They have argued that such a progression of activity could be related to visual recognition that starts by visual analysis and ends, via feed-forward connections, by recognition (and understanding) of the actions.

Wolpert et al. (2003), while examining the extent to which motor commands acting on the body can be equated with communicative signals acting on other people, have mapped such information flow onto a theoretical framework. The latter, named MOSAIC, is defined in terms of paired forward and inverse internal models, and it has been developed in the motor learning and control field (Wolpert et al., 2003). According to MOSAIC, internal models are input/output functions that mimic experience-dependent sensory-motor states. The inverse model is a controller that retrieves the motor plan necessary to achieve a desired sensory state (or goal), whereas the forward model is a predictor of the sensory consequences of a motor plan. Thus, the inverse model is updated on the basis of the forward model.

In neural terms (see Figure 3.7), the input of the inverse model would be the STS output that is sent to the frontoparietal mirror neuron system, and the output of the inverse model would be the output of the frontoparietal MNS down-stream to motor areas. Efference copies of motor commands originating from the frontoparietal MNS would provide the input of the forward model, whereas its output would be the matching process occurring in the STS (Iacoboni, 2005).

Such theoretical picture has also been endorsed by Hari and Kujala (2009) who assumed the experimental activation sequence of the core system to be a kind of hierarchical predictive processing, with reciprocal feedback connections between STS, parietal MNs and frontal MNs. In line with Bayesian modelling of brain functions that incorporates prior information (the context), the activation sequence is defined in terms of predictive coding (Kilner et al., 2007). The assumption is that neuronal processing at each stage both generates predictions of the input reaching the next phase and is sensitive to the prediction error resulting from lower-level computations. The prediction error (roughly, the difference between the real and predicted data) is then minimised.

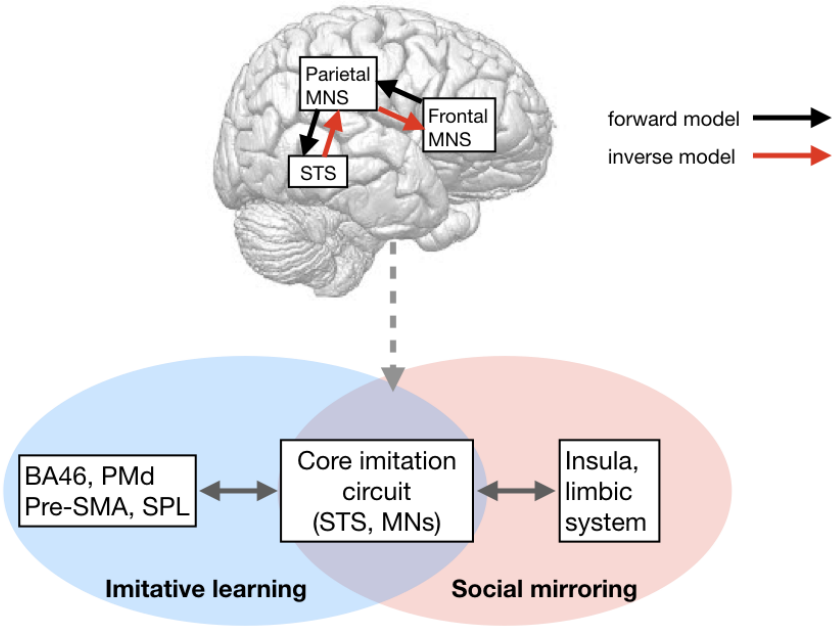


Figure 3.7: *Neural mechanisms of imitative learning and social mirroring. Top: a representation of the core circuitry for imitation on the lateral wall of the right cerebral hemisphere, together with the internal models the circuitry implements during imitation. Bottom: imitative learning is implemented by interactions among the core imitationcircuit, the dorsolateral prefrontal cortex (BA46) and a set of areas relevant to motor preparation (PMd, pre-SMA, SPL), whereas social mirroring is implemented by the interactions among the core imitation circuit, the insula and the limbic system. Abbreviations: MNS, mirror neuron system; STS, superior temporal sulcus, BA46, Brodmann area 46; MNS, mirror neuron system; PMd, dorsal premotor cortex; pre-SMA, pre-supplementary motor area; SPL, superior parietal lobule. Adapted from Iacoboni (2005)*

This scheme suggests that the core circuitry for imitation indeed interacts with the limbic system (the neural system concerned with emotions) during social mirroring (Iacoboni, 2005). fMRI studies of the observation and imitation of facial emotional expressions have revealed a large-scale neural network that comprises the core circuitry for imitation (the MNS and the STS), the insula and the limbic system (Carr et al., 2003). Subsequent studies of the imitation or observation of emotions have supported the idea that empathy has a sensory-motor, mirroring basis, e.g. the study on pain by Avenanti et al. (2005).

It is worth noticing that in such scheme STS activity goes beyond its classical function of visual motion processing network. It provides visual input to the MNs so to allow matching between sensory predictions of imitative motor plans and a visual description of observed actions. Molenberghs et al. (2010) have reported enhanced activity of the STS during imitation. Their results call for a role of the STS in registering the congruence between one's own actions and the actions of others, and perhaps enhancing the perceptual interpretation of others' actions when they match the observer's own motor plans. This way, the STS provides an interface between observer's own actions and the actions he or she observes, a neural substrate for scaffolding various

aspects of social cognition. More specifically, Lahnakoski et al. (2012), by using stimuli in the form of audiovisual movie clips depicting pre-selected social signals, have shown that the posterior superior temporal sulcus (pSTS) responded to all social features but not to any non-social features, and the anterior STS responded to all social features except bodies and biological motion. Interestingly, they have identified four partially segregated, extended networks for processing of specific social signals: (1) a fronto-temporal network responding to multiple social categories, (2) a fronto-parietal network preferentially activated to bodies, motion, and pain, (3) a temporo-amygdalar network responding to faces, social interaction, and speech, and (4) a fronto-insular network responding to pain, emotions, social interactions, and speech. Their results, achieved under conditions that resemble the complexity of real life, highlight the role of the pSTS in processing multiple aspects of social information.

In this perspective Chakrabarti et al. (2006), who considered the perception of dynamic facial expressions of emotion, have proposed that an intermediate module for "action perception is involved, namely the MNS. Their proposal is outlined in Figure 3.8.

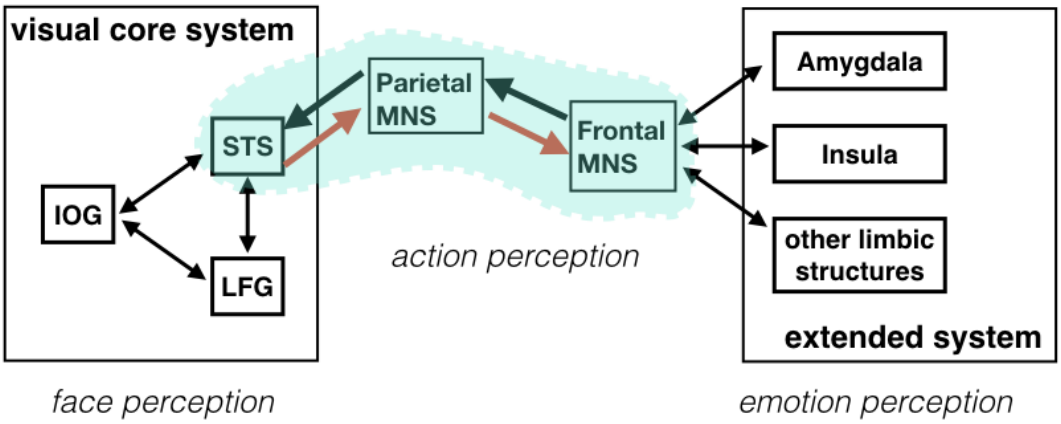


Figure 3.8: *Chakrabarti et al. (2006) modification of Haxby et al. (2000) model of distributed neural system for perception of dynamic facial expressions of emotion. The model incorporates a module for "action perception" based on the human MNs. Abbreviations: IOG, inferior occipital gyrus; LFG, lateral fusiform gyrus; STS, superior temporal sulcus. Adapted from (Chakrabarti et al., 2006)*

In the framework of this thesis, Chakrabarti et al. (2006) scheme represents a starting point to identify a *somatomotor route* for the perception and mirroring of dynamic facial expressions of emotion.

Clearly, the interaction between affective structures and the motor pathway to effectively generate the facial actions, for example along mimicry, is more complex than the scheme above suggests. Indeed one has to take into account the role played by lower level brain areas such as the hypothalamus and the brainstem. A clear example is provided by the amygdalo-motor pathways for the control of facial expressions (see Figure 3.9). As discussed by Gothard (2014), it is evident the distributed nature of the joint activation of multiple motor areas that initiate the production of a facial expression. Concomitantly multiple areas monitor ongoing overt behaviours (the expression

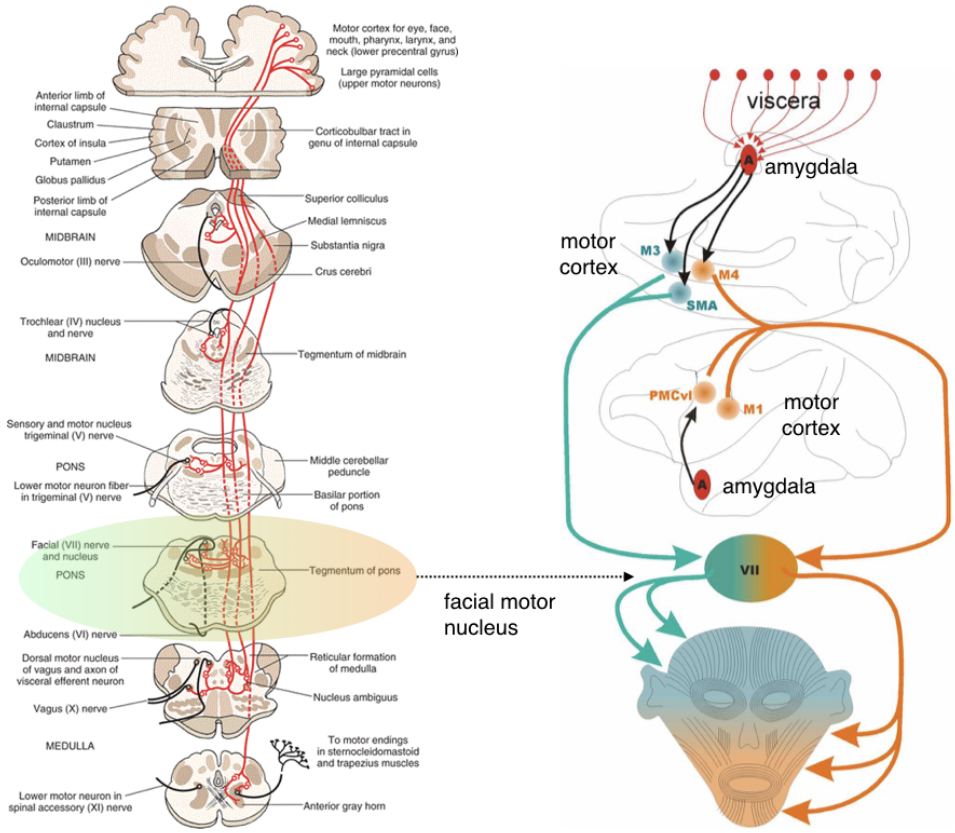


Figure 3.9: *Left: the classic motor pathway for controlling the facial expression. Right: amygdalo-motor pathways. The lower half of the face is controlled by the coordinated activity of three motor areas: M1, primary motor cortex; PMCvl, premotor cortex ventrolateral division; and M4, caudal face area of the midcingulate cortex. The upper half of the face is controlled by the coordinated activity of two motor areas: SMA, supplementary motor area; and M3, the anterior face area of the midcingulate cortex. The black arrows indicate direct projections from the basal nucleus of the amygdala to PMCvl, M3, M4, and SMA, The first segment of the orange and green lines indicate the corticobulbar tract. VII, pontine facial nucleus that contain the motor neurons that synapse on the muscles of facial expressions. The medial division of the facial nucleus contains the motor neurons that control muscles in that upper half of the face (in green) while the lateral division contains the neurons that control the muscles in the lower part of the face (in orange). Note that the amygdala receives multiple lines of viscerosensory input (red arrows, top) that are likely integrated in the output directed at facial motor areas. Adapted from Gothard (2014)*

itself) and the covert, autonomic responses that accompany emotional expressions. The ventrolateral regions of the premotor cortex (PMCvl) initiates movements triggered by external cues; the supplementary motor cortex (SMA) directly innervates motor neurons in the medial segment of the facial nucleus. The anterior face area of the midcingulate cortex, (designated as M3) gives rise to projections that target bilaterally the medial segments of the facial nucleus harbouring the motor neurons that supply the upper facial muscles; notably, projections originating from M3 also target the reticular formation of the brainstem that contains autonomic centers. These are likely to be

activated during emotional states, coordinating both the overt (behavioural) and covert (autonomic) expression of emotions. The caudal motor area, M4, (located at the border between the anterior and posterior midcingulate) targets the lateral regions of the facial nucleus, especially the motor neurons that supply the upper lip, and is mostly involved in the emotional control of facial expressions. The amygdala forms a closed processing loop with both the anterior cingulate cortex and with area M3, whilst M3 projects to the basal and accessory basal nuclei of the amygdala and the basal nucleus of the amygdala gives rise to feedback projections to all subdivisions of the cingulate cortex (Gothard, 2014).

The final control is provided by the facial motor nucleus in the brainstem (namely, the caudal portion of the ventrolateral pontine tegmentum) a collection of lower motor neurons that innervate the muscles of facial expression (cranial nerve VII) and the stapedius. It has a dorsal and ventral region, with neurons in the dorsal region innervating muscles of the upper face and neurons in the ventral region innervating muscles of the lower face. Nucleus axons take an unusual course, traveling dorsally and looping around the abducens nucleus, then traveling ventrally to exit the ventral pons medial to the spinal trigeminal nucleus. These axons form the motor component of the facial nerve, with parasympathetic and sensory components forming the intermediate nerve. Like all lower motor neurons, cells of the facial motor nucleus receive cortical input from the primary motor cortex in the frontal lobe of the brain. Upper motor neurons of the cortex send axons that descend through the internal capsule and synapse on neurons in the facial motor nucleus. Interestingly, the neurons in the dorsal aspect of the facial motor nucleus receive inputs from both sides of the cortex, while those in the ventral aspect mainly receive contralateral inputs (i.e. from the opposite side of the cortex). The result is that both sides of the brain control the muscles of the upper face, while the right side of the brain controls the lower left side of the face, and the left side of the brain controls the lower right side of the face.

At the same time, the midcingulate and the amygdala receive signals from the viscera via the nucleus of the solitary tract and parabrachial nuclei and via the insula, which integrates interoceptive and exteroceptive signals. Interoceptive afferents, therefore, may modulate both the perception and the production of facial expressions. Indeed, neurons in the amygdala and cingulate cortex discharge in phase with the cardiac and respiratory cycle.

Since at least in the amygdala, neurons respond primarily after the onset of facial activity, overall the mechanism we have described supports a mirror like emotion-to-motor transformation (Gothard, 2014).

Integrating “cold” and “hot” actions

To sum up, a wealth of evidence supports the hypothesis that a mirror/simulation mechanism is instrumental in coordinating facial expressions within dyadic interactions and more generally in social contexts.

Studies on the monkey and human MNs that by and large originally addressed the so-called “cold” actions, without a specific emotional content, more recently have prompted the idea that a mirror mechanism is also present in the cortical areas involved in coding emotions (Gallese et al., 2004), and markedly along affective facial expression processing (Tramacere and Ferrari, 2016). Mirror neurons were found in the

premotor cortex of the marmoset, a New World monkey, indicating that such mirroring mechanism has been highly preserved in the course of evolution. Indeed, neuronal mirroring of observed behaviours has been shown in phylogenetically ancient structures, such as the basal ganglia, and in subcortical regions related to visceromotor reactions, such as the insular and cingulate cortices. It is possible that the involvement of such brain structures during the direct experience and perception of others' emotions has been instrumental for sharing similar emotional experience and could represent the building block of the emergence of empathic behaviours in several species of primates.

In brief, a mechanism similar to somatic motor action mirroring can be hypothesized for internal "hot" actions which can act as a prelude to specific behavioural responses. The difference from the former is that what is matched is not only a gesture (facial expressions), but also internal states. Namely, motor knowledge required is grounded in visceromotor actions, that is motor commands directed to visceral organs (Fogassi and Ferrari, 2011). In both monkeys and humans, subcortical regions much involved in this kind of output are the insular and cingulate cortices (Fogassi and Ferrari, 2011).

Interestingly, recent evidence suggests that interoceptive experience may largely reflect limbic predictions about the expected state of the body that are constrained by ascending visceral sensations (see, Barrett and Simmons, 2015, for an in-depth review and discussion). They introduce the Embodied Predictive Interoception Coding (EPIC) model, which integrates an anatomical model of corticocortical connections with Bayesian active inference principles, to propose that agranular visceromotor cortices contribute to interoception by issuing interoceptive predictions. In a nutshell, visceromotor cortices generate autonomic, hormonal and immunological predictions to adjust how the internal systems of the body deploy autonomic, metabolic and immunological resources to deal with the sensory world, not as it is right now, but as the brain predicts it will be in a moment from now. These visceromotor predictions underlie most of the body's allostatic and anticipatory responses to moment-by-moment physical movements and mental challenges.

At the same time as visceromotor predictions are issued to maintain homeostasis or to enable allostasis, the deep layers of agranular visceromotor cortices send the same information via connections to another part of the interoceptive system: the supragranular layers of the (granular) mid-to-posterior insular cortex, which serves as the primary interoceptive cortex. Once there, these interoceptive predictions initiate a pattern of activity that represents expected interoceptive sensations. Interoceptive signals that result from changes in the viscera, muscles and skin ascend via the lamina I pathway and vagal afferents in the nucleus of the solitary tract, the parabrachial nucleus and the thalamus, before arriving at the granular layer (layer IV) of the primary interoceptive cortex. Granular cortex in primary interoceptive sensory regions of the mid- and posterior insula are architecturally well suited for computing and transmitting prediction error and for propagating prediction-error signals back to visceromotor regions to modify predictions (Barrett and Simmons, 2015).

The agranular visceromotor regions including the anterior insula (AI) are "rich-club hubs" (Barrett and Simmons, 2015) that are part not only of the interoceptive system, but also of the intrinsic control and attention networks. The AI is a key hub in olfactory and gustatory networks, as well as in the brain's multimodal network, which integrates visual, auditory and somatosensory networks. Thus, in turn, agranular visceromotor

cortices inform the rest of the brain of interoceptive changes by sending predictions to these intrinsic networks that are based on anticipated visceromotor consequences.

An important issue to take into account in such proposal is that, at a lower level, as viscerosensory signals undergo substantial signal conditioning in the brainstem (which mediates the viscerosensory signals that are sent to the cortex), it is likely that brainstem and subcortical structures contribute directly to active inference, for instance, by computing themselves a prediction error (Barrett and Simmons, 2015).

Indeed, lower lever brain areas involving hypothalamus and brainstem are likely to play a more key role than usually surmised. The brainstem, including the mesencephalon (midbrain), is a complete brain with somatosensory, auditory (inferior colliculus), visual (superior colliculus), vestibular, and motor systems (Holstege, 2016). After all, the cortex cerebri can be considered a copy of the brainstem - with somatosensory, auditory, visual, vestibular and motor parts -, the major difference being that the cortex contains a great many more neurons than the brainstem, which enable the various cortical systems to function much more precisely with a much larger amount of memory. Holstege (2016) has depicted an overall picture where the motor system is supported by two components (both at the cortex and and brainstem level): the voluntary, somatic motor system allows the individual to move its body parts voluntarily, whereas the emotional motor system controls basic motor activities such as blood pressure, heart rate, respiration (Figure 3.10.

Both subsystems consist of medial and lateral components either at the cortex level and subcortex level. The affective motor subsystem cooperates with the somatic/voluntary one. For instance in speech production, at the cortex level, the Broca area controls the mouth part of the motor cortex , which in turn signals, at the brainstem level, the lateral tegmental field of medulla and caudal pons (premotor interneurons). The latter controls different motor areas, e.g. the lateral part of the facial nucleus, which coordinate muscles for mouth opening, tongue movements, etc.

On the emotional side, mouth movements are controlled in the context of emotional expression. The amygdala and lateral hypothalamus control, at the brainstem level, the periaqueductal grey (PAG), which in this case serves as a central pattern generation of sound production. The PAG modulates the premotor neurons of nucleus ambiguous, which in turn activates several brainstem motor neuron areas eventually controlling the final vocalisation. Crucially, the amygdala and the PAG also have access to the somatic pathway so to tune activities occurring in the lateral tegmental field, involved in voluntary mouth control (Holstege, 2016). Again, in this example too, it is indeed remarkable the manifold modulator role of the amygdala, which is recruited both on the somatomotor side and on the visceromotor side.

All the issues touched in the above discussions can be synthetically subsumed under the architecture of the distributed neural system for perception of dynamic facial expressions of emotion outlined in Figure 3.11.

The scheme shows at a glance the three main routes contributing to affective facial expression processing: the *visual route*, the *visuomotor route* and the *visceromotor route*. The affective core is provided by the close interactions occuring among the amygdala, the insula and the orbitofrontal cortex. The visual input, namely, the expresser's facial action, is processed by early visual occipital cortices, and undergoes a visuomotor mapping provided by STS and IPL. Such motor representation is suit-

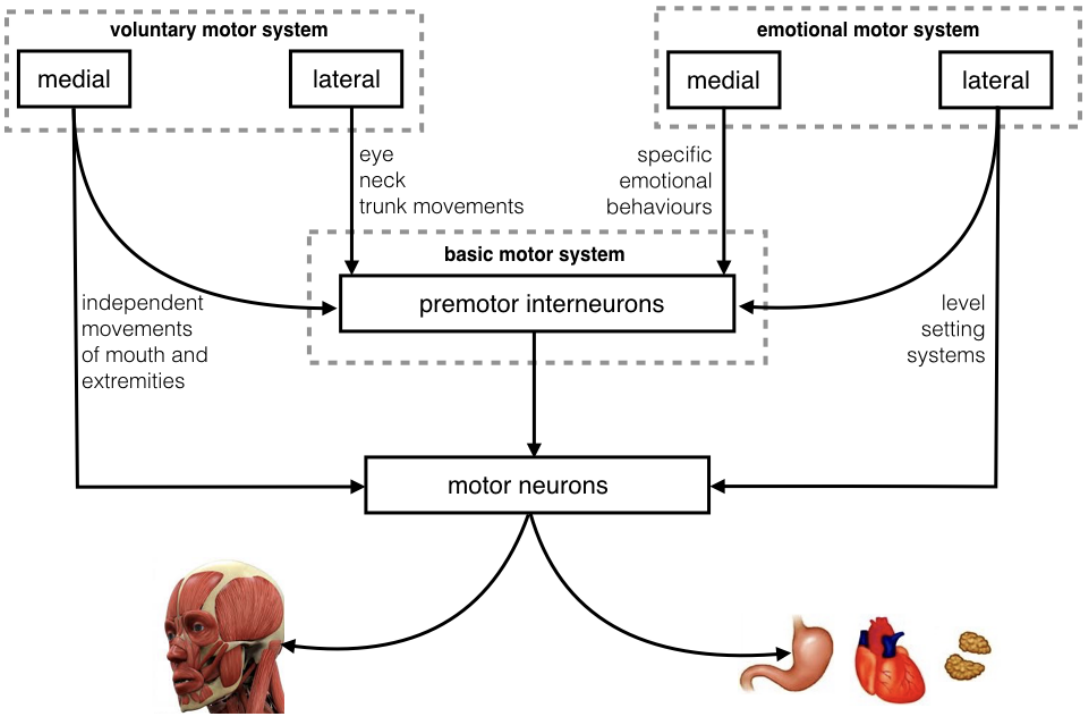


Figure 3.10: *A motor perspective of somatic and affective control at the brainstem level. The overall motor system consists of two subsystems, the voluntary and the emotional motor systems. The voluntary motor system allows the individual to move its body parts voluntarily, whereas the emotional motor system controls basic motor activities such as blood pressure, heart rate, respiration. Both consist of medial and lateral components. These subsystems have access to premotor interneurons and motoneurons. Adapted from Holstege (2016)*

able to trigger the core affect system, which, in turn, activates visceromotor actions. When both the visuomotor route and the visceromotor route are active, core affect structures can modulate perceptual representations via feedback connection; meanwhile, an emotional response is generated in the subject, via connections to motor structures, hypothalamus, and brainstem nuclei. On a short term the process of internal simulation supports observer's responses to expresser's facial action. These can be shaped in terms of observer's overt or covert facial mimicry and physiological responses. On a longer term the simulation of expresser's emotion proceeds via the generation of a somatosensory image of the associated body state, which may contribute, in subsequent processing stages to the elicitation of knowledge about the expresser's emotional state for expression recognition and/or to emotional contagion.

Moving towards the theoretical model: a functional architecture

The distributed neural architecture outlined in Figure 3.11 provides the necessary neurobiological underpinnings and constraints for devising a functional architecture to frame the theoretical model (computational theory level) of an observer in dyadic interaction with an expresser.

At the most general level, we state that the different processes based on such archi-

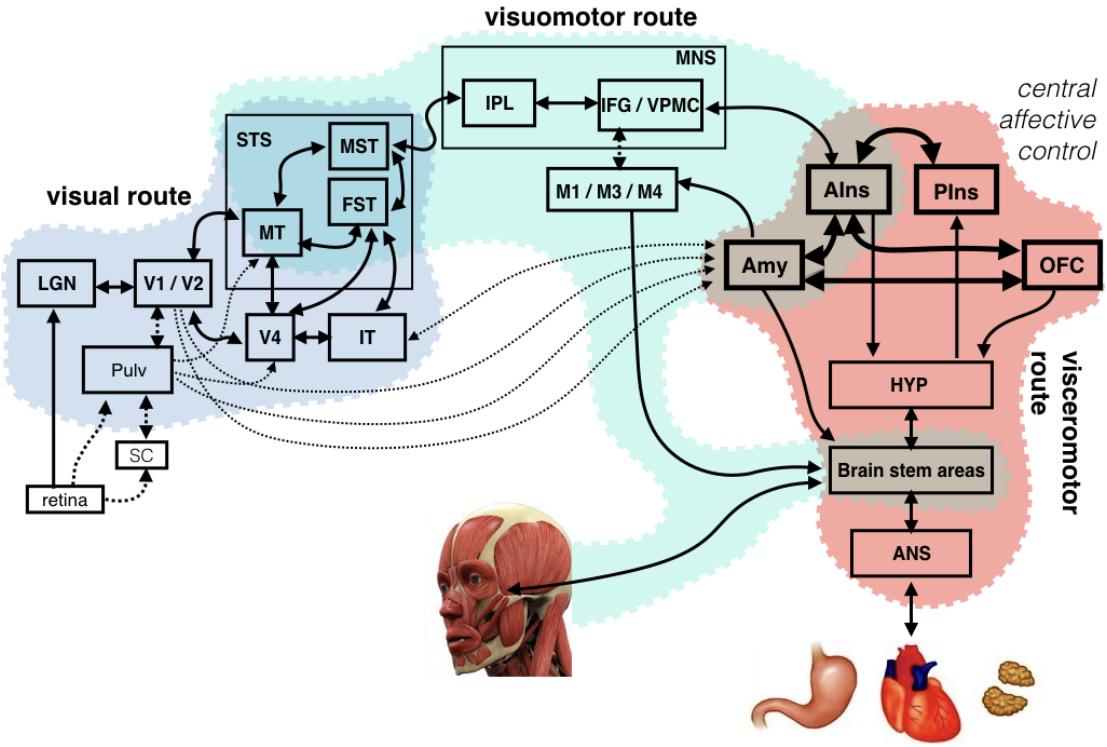


Figure 3.11: *Architecture of a distributed neural system for perception of dynamic facial expressions of emotion. Two, reciprocal, heavy arrowheads indicate "forward" and "backward" projections between areas (boxes). Light dotted projections indicate the possible subcortical dual route from SC/Pulvinar to limbic areas (not included in the current model, but discussed later). Only the main area of interest have been included. The architecture incorporates a module for "action perception" based on the human MNs, which mediates between the external stimuli (expresser's facial action), as processed along the visual route, and the internal motor/action representation. The MNs provides the necessary input to activate the core affect system, represented by the amygdala, the insula and the OFC. This system coordinates the dynamics of the activities occurring along the visuomotor and visceromotor loops, either by modulating perceptual representations via feedback, and by generating an emotional response in the subject, via connections to motor structures, hypothalamus, and brainstem nuclei, where components of an emotional response to the facial expression can be activated. Moment-by-moment, the output can be in terms of observer's overt or covert facial mimicry and physiological responses*

tecture can be best described as stochastic processes. The concepts of uncertainty and noise are fundamental in probabilistic modelling. Computational constraints restrict us to define probabilistic models that are only a crude simplification of the complex reality. Therefore, uncertainty is introduced by our specific modelling assumptions.

Cogently, a person always has core affect, and at each point in time, a person's emotional state can be described in terms of how pleasant or unpleasant and how activated the person is feeling. Across time, a person?s core affect describes a core affect trajectory, reflecting the typical pattern of affective changes and fluctuations that characterise an individual (Kuppens et al., 2010). Describing such trajectory as the realisation of

a stochastic process, is a viable account for changes and fluctuations that people's valence/arousal levels undergo across time, as well as the observed individual differences in core affect variability.

Thus, in order to account for observer's visuomotor and visceromotor dynamics, we make the following working assumptions:

Assumption 0 The core affect can be approached as resulting from a complex, open system. It is embedded in a larger context and is therefore subject to stochastic variability resulting from the many internal and external events that influence core affect.

Assumption 1 The activity of the amygdala-OFC-insula network provides the core affect dynamics, which can be summarised in terms of valence/arousal dimensions, that is the emotional primitives of the system. Formally, such dynamics occurs in a continuous stochastic core affect state-space. A trajectory in such state-space represents a joint visuomotor/visceromotor action.

Assumption 2 Facial actions are represented as trajectories in a stochastic facial action state-space; namely, a trajectory represents the continuous dynamics of a suitable encoding of a finite set of facial motor control parameters. Analogously, given a number N_V of autonomic subsystems, autonomic actions can be encoded as trajectories of action state control variable taking place in the corresponding N_V stochastic state-spaces.

Assumption 3 The internal, egocentric representation of the overall facial dynamics is represented as a continuous stochastic trajectory in a suitable somatomotor state-space; analogously, each stochastic trajectory evolving in a specific visceromotor state-space, represents the evolution of one observer's visceromotor sub-system.

Assumption 4 A perceptual state-space accounts for the overall facial action of the expresser sensed by the observer; thus, a trajectory in such space is the evolution of the observer's perception of the expresser's facial configuration, as inferred via early perceptual modules. A suitable visuomotor mapping provides the necessary transformation from the perceived expresser's facial configuration to the observer's egocentric motor representation.

The different state-spaces exchange information according to the functional architecture outlined in Fig. 3.12, which summarises at a higher, coarser explanation level the neural architecture presented in Figure 3.11

One-head arrows denote forward and backward interactions between state-spaces. Each state-space is endowed with its own dynamics - specified in the above assumptions- but the heart of the model is the "core affect $\rightarrow$ action $\rightarrow$ motor" hierarchy replicated in the visuomotor and visceromotor routes. The simulation aspect of the model strictly relates to such upward-downward cycle of information flow.

In this perspective, we assume that

Assumption 5 Any state-space can be conceived as a dynamic input-state-output model (cfr., Eqs. 3.1 and 3.2 below)

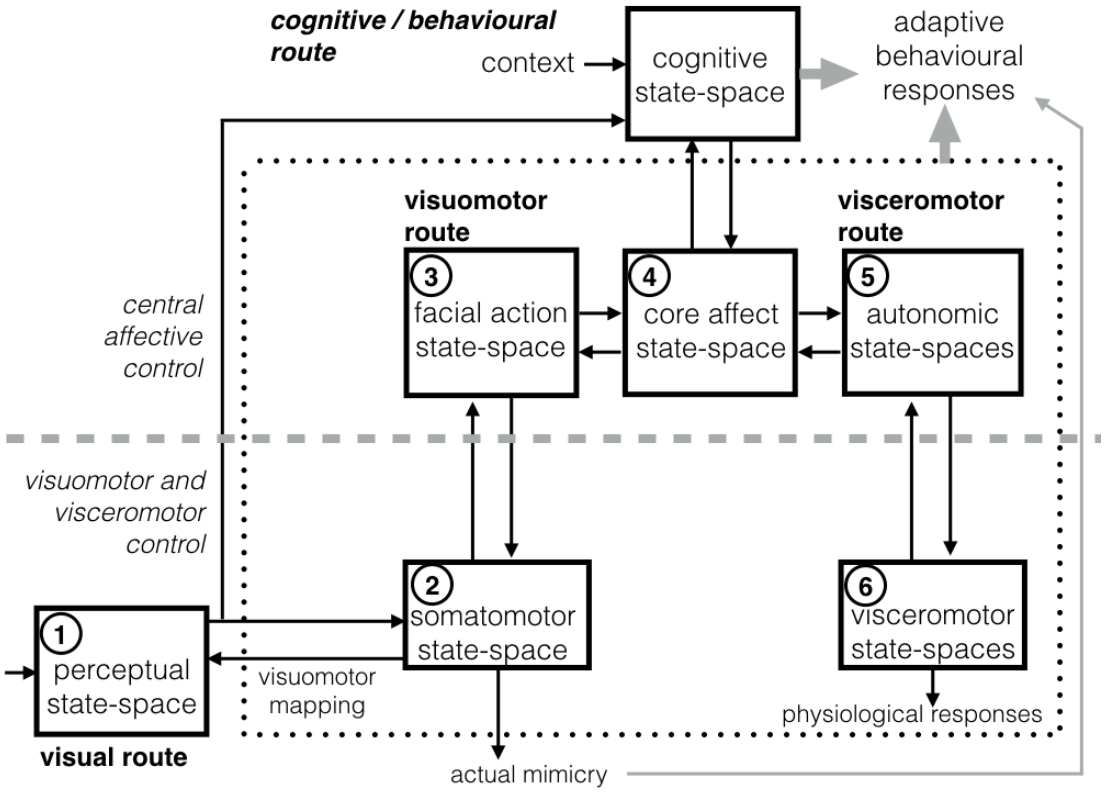


Figure 3.12: *A functional architecture for face-based emotion understanding. Numbered modules are those considered in this study. In brief, the visual system for dynamic facial expression perception interacts with an extended system, which involves the emotion system (dotted box) and high level cognitive/conceptual processes. Interaction is regulated by the visuomotor mediation of a module for action perception. The latter transforms the sensory information of observed facial actions into the observer's own somatomotor representations. The activation of the visuomotor route in turn triggers visceromotor reactions through the mediation of the core affect state-space. From there on simulation-based dynamics involving all components unfolds to support the whole process. Dashed grey lines distinguish between the hierarchical levels of control*

In forward mode, the input or control is provided top-down (TD) by an upper level state-space and the output is an emission, or prediction to a lower-level state-space. In backward mode, the input is a bottom-up (BU) observation or a prediction error, of the lower-level state.

For simplicity, but without loss of generality, we consider a discrete time system with Markovian dynamics. The latter can be written as the stochastic difference equations

$$\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t), \mathbf{c}(t+1), \boldsymbol{\epsilon}_{\mathbf{x}}(t+1)), \tag{3.1}$$

$$\mathbf{x}_{BU}(t+1) = \mathbf{g}(\mathbf{x}(t+1), \mathbf{c}(t+1), \boldsymbol{\epsilon}_{\mathbf{x_{BU}}}(t+1)), \tag{3.2}$$

Here, $\mathbf{x}(t)$ represents the state vector, and the emission $\mathbf{x}_{BU}(t+1)$ will serve as the BU measurement vector at the subsequent time step. $\mathbf{f}$ and $\mathbf{g}$ are two vector-valued functions, which are potentially time-varying and nonlinear, while $\boldsymbol{\epsilon}_{\mathbf{x}}$ and $\boldsymbol{\epsilon}_{\mathbf{x_{BU}}}$ denote the

dynamics and measurement noise, respectively. The vector $\mathbf{c}(t)$ represents the system control vector (also referred to as input). This can be shaped in many ways, for example as a function of both TD and BU signals (e.g. to introduce feedback); A TD signal can either be a suitable function of the measurement provided by an upper level state-space or an exogenous input (e.g., a labelling sequence provided along a supervised learning stage).

The dynamics can be simply described as a recursive *predict* step, where a state transition $\mathbf{x}(t) \rightarrow \mathbf{x}(t+\delta t)$ is proposed, and a *measure* step, where the next state $\mathbf{x}(t+1)$ is obtained by updating $\mathbf{x}(t+\delta t)$ through the noisy observation $\mathbf{x}_{BU}(t+1)$. In this view the state-space dynamics can be represented as the dynamic Probabilistic Graphical Model (PGM) shown in Fig. 3.13, where the predict step and recursion, are explicitly shown.

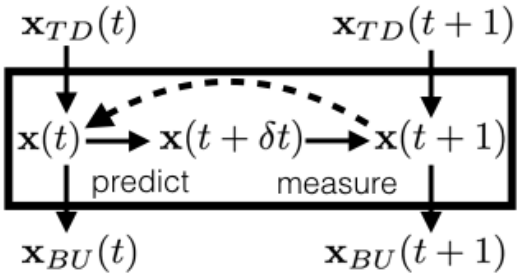


Figure 3.13: *State-space dynamics. Predict step and recursion, are explicitly shown*

Note that layers of the dynamic input-state-output models organised as in Figure 3.12, can be seen as form of hierarchical predictive coding (Wolpert et al., 2003).

More compactly, the state Eq.3.1 characterises the state transition probability $P(\mathbf{x}(t+1) \mid \mathbf{x}(t), \mathbf{c}(t))$ and thus the generative sampling

$$\widetilde{\mathbf{x}}(t+1) \sim P\left(\mathbf{x}(t+1) \mid \mathbf{x}(t), \mathbf{c}(t+1)\right) \tag{3.3}$$

Analogously, the measurement Eq. 3.2 represents the sampling

$$\widetilde{\mathbf{x}}_{BU}(t+1) \sim P(\mathbf{x}_{BU}(t+1) \mid \widetilde{\mathbf{x}}(t+1), \mathbf{c}(t+1)) \tag{3.4}$$

where $P(\mathbf{x}_{BU}(t+1) \mid \widetilde{\mathbf{x}}(t+1), \mathbf{c}(t+1))$ is related to the measurement noise (e.g., by weighing the observation $\widehat{\mathbf{x}}_{BU}$ estimated under prediction against $\mathbf{x}_{BU}$, the observed one).

In a probabilistic formulation such dynamics allows, at any level, to clearly state the inference of the hidden state at that level (inverse probabilistic model), and model parameter learning.

The overall dynamics of the observer's facial analysis is assumed to unfold as follows. At the onset (see the numbering sequence in Figure 3.12), visual facial cues are inferred initialising the *perceptual state-space* (1); a visuomotor mapping is used to couple visual cues to internal motor states of the *somatomotor state-space* (2); somatomotor dynamics is controlled by facial motor parameters, and accounts for the facial

gestures; the evolution of motor control parameters is generated within the *facial action state-space* (3), where trajectories encode facial actions; facial actions activate the *core affect state-space* (4), which mediates between the somatomotor and visceromotor components; the latter, in turn, activates the *autonomic state-space* (5), encoding visceromotor actions; the visceromotor actions govern the evolution of the *visceromotor state-space* (6) that eventually can generate physiological responses. Analogously, the evolution of the core affect can be propagated forward along the somatic route to generate actual facial mimicry.

Thereafter, simulation-based dynamics unfolds to support the process, jointly involving all the introduced components in the simulation loop, relying on forward prediction/generation and backward inference.

It is worth remarking that in this thesis we will not consider the interaction between high level cognitive and affective processes, notwithstanding we have represented a cognitive state-space in Figure 3.12 for completeness sake. Yet, during primate phylogeny, neural circuits with mirror properties and other circuits involved in mentalising might have been exapted to support more abstract cognitive functions, such as cognitive empathy (Tramacere and Ferrari, 2016). Indeed, increased activity in the anterior superior temporal gyrus and the medial prefrontal cortex are consistently reported in studies that involve some kind of social judgements such as attributing mental states and thinking about others. Thus, when watching an emotional face, observers not only participate through synchronized mimicry, they might also understand the relevance of the expresser's mental state for social interaction (Tramacere and Ferrari, 2016).

Also we are not taking into account the feedforward subcortical route from SC/Pulvinar to limbic areas represented in Figure 3.11 as light dotted lines. This certainly is an important issue and meanwhile controversial.

For instance, it is well known that visual attention can be non-consciously driven by the affective significance of visual stimuli before full-fledged processing of the stimuli. Two kinds of models have been proposed to explain this phenomenon: models involving sequential processing along the ventral visual stream, with secondary feedback from emotion-related structures or "two-stage models"; and models including additional short-cut pathways directly reaching the emotion-related structures or"two-pathway models" (Rudrauf et al., 2008).

The standard two-pathway (or dual route) models include a "retino-tectal" subcortical pathway from the retina to the amygdala, via the superior colliculus and the pulvinar involved in the rapid processing of complex visual scenes and in the rapid modulation of visual attention by emotion-related information (light dotted lines, cfr. Figure 3.11). Rudrauf et al. (2008) have tested, via simulation, which type of model would best predict real magnetoencephalographic responses in subjects presented with arousing visual stimuli, using realistic models of large-scale cerebral architecture and neural biophysics, and the results strongly support a "two-pathway" hypothesis.

However, on the one hand, alternative models can be instantiated to account for the dual route, for example long-range forward cortical-cortical pathways connecting the early visual cortices (V1/V2) with the to the anterior affective system (Rudrauf et al., 2008). The existence of these fasciculi in humans is supported by recent fiber tract analyses based on diffusion tensor imaging.

On the other hand, Pessoa and Adolphs (2010) have reviewed anatomical and phys-

iological data and argued against the notion that such a pathway plays a prominent part in processing affective visual stimuli in humans, since the visual processing of emotion stimuli occurs no faster than visual processing in the cortex in general.

Actually, in a recent work (Ceruti et al., 2017) we have addressed the problem of providing a fast, automatic, and coarse processing of the early mapping from emotional facial expression stimuli to the basic continuous dimensions of the core affect. Differently from the vast majority of approaches in the field of affective facial expression processing, we assumed and designed such a feedforward mechanism as a preliminary step to provide a suitable probabilistic prior to the subsequent stochastic core affect dynamics. However, such results have not been yet fully integrated in the model we are presenting here.

Summary

Under the rationale of a multilevel framework, stemming from Marr's analysis of computational explanations in the cognitive and behavioural sciences (Marr, 1982), we have devised an architecture of a distributed neural system for perception of dynamic facial expressions of emotion. Such architecture can be seen as an extension, or a further specification, of Adolph's original proposal (Adolphs, 2002b), and it takes into account the large body of evidence from affective neuroscience concerning the role of mirroring mechanisms along social interactions (Gallese et al., 2004; Tramacere and Ferrari, 2016).

On such basis, under a number of assumptions, a hierarchical functional architecture has been devised. In the following Chapter 4, the latter will be used to shape and constrain the Probabilistic Graphical Model at the core of our explanation at the computational theory level. Subsequently, we provide, at the implementation level of explanation, a possible realisation of the different state-spaces, which relies on the compositionality of the probabilistic graph.

CHAPTER 4

The model

ASSUME a face-to-face interaction between two agents, an *expresser* ($\mathcal{E}$) and an *observer* ($\mathcal{O}$), that share the common model underlying the state-spaces and related dynamics, as figured out in Fig. 3.12.

We are interested in reasoning about the state of the agents, in particular that of the observer $\mathcal{O}$, as it evolves over time, in terms of a system state whose value at time t is a snapshot of its relevant attributes, hidden or observed, at that time. We can model such setting in terms of a dynamic Probabilistic Graphical Model (PGM) (Section 4.1) shaped on the basis of the functional architecture previously devised, which constrains conditional independence assumptions between the graph nodes. Relying on such compositional representation, we can therefore devise an effective realisation (Section 4.2) of each sub-graph corresponding to the main functional components. The implementation will reflect one by one all the random variables (RVs) and state-spaces introduced in PGM formulation, going from the central affect level to the peripheral visuomotor and visceromotor responses. Moreover, this will be done taking into account the specificities of each module as well as the duality of the entire architecture, with the backward/inferential and forward/generative information flow.

Theoretical model

Due to the stochastic nature of the system at hand, we represent the observer's state at time t as a collection of RVs $\mathcal{X}_{\mathcal{O}}(t)$ and we denote by $\mathcal{X}_{\mathcal{O}}(t_1 : t_2)$ the random process $\{\mathcal{X}_{\mathcal{O}}(t) : t \in [t_1, t_2]\}$ indexed over the subset of reals $[t_1, t_2]$.

An assignment of values to each RV of interest, collected in $\mathcal{X}_{\mathcal{O}}$ for each relevant time t, correspond to a *trajectory* in our probability space.

We then introduce two simplifying assumptions to make the problem more treatable (see Section **??** for a remark on this).

The first simplification concerns the discretisation of the timeline into a set of time slices, that is, we take measurements of the system state at intervals that are regularly spaced with a predetermined time granularity Δ. Thus, we can restrict our set of RVs to $\mathcal{X}_{\mathcal{O}}(0), \mathcal{X}_{\mathcal{O}}(1), \cdots$, where $\mathcal{X}_{\mathcal{O}}(t)$ are the ground random variables that represent the system state at time $t \cdot \Delta$. Without loss of generality, this assumption simplifies our problem from representing distributions over a continuum of RVs to representing distributions over countably many RVs, sampled at discrete intervals.

Under such assumption, consider a distribution $P(\mathcal{X}_{\mathcal{O}}(0 : T))$ over trajectories sampled over a prefix of time $t = 0, ..., T$. We can reparametrise the distribution using the chain rule for probabilities, in a direction consistent with time:

$$P(\mathcal{X}_{\mathcal{O}}(0:T)) = P(\mathcal{X}_{\mathcal{O}}(0)) \prod_{t=0}^{T-1} P(\mathcal{X}_{\mathcal{O}}(t+1) \mid \mathcal{X}_{\mathcal{O}}(0:t)) \tag{4.1}$$

Thus, the distribution over trajectories is the product of conditional distributions, for the variables in each time slice given the preceding ones.

The second simplification entails the Markovianity of the process.

Definition 4.1 (Markov property)**.** A dynamic system over the variable $\mathcal{X}$ satisfies the Markov assumption if, $\forall t \geq 0$,

$$(\mathcal{X}(t+1) \perp \mathcal{X}(t-1) \mid \mathcal{X}(t)).$$

Such systems are called Markovian.

The Markov assumption allows us to define a more compact representation of the distribution in Eq. 4.1:

$$P(\mathcal{X}_{\mathcal{O}}(0:T)) = P(\mathcal{X}_{\mathcal{O}}(0)) \prod_{t=0}^{T-1} P(\mathcal{X}_{\mathcal{O}}(t+1) \mid \mathcal{X}_{\mathcal{O}}(t)). \tag{4.2}$$

The conditional distribution $P(\mathcal{X}_{\mathcal{O}}(t+1) \mid \mathcal{X}_{\mathcal{O}}(t))$ represents the dynamics of the system and captures the Markov assumptions that the variables in $\mathcal{X}_{\mathcal{O}}(t+1)$ cannot depend directly on variables in $\mathcal{X}_{\mathcal{O}}(t')$ for $t' < t$.

Next, we need to specify the state of the observer $\mathcal{X}_{\mathcal{O}}$. In our case, the actual inputs to the system are provided by the facial display of the expresser, in the form of a time-varying random variable (RV) $\mathbf{I}_{\mathcal{E}}(t)$, and a set of suitable control variables $\mathbf{c}(t)$, which will be used in the learning stage to account for ground-truth valence/arousal pairs.

The outputs are the facial display or actual mimicry $\mathbf{I}_{\mathcal{O}}(t)$ of the observer and a set of physiological responses $\mathcal{U}_{\mathcal{O}}(t) = \{u^{(j)}(t)\}_{j=1}^{N_V}$, namely a number N_V of measurable physiological signals $u^{(j)}(t)$ originated along the observer's autonomic activity.

Such measurable outputs are generated by the internal state of the observer that, by taking into account the phenomenological model outlined in Section 3.4, can be described by the following:

- $\mathbf{e}(t)$: the time-varying RV spanning the latent core affect state-space, at time t;

- $\mathcal{S}_M(t)$: the collection of the time-varying RVs involved in the somatic visuomotor simulation (concerning facial expression dynamics);
- $\mathcal{S}_V(t)$: the collection of the time-varying RVs involved in the visceromotor simulation (concerning autonomic activities).

More precisely, we can provide the following definitions:

Definition 4.2. The triple

$$\mathcal{X}_{\mathcal{O}}^{hidden} = \langle \mathbf{e}, \mathcal{S}_M, \mathcal{S}_V \rangle$$

represents the hidden or latent state variables of the observer.

Definition 4.3. The couple

$$\mathcal{X}_{\mathcal{O}}^{obs} = \langle \mathbf{I}_{\mathcal{O}}, \mathcal{U}_{\mathcal{O}} \rangle$$

represents the observable variables of the observer.

Remark. Note that in the actual dynamics as occurring, for example, along one experimental setting, physiological signals $\mathcal{U}$ are not explicitly observable (in the sense of visible) neither for the observer nor for the expresser. However, they are observable in the sense of being measurable through suitable sensors.

Under such basic setup, the probabilistic model of the observer (and its dynamics) is captured by the joint distribution

$$P(\mathcal{X}_{\mathcal{O}}^{hidden}(0:T), \mathcal{X}_{\mathcal{O}}^{obs}(0:T)) = \\ P(\mathbf{e}(0:T), \mathcal{S}_M(0:T), \mathcal{S}_V(0:T), \mathbf{I}_{\mathcal{O}}(0:T), \mathcal{U}_{\mathcal{O}}(0:T)). \quad (4.3)$$

The factorisation of such distribution can be simplified by introducing a set of conditional independence assertions associated with it.

The observable/hidden distinction allows to further refine the Markovian assumption over the state variables $\mathcal{X}_{\mathcal{O}}$, through the following conditional independence assumptions:

Hyp. 4.1. *The latent state variables evolve in a Markovian way*

$$(\mathcal{X}_{\mathcal{O}}^{hidden}(t+1) \perp \mathcal{X}_{\mathcal{O}}^{hidden}(0:t-1) \mid \mathcal{X}_{\mathcal{O}}^{hidden}(t)). \quad (4.4)$$

Hyp. 4.2. *The observation variables at time t are conditionally independent of the entire hidden state sequence given the state variables at time t:*

$$(\mathcal{X}_{\mathcal{O}}^{obs}(t+1) \perp \mathcal{X}_{\mathcal{O}}^{hidden}(0:t-1), \mathcal{X}_{\mathcal{O}}^{hidden}(t+1:\infty) \mid \mathcal{X}_{\mathcal{O}}^{hidden}(t)). \quad (4.5)$$

Remark. Hyp. 4.1 and Hyp. 4.2 allow to account for the affective dynamics of the observer as a state-observation model. In a state-observation model, we view the system as evolving naturally on its own, with observations of it occurring in a separate process (Koller and Friedman, 2009).

Recalling the rationales discussed in Section 3.4, we can further assume that a trajectory $\mathbf{e}(0:t)$ in the core affect state space gives rise to independent trajectories $\mathcal{S}_M(0:t), \mathcal{S}_V(0:t)$ in the somatomotor and visceromotor state-spaces, respectively. Thus,

Hyp. 4.3 (Somatomotor and visceromotor independence)**.**

$$(\mathcal{S}_M(t) \perp \mathcal{S}_V(t) \mid \mathbf{e}(t)). \tag{4.6}$$

Then, under the above assumptions the following independence properties also hold by construction:

Prop. 4.4.

$$(\mathbf{I}_{\mathcal{O}}(t) \perp \mathcal{S}_V(t) \mid \mathcal{S}_M(t)), \tag{4.7}$$

$$(\mathbf{U}_{\mathcal{O}}(t) \perp \mathcal{S}_M(t) \mid \mathcal{S}_V(t)). \tag{4.8}$$

The latter property states that the somatomotor and visceromotor state-spaces independently control the observable outputs $\mathbf{I}_{\mathcal{O}}(0:t), \mathcal{U}_{\mathcal{O}}(0:t)$, respectively.

Then, we can provide the following:

Definition 4.4. Hyp. 4.2, 4.3 and 4.4 define the set $\mathcal{I}(P)$ of conditional independence assertions that hold in the joint distribution P (Eq. 4.3).

It is straightforward, at this point, to devise a probabilistic graph $\mathcal{G}$, in the form of a dynamic direct Graphical Model, which encodes the RVs of the joint distribution P (Eq. 4.3) and the set $\mathcal{I}(P)$ defined above. The graph $\mathcal{G}$ is shown in Fig. 4.1. For simplicity, only the time slice $(t, t+1)$ is outlined.

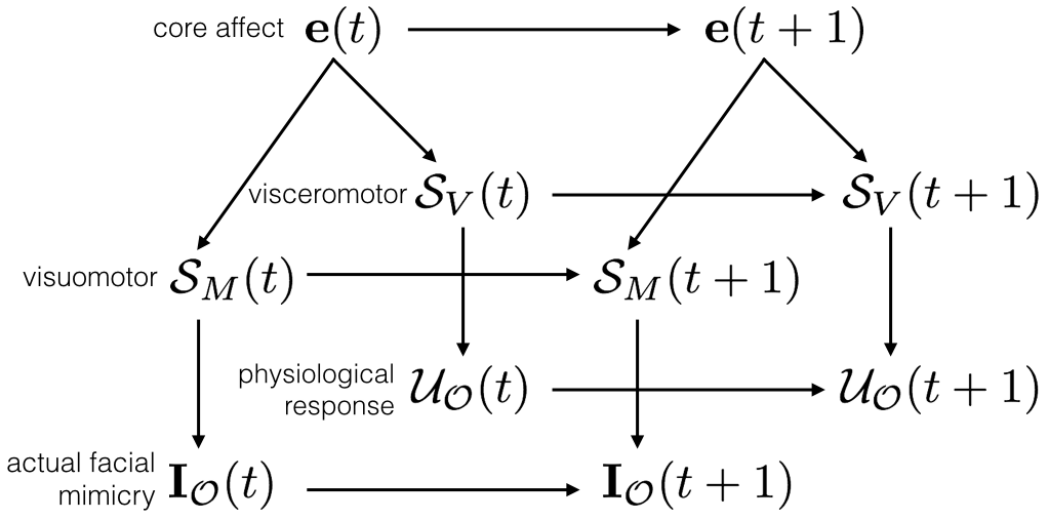


Figure 4.1: *The relations among the core components of the model (cfr. Fig. 3.12) and their generative dynamics represented as a dynamic Probabilistic Graphical Model (PGM). Graph nodes denote RVs and directed arcs encode conditional dependencies between RVs.*

By construction, $\mathcal{G}$ is an I-map (independency map, Def. A.3) for P, that is $\mathcal{I}_\ell(\mathcal{G}) \subseteq \mathcal{I}(P)$, $\mathcal{I}_\ell(\mathcal{G})$ being the set of local independencies associated with $\mathcal{G}$, (Koller and Friedman, 2009).

Eventually, the following holds:

Prop. 4.5. *Given the set $\mathcal{I}(P)$, the joint distribution P (Eq. 4.3) factorises as*

$$P(\mathcal{X}^h_{\mathcal{O}}(0:T), \mathcal{X}^o_{\mathcal{O}}(0:T)) = P(\mathcal{X}_{\mathcal{O}}(0)) \prod_{t=0}^{T-1} \begin{aligned} & P(\mathbf{I}_{\mathcal{O}}(t+1) \mid \mathcal{S}_M(t+1), \mathbf{I}_{\mathcal{O}}(t)) \times \\ & P(\mathcal{U}_{\mathcal{O}}(t+1) \mid \mathcal{S}_V(t+1), \mathcal{U}_{\mathcal{O}}(t)) \times \\ & P(\mathcal{S}_M(t+1) \mid \mathcal{S}_M(t), \mathbf{e}(t+1)) \times \\ & P(\mathcal{S}_V(t+1) \mid \mathcal{S}_V(t), \mathbf{e}(t+1)) \times \\ & P(\mathbf{e}(t+1) \mid \mathbf{e}(t)) \end{aligned} \tag{4.9}$$

Proof. By applying probability chain rule, Markov property, and conditional independence statements in set $\mathcal{I}(P)$. More simply, since by construction the graph $\mathcal{G}$ is an I-map for P, and also is a directed acyclic graph then P factorises according to $\mathcal{G}$, from Theorem A.1 (Koller and Friedman, 2009). □

Equation 4.9 provides the core probabilistic model for the agent. To further specify the complete model note, we need to precisely define the RVs collected by the time-varying state ensembles of the visuomotor and visceromotor routes $\mathcal{S}_M$ and $\mathcal{S}_V$, respectively. Note that these are subgraphs of model $\mathcal{G}$ and recall that they are conditionally independent given the current core affect state. As a consequence, if we define each subgraph as a directed acyclic graph, then the procedure used before to define the joint factorisation can be recursively applied at any level of the subgraph.

To such end, we use the following random variables related to the different state spaces.

Action state-spaces:

- $\mathbf{a}_M(t)$: the RV denoting the somatomotor action;
- $\mathbf{a}_V^{(j)}(t)$: the RV denoting the j-th visceromotor action.

Somatomotor state-space:

- $\mathbf{m}(t)$: the facial deformation due to muscle action "shared" between the two agents;
- $\boldsymbol{\theta}(t)$: the head pose parameters; in principle they need not to be shared between the agents, thus $\boldsymbol{\theta}(t) = \{\boldsymbol{\theta}_{\mathcal{O}}(t), \boldsymbol{\theta}_{\mathcal{E}}(t)\}$, though here in practice we will use the same set of pose parameters for both;
- $\mathbf{w}(t)$: the RV accounting for somatomotor state-space dynamics;
- $\mathbf{s}_{\mathcal{I}}$: a set of static parameters encoding the biometric characteristics of each individual $\mathcal{I} \in \{\mathcal{E}, \mathcal{O}\}$, namely $\{\mathbf{s}_{\mathcal{E}}, \mathbf{s}_{\mathcal{O}}\}$; expresser's $\mathbf{s}_{\mathcal{E}}$ are inferred by the observer at the onset of the interaction, while observer's parameters are given;
- $\mathbf{y}_{\mathcal{E}}(t)$: the predicted visual facial cues of the expresser.

Visceromotor state-space:

- $\mathbf{v}^{(j)}(t)$: the RV spanning the j-th visceromotor state space;

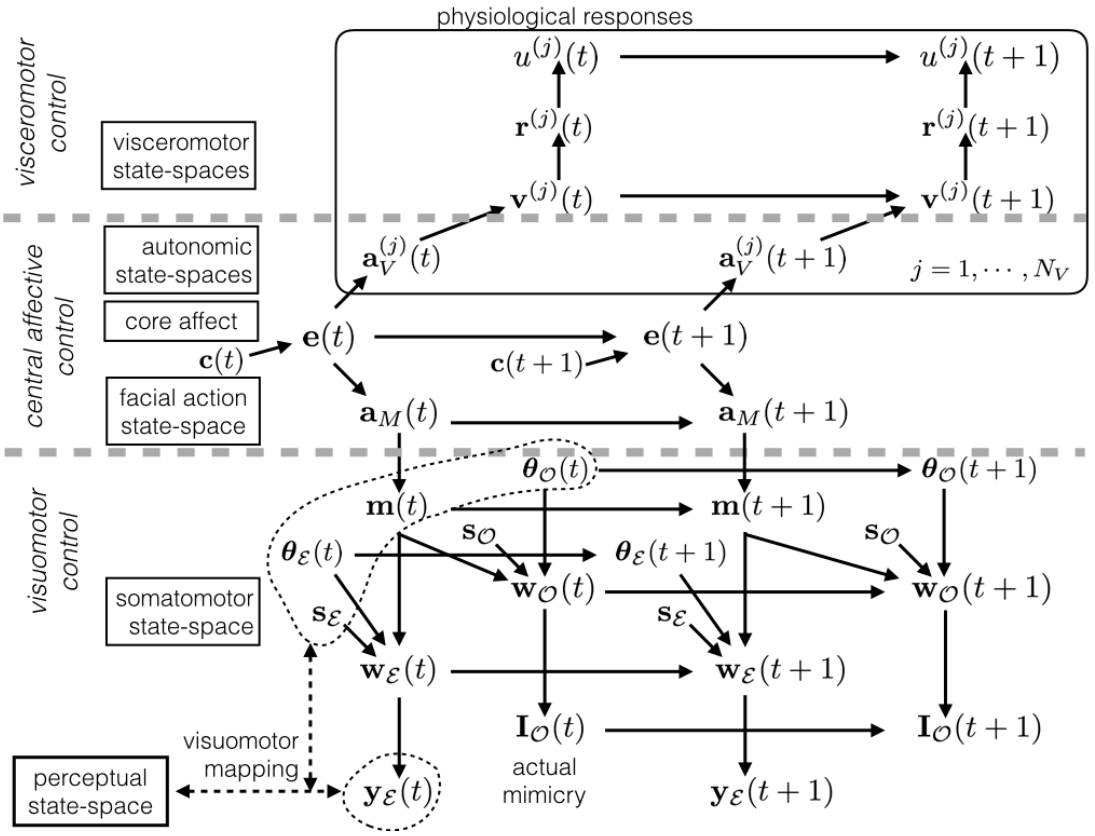


Figure 4.2: *The conditional dependencies among the core components of the model (cfr. Fig. 3.12) and their generative dynamics represented as a dynamic PGM over the time slice* $(t, t+1)$. *Dashed grey lines emphasise the hierarchical levels of control of the dynamics of the system.*

- $\mathbf{r}^{(j)}(t)$: the internal observation, or feature vector, of the j-th physiological response $u^{(j)}(t)$.

The model builds on the backward/forward hierarchical control between levels outlined in Fig. 3.12 that governs the dynamics of the whole system (Fig. 4.2). At any time t the somatomotor and visceromotor state-spaces, on the basis of their observations, provide back to the action state-spaces bottom-up inputs, namely the currently estimated motor control parameters $\widehat{\mathbf{m}}(t)$ and autonomic states $\{\widehat{\mathbf{v}}^{(j)}(t)\}_{j=1}^{N_V}$, respectively. These are used to estimate current action states $\widehat{\mathbf{a}}_M(t)$ and $\{\widehat{\mathbf{a}}_V^{(j)}(t)\}_{j=1}^{N_V}$ and, in turn, the most likely core affect state $\widehat{\mathbf{e}}(t)$. Based on this, a next-state prediction $\widehat{\mathbf{e}}(t) \rightarrow \widetilde{\mathbf{e}}(t+1)$ is obtained and, subsequently $\widetilde{\mathbf{a}}_M(t+1)$ and $\{\widetilde{\mathbf{a}}_V^{(j)}(t+1)\}_{j=1}^{N_V}$ are proposed. Then, $\widetilde{\mathbf{m}}(t+1)$ and $\{\widetilde{\mathbf{v}}^{(j)}(t+1)\}_{j=1}^{N_V}$ are sampled to provide a top-down control (filled arrows) for the somatomotor and visceromotor state-space dynamics, so that next states $\mathbf{w}(t+1)$ and $\{\mathbf{v}^{(j)}(t+1)\}_{j=1}^{N_V}$ can be actually predicted. From the latter, facial mimicry and physiological responses can be eventually generated.

It may be an helpful insight, as suggested by Minka (1999), to view the graph as a parallel machine where each state-node is a processor and links are communication paths. At each state-space level, information is propagated "horizontally" at that

level by standard forward-backward recursions; however, along the overall perception-simulation cycle, information also flows upward-downward. As a result, the state-spaces of the core affect and of the visuomotor and visceromotor routes will get shared information of the observation-measurement sequence.

This is the model in a nutshell. The latent spaces of admissible visuomotor and visceromotor actions, together with the core affect space are constructed at the learning stage. Specifically the visuomotor space of the observer $\mathcal{O}$ is learned bottom-up by observing an expresser $\mathcal{E}$ in the course of the affective interaction (a sort of mother and child interaction), whilst the visceromotor dynamics can be learned provided that $\mathcal{E}$'s physiological data are available and measured throughout the face-to-face interaction. As to the core affect space, this can be learned either in unsupervised mode or supervised. The latter requires that a continuous valence/arousal labelling of the interaction has been performed. This, in turn, conditions top-down components at lower levels.

It is worth noting that affective interactions are possible in nature since the observer and the expresser are individuals endowed with a similar brain-body system; thus, they share the key components of such interaction, e.g. visceromotor reactions along newborns' mimicry to mother's expressions mentioned before. Here, in the case of artificial agents, the same assumption holds: both the expresser and the observer rely on the same generative model. However, in this case, prewired representations and dynamics entailed by autonomic, visceromotor responses need to be learned via suitable physiological data experimentally collected.

After learning, the interaction at test time only relies upon the visible facial dynamics of a new expresser $\mathcal{E}$ as perceived by $\mathcal{O}$, who is in a specific internal physiological state. The goal is to enact within $\mathcal{O}$ a core affect dynamics similar to that of $\mathcal{E}$.

Dynamics of affect enactment

The dynamics relies on a nested, double simulation loop. The outer loop is a perception-action cycle based on the current observation of expresser's facial display. The generative properties of the model are exploited to hierarchically predict core affect states and in turn visuomotor and visceromotor states that will eventually determine facial mimicry and physiological responses. To such end, the inner loop of measurements and predictions within the central affect state-spaces implements a kind of "as if" internal simulation (see Goldman and Sripada (2005) for a discussion) to jointly optimise variables $\widetilde{\mathbf{m}}, \widetilde{\mathbf{v}}$. At the end of such inner loop, optimal $\widetilde{\mathbf{m}}^*, \widetilde{\mathbf{v}}^*$ are provided as top-down controls to motor and visceromotor state-spaces in the outer loop. Such dynamics, relying on prediction and measurement steps, is outlined in Algorithm 1 and detailed in the following.

The construction of the latent affect space model grounds in the probabilistic dependencies that relate visuomotor and visceromotor components to the core affect, namely $\mathbf{e} \rightarrow \mathbf{a}_M \rightarrow \mathbf{m}$ and $\mathbf{e} \rightarrow \mathbf{a}_V^{(j)} \rightarrow \mathbf{v}^{(j)}$, respectively.

Here the dynamics can be summarised as sampling a time dependent affect state from the latent core affect space

$$\widetilde{\mathbf{e}}(t+1) \sim P(\mathbf{e}(t+1) \mid \mathbf{e}(t), \mathbf{c}(t+1)); \tag{4.10}$$

Algorithm 1 Simulation-based dynamics

Input: - Dynamic sequence of expresser's facial display $\mathbf{l}_{\mathcal{E}}(t)$ at times denoted by $t = 1, 2, 3, \ldots$ and corresponding to multiple of frame interval Δt
- suitable initialisation of state-space parameters obtained from the training step

Output: Predictions $\widetilde{\mathbf{I}}_{\mathcal{O}}(t')|_{t'>t}$ and $\{\widetilde{u}^{(j)}(t')|_{t'>t}\}$ (Eq. 4.23)

$t \leftarrow 1$
while in interaction **do**
  {*Visuomotor mapping*}
  Given $\mathbf{I}_{\mathcal{E}}(t)$ measure $\widehat{\mathbf{l}}_{\mathcal{E}}(t), \widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t)$ (Eqs. 4.19, 4.20)
  if $t = 1$ **then**
    Measure $\mathbf{s}_{\mathcal{E}}$ and scale parameters
  end if
  $k \leftarrow 0$; $\tau_k \leftarrow t$
  {*Internal "as if" simulation*}
  repeat
    {*Backward / bottom-up measure step*}
    $\widehat{\mathbf{a}}_M(\tau_k) \sim P(\mathbf{a}_M(\tau_k) \mid \widehat{\mathbf{m}}(\tau_k))$
    $\{\widehat{\mathbf{a}}_V^{(j)}(\tau_k)\} \sim P(\{\mathbf{a}_V^{(j)}(\tau_k)\} \mid \{\widehat{\mathbf{v}}^{(j)}(\tau_k)\})$;
    $\widehat{\mathbf{e}}(\tau_k) \sim P(\mathbf{e}(\tau_k) \mid \widehat{\mathbf{a}}_M(\tau_k), \{\widehat{\mathbf{a}}_V^{(j)}(\tau)\})$
    {*Forward / top-down core affect step*}
    $\mathbf{c}(\tau_k) \leftarrow \widehat{\mathbf{e}}(\tau_k)$
    Predict $\widetilde{\mathbf{e}}(\tau_{k+1})$ (Eq. 4.10)
    {*Forward / top-down visuomotor step*}
    Predict $\widetilde{\mathbf{a}}_M(\tau_{k+1})$, then $\widetilde{\mathbf{m}}^*(\tau_{k+1})$ (Eqs. 4.11, 4.13);
    {*Forward / top-down visceromotor step*}
    Predict $\widetilde{\mathbf{a}}_V(\tau_{k+1})$, then $\{\widetilde{\mathbf{v}}^{(j),*}(\tau_{k+1})\}$ (Eqs. 4.12, 4.21)
    Save pred. $\widetilde{\mathbf{m}}^*, \widetilde{\mathbf{v}}^*$ as the current observed $\widehat{\mathbf{m}}, \widehat{\mathbf{v}}$
    $k \leftarrow k + 1$; $\tau_k \leftarrow t + k\delta t$
  until $\tau_k - t \leq \Delta t$
  Use predicted states as the current ones
  {*Forward / top-down visuomotor step*}
  Predict $\widetilde{\boldsymbol{\theta}}(t+1)$
  Predict $\widetilde{\mathbf{w}}(t+1)$ (Eq. 4.15), then $\widetilde{\mathbf{y}}_{\mathcal{E}}(t+1)$ (Eq. 4.16);
  {*Forward / top-down visceromotor step*}
  Use predicted $\{\widetilde{\mathbf{v}}^{(j),*}\}$ as control parameters and predict actual $\{\widetilde{\mathbf{v}}^{(j)}(t+1)\}$ (Eq. 4.21)
  Predict $\{\widetilde{\mathbf{r}}^{(j)}(t+1)\}$ (Eq. 4.22)
  {*Mimicry and physiological responses*};
  Predict $\widetilde{\mathbf{I}}_{\mathcal{O}}(t+1)$
  Predict $\{\widetilde{u}^{(j)}(t+1)\}$ (Eq. 4.23)
  $t \leftarrow t + 1$;
end while

then, due to local independency, sampling in parallel somatic and visceromotor actions

$$\widetilde{\mathbf{a}}_M(t+1) \sim P(\mathbf{a}_M(t+1) \mid \mathbf{a}_M(t), \widetilde{\mathbf{e}}(t+1)), \tag{4.11}$$

$$\widetilde{\mathbf{a}}_V^{(j)}(t+1) \sim P(\mathbf{a}_V^{(j)}(t+1) \mid \mathbf{a}_V^{(j)}(t), \widetilde{\mathbf{e}}(t+1)), \tag{4.12}$$

where $j = 1, \ldots, N_V$. This way a core affect trajectory $\{\mathbf{e}(t)\}_{t=1}^T$ generates specific action trajectories $\{\mathbf{a}_M(t)\}_{t=1}^T$ and $\{\mathbf{a}_V^{(j)}(t)\}_{t=1}^T$ to be taken in the somatomotor and visceromotor routes.

Note that the control variable $\mathbf{c}$ in Eq. 4.10 serves the purpose of accounting for either exogenous inputs if given - e.g. valence/arousal pairs at the learning stage -, and bottom-up feedbacks $\widehat{\mathbf{a}}_M(t), \{\widehat{\mathbf{a}}_V^{(j)}(t)\}$ that can be inferred by using posterior distributions $P(\mathbf{a}_M(t) \mid \widehat{\mathbf{m}}(t)), P(\{\mathbf{a}_V^{(j)}(t)\} \mid \{\widehat{\mathbf{v}}^{(j)}(t)\})$, respectively.

The somatic visuomotor route

A trajectory $\{\mathbf{a}_M(t)\}_{t=1}^T$ in the latent space of facial actions is used to sample and constrain a sequence of facial motor control parameters $\mathbf{m}(t)$. These tune the facial action unfolding in the motor state-space spanned by $\mathbf{w}(t)$, namely the observer's internal representation of the face. We instantiate $\mathbf{w}(t)$ as a 3D deformable shape model. Thus, $\mathbf{w}(t)$ is a vector of vertices such that the evolution of the face model at time t is represented by the ensemble of vertex state vectors $\mathbf{w}_i(t) = [X_i(t), Y_i(t), Z_i(t)]^T$. In brief, control parameters $\mathbf{m}(t), \boldsymbol{\theta}(t), \mathbf{s}_\mathcal{I}$, dynamically shape the egocentric motor representation of the agent's face, namely, $\mathbf{w}(t) := \mathbf{w}(\mathbf{m}(t), \boldsymbol{\theta}(t), \mathbf{s}_\mathcal{I})$.

The biometric parameters $\mathbf{s}_\mathcal{O}$ are used for building the observer's inner representation of his own face, whilst $\mathbf{s}_\mathcal{E}$ are adopted for predicting the evolution of expresser's facial action. The motor control parameters $\boldsymbol{\theta}(t)$ and $\mathbf{m}(t)$, the head pose and the facial deformation due to muscle action, respectively, are "shared" between the two agents. Namely, they are inferred/perceived by $\mathcal{O}$ looking at $\mathcal{E}$ along the interaction, and used as his own parameters, a process which we address as *visuomotor mapping* (Lopes and Santos-Victor, 2005).

Assume that an estimate of global head motion and face deformation parameters $\widehat{\boldsymbol{\theta}}(t), \widehat{\mathbf{m}}(t)$, respectively, are available at time t after the perceptual stage. The somatomotor space will be characterised by the following dynamics. First, sample facial action control parameters:

$$\widetilde{\mathbf{m}}(t+1) \sim P(\mathbf{m}(t+1) \mid \widehat{\mathbf{m}}(t), \widetilde{\mathbf{a}}_M(t+1)), \tag{4.13}$$

$$\widetilde{\boldsymbol{\theta}}(t+1) \sim P(\boldsymbol{\theta}(t+1) \mid \hat{\boldsymbol{\theta}}(t)). \tag{4.14}$$

Then, by using sampled control parameters, set $\mathbf{w}(t+1) := \mathbf{w}(\widetilde{\mathbf{m}}(t+1), \widetilde{\boldsymbol{\theta}}(t+1), \mathbf{s}_\mathcal{E})$, predict the facial configuration of $\mathcal{E}$,

$$\widetilde{\mathbf{w}}(t+1) \sim P(\mathbf{w}(t+1) \mid \mathbf{w}(t), \widetilde{\mathbf{m}}(t+1), \widetilde{\boldsymbol{\theta}}(t+1)), \tag{4.15}$$

and sample a predicted observation of $\mathcal{E}$'s facial cues (landmarks)

$$\widetilde{\mathbf{y}}_\mathcal{E}(t+1) \sim P(\mathbf{y}_\mathcal{E}(t+1) \mid \mathcal{T}(\widetilde{\mathbf{w}}(t+1))), \tag{4.16}$$

where $\mathcal{T}(\cdot)$ is a projection of the $3D$ model in the expresser's $2D$ visual space.

Facial mimicry is obtained by setting $\mathbf{w}(t+1) := \mathbf{w}(\widetilde{\mathbf{m}}(t+1), \widetilde{\boldsymbol{\theta}}(t+1), \mathbf{s}_{\mathcal{O}})$, using Eq. 4.15, and generating $\mathcal{O}$'s facial expression:

$$\widetilde{\mathbf{I}}_{\mathcal{O}}(t+1) \sim P(\mathbf{I}_{\mathcal{O}}(t+1) \mid \mathbf{w}(t+1), \mathbf{I}_{\mathcal{O}}(t)). \tag{4.17}$$

The observer's perception of the expresser

The goal for $\mathcal{O}$ is to infer the expresser's facial action visual cues that are then mapped onto internal parameters $(\widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t))$. To such end we need to estimate:

1. $\mathcal{E}$'s actual facial landmarks $\mathbf{l}_{\mathcal{E}}(t)$, conditioned on the set of facial patch feature responses $\mathbf{X}_{\mathcal{E}}(t)$ computed on frame $\mathbf{I}_{\mathcal{E}}(t)$, and on the currently predicted facial shape state $\widetilde{\mathbf{w}}_{\mathcal{E}}(t)$ (in the action stage);
2. the hidden motor control parameters $\boldsymbol{\theta}(t)$ (facial pose) and $\mathbf{m}(t)$ (facial deformation) that most likely modulate the visible facial configuration of the expresser.

Inference relies on the joint posterior

$$P(\mathbf{l}_{\mathcal{E}}(t), \boldsymbol{\theta}(t), \mathbf{m}(t) \mid \widetilde{\mathbf{y}}_{\mathcal{E}}(t), \mathbf{X}_{\mathcal{E}}(t), \mathbf{I}_{\mathcal{E}}(t)) = P(\boldsymbol{\theta}(t), \mathbf{m}(t) \mid \mathbf{l}_{\mathcal{E}}(t), \widetilde{\mathbf{y}}_{\mathcal{E}}(t)) \times P(\mathbf{l}_{\mathcal{E}}(t) \mid \mathbf{X}_{\mathcal{E}}(t), \mathbf{I}_{\mathcal{E}}(t)). \tag{4.18}$$

The first factor on the r.h.s of Eq. 4.18 substantiates the visuomotor mapping; the second factor supports the visual processing of facial landmarks. Hence, the perception stage boils down to the following.

1. Computing the most likely configuration of actual landmarks:
$$\widehat{\mathbf{l}}_{\mathcal{E}}(t) = \arg\max P(\mathbf{l}_{\mathcal{E}}(t) \mid \mathbf{X}_{\mathcal{E}}(t), \mathbf{I}_{\mathcal{E}}(t)); \tag{4.19}$$
2. Backprojecting into the expresser's image space the current predicted observer's facial state $\widetilde{\mathbf{y}}_{\mathcal{E}}(t) = \mathcal{T}(\mathbf{w}(\widetilde{\mathbf{m}}(t), \widetilde{\boldsymbol{\theta}}(t), \mathbf{s}_{\mathcal{E}}))$ for estimating control parameters that best explain observed landmarks $\widehat{\mathbf{l}}_{\mathcal{E}}(t)$:
$$(\widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t)) = \arg\max_{\mathbf{m}, \boldsymbol{\theta}} P(\widehat{\mathbf{l}}_{\mathcal{E}}(t) \mid \widetilde{\mathbf{y}}_{\mathcal{E}}(t)). \tag{4.20}$$

The visceromotor route

In this case, we deal with a number N_V of measurable physiological signals $\mathcal{U}(t)_{\mathcal{O}} = \{u^{(j)}(t)\}_{j=1}^{N_V}$. Typically, $u^{(j)}(\cdot)$ are 1-D time-series, that can be assumed to be the realisation of N_V independent stochastic processes. Prediction and observation steps can be performed for the $j = 1, \cdots, N_V$ spaces through the following steps:

1. predict the autonomic visceromotor state conditioned on the current action $\widetilde{\mathbf{a}}_V^{(j)}$
$$\widetilde{\mathbf{v}}^{(j)}(t+1) \sim P(\mathbf{v}^{(j)}(t+1) \mid \mathbf{v}^{(j)}(t), \widetilde{\mathbf{a}}_V^{(j)}(t+1)), \tag{4.21}$$
2. predict the observation of physiological features
$$\widetilde{\mathbf{r}}^{(j)}(t+1) \sim P(\mathbf{r}^{(j)}(t+1) \mid \widetilde{\mathbf{v}}^{(j)}(t+1)), \tag{4.22}$$
3. sample a physiological response
$$\widetilde{u}^{(j)}(t+1) \sim P(u^{(j)}(t+1) \mid \widetilde{\mathbf{r}}^{(j)}(t+1)). \tag{4.23}$$

Implementation model

The probabilistic formalisation provided in the previous section represents the foundations of the model presented in this work. As shown, the whole architecture relies upon a series of state-spaces that permits to realise the heart of the model, namely the "core affect $\rightarrow$ action $\rightarrow$ motor" hierarchy. More precisely, the construction of the latent affect space model grounds in the probabilistic dependencies that relate visuomotor and visceromotor components to the core affect, namely $\mathbf{e} \rightarrow \mathbf{a}_m \rightarrow \mathbf{m}$ and $\mathbf{e} \rightarrow \mathbf{a}_V^{(j)} \rightarrow \mathbf{v}^{(j)}$, respectively.

The requirement of the construction of the core affect latent spaces and relative dependencies is to devise an effective and efficient nonlinear mapping such that trajectories of similar control parameters and, in turn, of actions, are placed nearby in the core affect space whilst dissimilar trajectories are far away. To meet such requirement, in current work we use a Deep Gaussian Process (deep GP) approach (Damianou and Lawrence, 2013). A deep GP is a deep belief network (DBN) that aims at mapping a potentially observed input to an observed output via a cascade of multiple hidden layers of latent variables. To govern the aforementioned mapping between consecutive layers it employs Gaussian Processes (GPs). Precisely, a single layer of the deep GP can be either a standard GP or a Gaussian process latent variable model (GPLVM) (refer to the Appendix C for a wider dissertation of the grounding models).

From a general point of view, the multivariate parent node $\mathbf{Z}^{(H)} \in \mathbb{R}^{N \times Q^{(H)}}$ can be unobserved leading to an unsupervised scenario or could be considered as an observed input for a supervised learning. The intermediate hidden variables $\mathbf{Z}^{(h)} \in \mathbb{R}^{N \times Q^{(h)}}$, $h = 1, \ldots, H-1$ model the transformations between layers until reaching the observed output $\mathbf{Y} \in \mathbb{R}^{N \times D}$, placed in the leaf nodes.

Each hidden layer can be treated as a GP-LVM by employing a product of $Q^{(h)}$ independent GPs as prior for the latent mapping, from which the latent functions $\mathbf{F}^{(h)} = \{\mathbf{f}_q^{(h)}\}_{q=1}^{Q^{(h)}}$, with $f_{nq}^{(h)} = f_q^{(h)}(\mathbf{z}_n^{(h)})$, are sampled. Thus, $f_q^{(h)} \sim \mathcal{GP}(0, k^{(h)}(\mathbf{z}_i^{(h)}, \mathbf{z}_j^{(h)}))$, where $k^{(h)}(\mathbf{z}_i^{(h)}, \mathbf{z}_j^{(h)})$ is the kernel function. For the down-most layer, $\mathbf{F}^{(1)} = \{\mathbf{f}_d^{(1)}\}_{d=1}^{D}$, with $f_{nd}^{(1)} = f_d^{(1)}(\mathbf{z}_n^{(1)})$.

The generative process from the upper-most to the down-most layer is given by the following:

$$z_{nq}^{(h-1)} = f_q^{(h)}(\mathbf{z}_n^{(h)}) + \epsilon_{nq}^{(h)}, \quad q = 1 \cdots Q^{(h)} \tag{4.24}$$

$$y_{nd} = f_1(\mathbf{z}_n^{(1)}) + \epsilon_{nd}^{(1)}, \quad d = 1 \cdots D \tag{4.25}$$

where $\epsilon_{nq}^{(h)} \sim \mathcal{N}(0, (\sigma_{nq}^{(h)})^2)$ and $\epsilon_{nd}^{(1)} \sim \mathcal{N}(0, (\sigma_{nd}^{(1)})^2)$.

Clearly, the size of each latent layer is crucial but does not need to be a priori defined. It has been shown (Damianou and Lawrence, 2013) that it is possible to define automatic relevance determination (ARD) covariance functions for the GPs

$$k^{(h)}(\mathbf{z}_i^{(h)}, \mathbf{z}_j^{(h)}) = \sigma_{ARD}^2 \exp\left[-\frac{1}{2}\sum_{q=1}^{Q^{(h)}} w_q^{(h)}(z_{i,q}^{(h)} - z_{j,q}^{(h)})^2\right]$$

such that a different weight $w_q^{(h)}$ is assumed for each latent dimension. This can be exploited at the training stage in order to prune irrelevant dimensions by driving their

corresponding weight to zero thus helping towards automatically finding the structure of complex models. In other terms an automatic Occam's razor is obtained, while ending up with significantly lower number of model parameters.

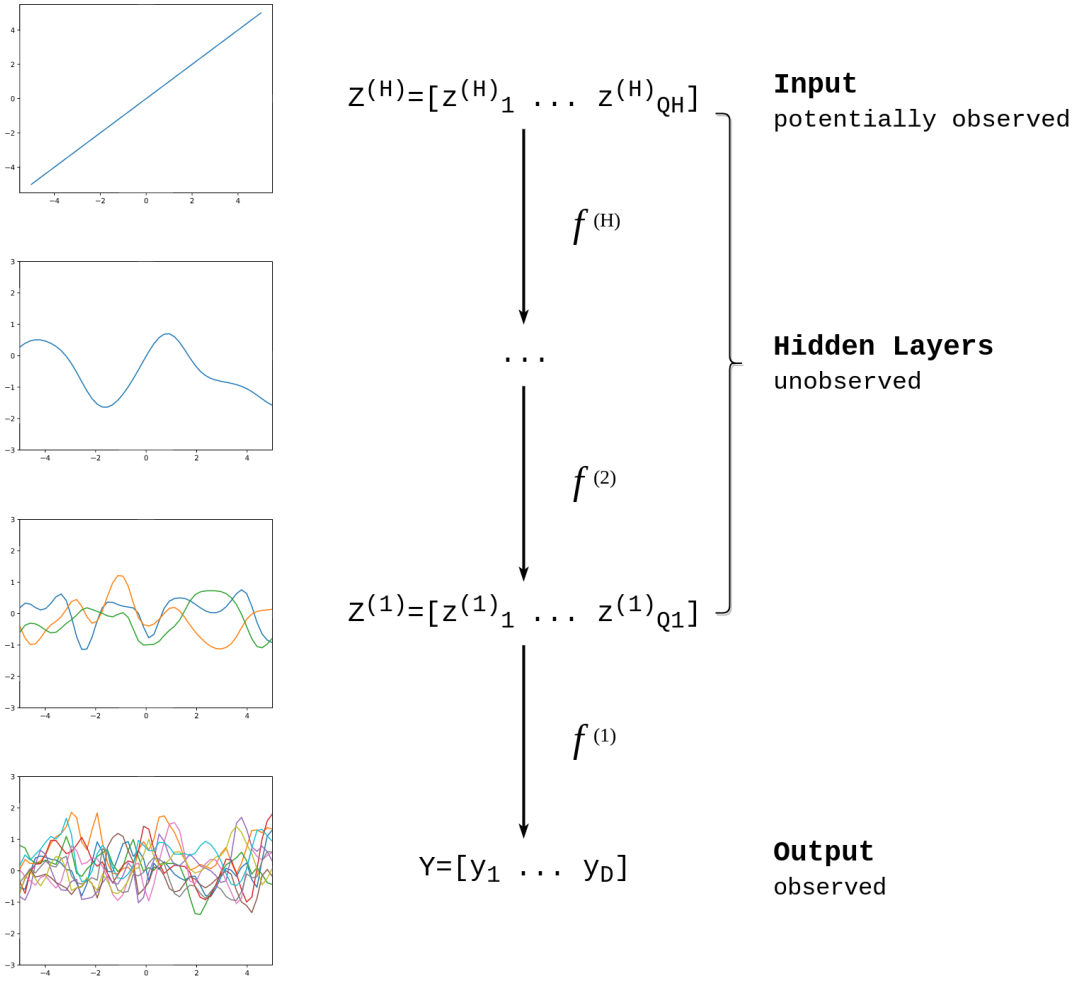


Figure 4.3: *Visualisation of a Deep Gaussian Process Model with a cascade of H hidden layers. The uppermost layer $Z^{(H)}$ is observed with a linear input and the kernel functions adopted in each layer are the squared exponential (or RBF).*

What is important for our work is that this deep structure can be naturally extended "horizontally" by segmenting each layer into different partitions. Thus, the latent space at level h is partitioned into $\pi^{(h)}$ conditionally independent subsets, matching exactly the conditional independence statements assumed from the beginning to design the PGM representation provided in Fig. 4.2. Eventually, it allows to handle in a principled and efficient way the multimodal nature of visual cues and of the different physiological signals. This can be achieved at this point by defining the down-most layer of the deep GP as a $N \times D$ matrix $\mathbf{Y} = [\mathbf{m} \mid \mathbf{v}^{(1)} \mid \mathbf{v}^{(2)} \mid \cdots]$, being N the dimension of the considered time series and D the sum of the dimensions of state/control parameters of each modality. To achieve such requirement we adopted in the down-most layer the Manifold Relevance Determination (MRD, Damianou et al., 2012), a latent variable model

which incorporates observations from several views. The Bayesian training of these layers automatically segments the latent space into private and shared parts among the considered modalities (Damianou et al., 2012), relying on separate covariance functions per view $k^{(h)(j)}$, with h indexing the layer, as usual, and j indexing the specific modality (cfr. Fig. 4.4 for an example). The ARD kernel function, introduced above, therefore requires a reformulation that takes in consideration the different view-dependent parameters to optimise, namely $\left(\sigma_{ARD}^{(h)(j)}\right)^2$ and the set of weights $\mathbf{w}^{(h)(j)} = \{w_q^{(h)(j)}\}_{q=1}^{Q^{(h)}}$.

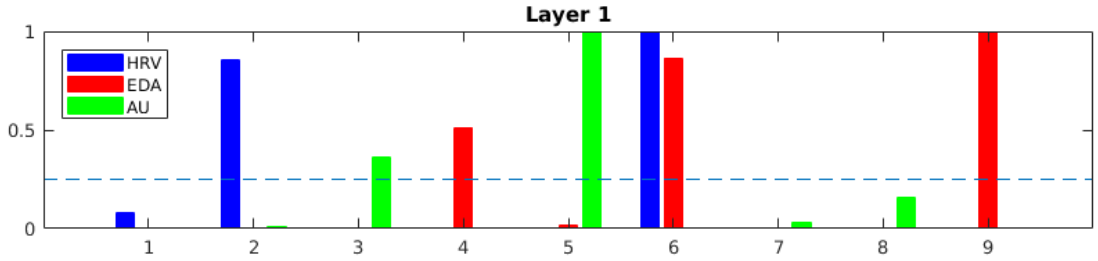


Figure 4.4: *Example of the ARD weights (vertical axis) as function of latent space dimensions (horizontal axis) resulting after training a MRD layer over multimodal data. This includes one modality obtained from facial actions (AU) and two different autonomic states, referring to heart rate variability (HRV) and electrodermal activity (EDA). Considering a threshold of* 0.25 *(dashed line) it is possible to say that two of these modalities (HRV, EDA) share the sixth dimension of the learnt latent space, while keeping private some other. This choice provides a suitable implementation model of the close exchange of information (at the state-space level) akin to the interactions occurring at the neural level between orbitofrontal cortex, the amygdala and the anterior insula.*

It is worth noting that the deep GP has the expressive power that indeed we need to map the trajectories taking place in the core affect state-space, onto trajectories at the motor state-space level. Because of the recursive warping of latent variables through the hierarchy, it allows for modelling non-stationarities and cumbersome non-parametric functional properties (Damianou and Lawrence, 2013).

In the supervised learning scenario, which is the one addressed here, the inputs to the top hidden layer $\mathbf{Z}^{(H)}$ are observed, namely they are the valence/arousal time sequences or trajectories $\mathbf{e}(t)$ provided at learning stage.

When the latent space is set up, then new estimated controls $\widehat{\mathbf{m}}, \{\widehat{\mathbf{v}}^{(j)}\}$ can be stochastically backprojected through the latent space layers up to the core affect state-space, namely,

$$\begin{aligned}
\widehat{\mathbf{a}}_M(t) &\sim P(\mathbf{a}_M(t) \mid \widehat{\mathbf{m}}(t)), \\
\{\widehat{\mathbf{a}}_V^{(j)}(t)\} &\sim P(\{\mathbf{a}_V^{(j)}(t)\} \mid \{\widehat{\mathbf{v}}^{(j)}(t)\}), \\
\widehat{\mathbf{e}}(t) &\sim P(\mathbf{e}(t) \mid \widehat{\mathbf{a}}_M(t), \{\widehat{\mathbf{a}}_V^{(j)}(t)\}).
\end{aligned}$$

This is achieved by using variational posteriors available from the model learning stage (Damianou et al., 2016), but using a bottom up sampling-like approach, in order not to disregard the uncertainty predicted at each time.

As to pose evolution formalised in Eq. 4.14, we simply put

$$P(\boldsymbol{\theta}(t+1) \mid \hat{\boldsymbol{\theta}}(t)) = \delta(\boldsymbol{\theta}(t+1), \hat{\boldsymbol{\theta}}(t))$$

hence straightforwardly exploiting the inferred $\hat{\boldsymbol{\theta}}(t)$.

Somatic motor space and visuomotor mapping

To formalise the parametric face model $\mathbf{w}(t) := \mathbf{w}(\boldsymbol{\Theta}(t))$, being $\boldsymbol{\Theta}(t)$ the vector of all involved parameters, we exploit the 3D face model Candide-3 (Ahlberg, 2010) that reasonably suits our needs. This is a 3D deformable wireframe model consisting of 113 vertices $\mathbf{w}_i$ and 184 triangles (cfr. Fig. 4.7, top row), represented by $\overline{\mathbf{w}} = [\overline{\mathbf{w}}_1 \cdots \overline{\mathbf{w}}_N]$ as a vector of vertices $\overline{\mathbf{w}}_i = \left[\overline{X}_i, \overline{Y}_i, \overline{Z}_i\right]^T$, where i indexes the i-th vertex of the model. The face shape $\mathbf{w}_{\mathcal{I}}$ can be generated from the standard mean shape $\overline{\mathbf{w}}$ which is deformed by both individual biometric characteristics and the facial action (expression) performed at time t. The evolution of the face model at time t is represented by the state vector $\mathbf{w}_i(t) = [X_i(t), Y_i(t), Z_i(t)]^T$.

Denote $\mathbf{dW}_i^S$ and $\mathbf{dW}_i^M$, biometric and facial action-based deformations. Precisely,

$$\mathbf{dW}_i^S = \left[\mathbf{dw}_{i,1}^S, \cdots, \mathbf{dw}_{i,N_s}^S\right],$$
$$\mathbf{dW}_i^M = \left[\mathbf{dw}_{i,1}^M, \cdots, \mathbf{dw}_{i,N_m}^M\right]$$

are constant $3 \times N_s$ and $3 \times N_m$ matrices, respectively, where each vector $\mathbf{dw}_{i,j}^S$ and $\mathbf{dw}_{i,k}^M$ represents the single Shape Unit (SU) and Action Unit (AU) deformation at vertex i, respectively. Here, $N_s = 14$ and $N_m = 11$. The columns of $\mathbf{dW}_i^S$ are vectors of control point displacements due to biometric traits of the individual (mouth width, eye distance, etc.) (details in Tab. 4.1). The columns of $\mathbf{dW}_i^M$ encode vectors of point displacements, each vector corresponding to AUs related to Ekman's FACS (Facial Action Coding System (Ekman and Rosenberg, 1997)) (details in Tab. 4.2); these describe the change in face geometry when the corresponding AU is enabled due to the motor activation of facial muscles. The effect of the SUs and AUs is controlled via the shape and motor/action parameter vectors $\mathbf{s}_{\mathcal{I}} = [s_1, \cdots, s_{N_s}]^T$, $\mathbf{m}_{\mathcal{I}} = [m_1, \cdots, m_{N_m}]^T$, respectively.

Under the face-to-face interaction assumption, the shape model dynamics is that of a deformable (i.e., not rigid) body (von Helmholtz, 1858) and by assuming small rotations, it can be shown that at any time t, facial movements will locally move a 3D vertex $\mathbf{w}_i$ to position $\mathbf{w}_i(t + \delta t) = \mathbf{w}_i(t) + \mathbf{dw}_i$ (for unitary time step $\delta t = 1$, without loss of generality) according to the law:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \mathbf{R}(t)\mathbf{w}_i(t) + \mathbf{dW}_i^S \mathbf{s}_{\mathcal{I}} + \mathbf{dW}_i^M \mathbf{m}(t) + \mathbf{t}(t), \tag{4.26}$$

where $\mathbf{R}$ and $\mathbf{t}$ represent the rotation matrix and the translation vector, respectively, that is the global rigid motion constrained by cranial pose dynamics. For what concerns $\mathbf{R}$, a 3D rotation can be represented by vector $\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]^T$. The exponential twist representation (Gallego and Yezzi, 2015; Murray et al., 1994) considers the rotation as the infinite limit of k rotations

$$\mathbf{R}(\boldsymbol{\omega}) = \lim_{k \to \infty} \left(\mathbb{I} + \frac{1}{k} [\boldsymbol{\omega}]_\times \right)^k = \exp([\boldsymbol{\omega}]_\times), \tag{4.27}$$

where $[\boldsymbol{\omega}]_\times$ denotes the 3×3 cross product (skew-symmetric) matrix

Table 4.1: *List of the Candide Shape Units as defined in Ahlberg (2010) and updated to version Candide-3.1.6.*

Shape Unit Vector	Name	Comment
0	Head height	Does not influence eyes, mouth, ...
1	Eyebrows, vertical position	
2	Eyes, vertical position	
3	Eyes, width	
4	Eyes, height	
5	Eye separation distance	
6	Cheeks z	Z-extension of the cheek bone
7	Nose z-extension	Z-extension of the nose
8	Nose vertical position	
9	Nose, pointing up	Vertical position of nose tip
10	Mouth vertical position	
11	Mouth width	
12	Eyes vertical difference	
13	Chin width	

Table 4.2: *List of the numerical formalised Candide Action Unit Vectors and corresponding Action Units from FACS (Ekman and Rosenberg, 1997) as defined in Ahlberg (2010), updated to version Candide-3.1.6.*

Action Unit Vector	AU (FACS)	Name
0	10	Upper lip raiser
2	20	Lip stretcher
3	4	Brow lowerer
5	2	Outer brow raiser
6	42 / 43 / 44 / 45	Eyes closed
7	7	Lid tightener
8	9	Nose wrinkler
9	23 / 24	Lip presser
10	5	Upper lid raiser
11	26	Jaw drop
14	13 / 15	Lip corner depressor

$$[\boldsymbol{\omega}]_\times = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix}, \tag{4.28}$$

such that $[\boldsymbol{\omega}]_\times \mathbf{v} = \boldsymbol{\omega} \times \mathbf{v}$, $\mathbf{v} \in \mathbb{R}^3$ being a vector to which rotation is applied, and $\mathbb{I}$ is the 3×3 identity matrix.

By expanding the matrix exponential as a Taylor series, $\exp([\boldsymbol{\omega}]_\times) = \mathbb{I} + [\boldsymbol{\omega}]_\times +$

$\frac{([\boldsymbol{\omega}]_\times)^2}{2!} + \cdots$ and using the first order approximation, we can write small incremental rotations as $\mathbf{R}(\boldsymbol{\omega}) \approx \mathbb{I} + [\boldsymbol{\omega}]_\times$. Note that, under such approximation one can conveniently represent $\mathbf{R}$ as

$$\mathbf{R} \approx \mathbb{I} + \omega_x \mathbf{G}_x + \omega_y \mathbf{G}_y + \omega_z \mathbf{G}_z, \tag{4.29}$$

where $\mathbf{G}_x, \mathbf{G}_y, \mathbf{G}_z$ are Lie algebra matrices or $SO(3)$ generators.

Eq. 4.26, applied to all vertices i, represents the state of the 3D face model evolving in time, i.e. the forward model, which is used in the action stage (Eq. 4.15). Its dynamic control parameters are the pose parameters $\boldsymbol{\theta}(t) = (\mathbf{R}(t), \mathbf{t}(t))$ and deformation parameters $\mathbf{m}(t)$. Individual biometric control parameters $\mathbf{s}_\mathcal{I}$ are considered fixed along the interaction. More precisely, observer's identity encoded by $\mathbf{s}_\mathcal{O}$ is given (practically, estimated offline), whilst expresser's parameters $\mathbf{s}_\mathcal{E}$ must be inferred through the perceptual process at the onset of the interaction.

As referred in robotics (Lopes and Santos-Victor, 2005), a visuomotor mapping defines a correspondence between perception and action of the type $\mathbf{VM} : \mathbf{F}^V \rightarrow \mathbf{F}^M$, where $\mathbf{F}^V, \mathbf{F}^M$ denote visual and motor features, in our case the inferred landmarks and the time-variant parameters controlling the internal motor state, respectively. Recall that, in order to estimate parameters $(\widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t))$, given the computed landmarks $\widehat{\mathbf{y}}_\mathcal{E}(t)$, the second step of the perceptual stage (Eq. 4.20) relies on projecting into the image space the predicted facial configuration of the expresser, namely $\widehat{\mathbf{w}}_\mathcal{E}$. The latter is obtained by sampling in the action stage the current state of the face model $\widetilde{\mathbf{w}}(\mathbf{m}, \boldsymbol{\theta})$, and assigning expresser's identity parameters, i.e. $\widetilde{\mathbf{w}}_\mathcal{E}(t) := \mathbf{w}(\widetilde{\mathbf{m}}(t), \widetilde{\boldsymbol{\theta}}(t), \mathbf{s}_\mathcal{E})$. Then, as to projection $\mathcal{T}(\widetilde{\mathbf{w}}_\mathcal{E})$ of the 3D vertices on the 2D image coordinate system, a weak perspective projection can be adopted given the small depth of the face (Orozco et al., 2013), thus

$$\widetilde{\mathbf{y}}_{\mathcal{E},l} = \mathcal{T}(\widetilde{\mathbf{w}}_{\mathcal{E},l}) = s\Pi\widetilde{\mathbf{w}}_{\mathcal{E},l} = s \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \widetilde{\mathbf{w}}_{\mathcal{E},l} \tag{4.30}$$

where s is the weak-perspective scale parameter, and l indexes the L vertices that are in correspondence with extracted facial landmarks.

Denote $\boldsymbol{\Theta} = s[1, \boldsymbol{\omega}^T, \mathbf{s}^T, \mathbf{m}^T, \mathbf{t}^T]^T$ the full parameter vector. Then, $P(\widehat{\mathbf{l}}_\mathcal{E}(t) \mid \widetilde{\mathbf{y}}_\mathcal{E}(t), \boldsymbol{\Theta})$ is used at this point to estimate parameters via maximum likelihood.

Under Gaussian noise assumption,

$$\widehat{\mathbf{l}}_{\mathcal{E},l} = \widetilde{\mathbf{y}}_{\mathcal{E},l} + \boldsymbol{\epsilon}_{\widetilde{\mathbf{y}}} \tag{4.31}$$

parameter estimation via Eq. 4.20 boils down to the negative log-likelihood minimisation problem,

$$(\widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t)) = \arg\min_{\boldsymbol{\Theta}} \frac{1}{2\sigma^2_{\mathbf{y}_\mathcal{E}}} \sum_{l=1}^{L} \|\widehat{\mathbf{l}}_{\mathcal{E},l} - \widetilde{\mathbf{y}}_{\mathcal{E},l}\|^2 + L\log(2\pi\sigma^2_{\mathbf{y}_\mathcal{E}}), \tag{4.32}$$

which can be easily solved in closed matrix form. Note that the full parameter vector $\boldsymbol{\Theta}$ needs to be estimated only at the onset of the interaction; in subsequent steps only $(\widehat{\mathbf{m}}(t), \widehat{\boldsymbol{\theta}}(t))$ are needed.

To characterise the behaviour of the observer's visuomotor simulation component *per se*, in terms of its mimicking abilities we conceived an experimental setup where, an "expert" evaluates the unfolding of the facial dynamics of the observer $\mathcal{O}$ while mimicking a number of expressers $\mathcal{E}$ different from those exploited to learn the $\mathcal{O}$'s action space. In this perspective, a human expert (e.g., a FACS certified psychologist) would compare AUs' behaviour along the interaction between $\mathcal{E}$ and $\mathcal{O}$.

To quantify AU signalling over time (Jack et al., 2014), we computed the activation probability of each AU at frame t, $P(AU_k(t))$ calculated across all expressers, and the same for related observer's responses. This was done for each of the six basic emotion categories.

Figure 4.5 illustrates the results achieved in terms of time-varying AU activation maps (each row denoting a single AU activation in time, brighter colour corresponding to higher activation). It can be noted at a glance that observer's patterns of activation, for each expression, are similar to expresser's pattern. There are however some differences that are mostly due to a lack of exact correspondence between the set $\{AU_k\}$ used by the detector, and the AU set $\mathbf{dW}^M$ considered in the Candide model. For example, AU_{12} (Lip Corner Puller) is not accounted for by our Candide implementation, thus it is never activated for any expression. Interestingly enough, associated AUs can be activated as surrogates, for instance AU_{14} (Dimpler, that forms in the cheeks when one smiles) in place of AU_{12}. This effect can be easily noticed for the "happy" expression.

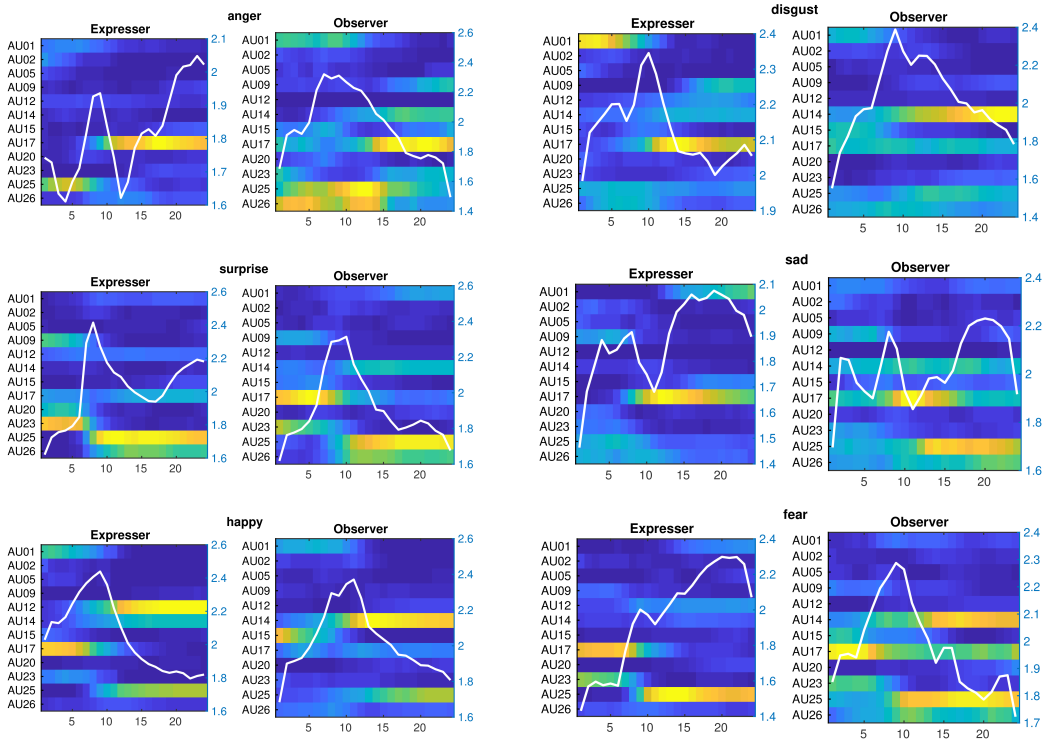


Figure 4.5: *AU activation maps for the expressers (left) and observer's related mimicry over time, for each of the six basic emotions (brighter colours for higher activations). Each row represents the activation over time of a single AU. The white line denotes time-varying Shannon's entropy $H(t)$.*

In the same vein of Jack et al. (2014), we also computed Shannon's entropy, $H(t) = -\sum_k P(AU_k(t)) \log P(AU_k(t))$, to quantify the average uncertainty of AU signalling over time. The rationale behind this choice is that signalling dynamics is an evolutionary pressure outcome to reliably transmit specific information to observers to support a near-optimal system of signalling and decoding Jack et al. (2014). For each facial expression, the $H(t)$ curve is superimposed on the corresponding activation map for both $\mathcal{E}$ and $\mathcal{O}$ (white lines in each panel of Fig. 4.5). It can be noted that $H(t)$ follows a common pattern over time (increase/decrease), but with a peculiar trend for each discrete emotion expression, which is similar for both the expresser and the observer. One notable exception is that represented by the "anger" case where the $H(t)$ behaviour in the observer looks different from the expresser.

Visual perception

To compute the $\widehat{\mathbf{y}}_{\mathcal{E}}(t)$ facial landmarks (cfr. Fig. 4.6), we adopt the Constrained Local Neural Field (CLNF) undirected graphical model (see Baltrušaitis et al. (2013), for details). The model uses the multivariate normal distribution to specify the landmark prediction probability at $L = 68$ locations $\mathbf{l}_{\mathcal{E}}(t) = [\ell_1, \cdots \ell_L]$ given patches $\mathbf{X}_{\mathcal{E}}(t) = [\mathbf{x}_1(t), \cdots, \mathbf{x}_L(t)]$ selected at frame $\mathbf{I}_{\mathcal{E}}(t)$

$$P(\mathbf{l}_{\mathcal{E}}(t) \mid \mathbf{X}_{\mathcal{E}}(t), \mathbf{I}_{\mathcal{E}}(t)) = \mathcal{N}(\mathbf{l}_{\mathcal{E}}(t); \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}) \tag{4.33}$$

where

$$\mathcal{N}(\mathbf{l}_{\mathcal{E}}(t); \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}) = \frac{\exp(\boldsymbol{\Psi}(\mathbf{l}_{\mathcal{E}}, \mathbf{X}))}{\int \exp(\boldsymbol{\Psi}(\mathbf{y}_{\mathcal{E}}, \mathbf{X})}$$

and $\boldsymbol{\Psi}(\mathbf{l}_{\mathcal{E}}, \mathbf{X})$ is the potential function used by the LNF patch expert, while the mean vector $\boldsymbol{\mu}_{\mathbf{X}}$ captures the feature extractor responses on patches $\mathbf{X}_{\mathcal{E}}(t)$ after a face detection step. A viable alternative is the method proposed by Zhu and Ramanan (2012) (or its sparse variants, e.g. Cuculo et al. (2014)) which has the merit of jointly inferring face and landmarks. However, in our experiments CLNF has proven to be more effective and computationally efficient.

Visceromotor state-space

For what concerns physiological signals $\mathcal{U}(t)_{\mathcal{O}} = \{u^{(j)}(t)\}_{j=1}^{N_V}$, in order to implement the predict and update step of Eq. 4.21, 4.22, we use an input driven linear dynamical system model (Beal et al., 2004)

$$\mathbf{v}^{(j)}(t+1) = \mathbf{A}\mathbf{v}^{(j)}(t) + \mathbf{B}\mathbf{c}^{(j)}(t) + \boldsymbol{\epsilon}_{\mathbf{v}}^{(j)}(t), \tag{4.34}$$

$$\mathbf{r}^{(j)}(t+1) = \mathbf{C}\mathbf{v}^{(j)}(t+1) + \mathbf{D}\mathbf{c}^{(j)}(t) + \boldsymbol{\epsilon}_{\mathbf{r}}^{(j)}(t+1), \tag{4.35}$$

with $\boldsymbol{\epsilon}_{\mathbf{v}}^{(j)}(t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{v}}^{(j)})$, $\boldsymbol{\epsilon}_{\mathbf{r}}^{(j)}(t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{r}}^{(j)})$, and, assuming a k-dimensional state vector $\mathbf{v}$, $\mathbf{A}$ is the $(k \times k)$ state dynamics matrix, $\mathbf{B}$ $(k \times d)$ and $\mathbf{D}$ $(p \times d)$ are the input-to-state and input-to-observation matrices, and $\mathbf{C}$ is the $(p \times k)$ observation matrix. In current implementation, the p-dimensional feature vector $\mathbf{r}^{(j)}(t)$ is obtained via the wavelet transform of the physiological response $u^{(j)}(t)$ (further details in Experiments chapter, Section 5.2).

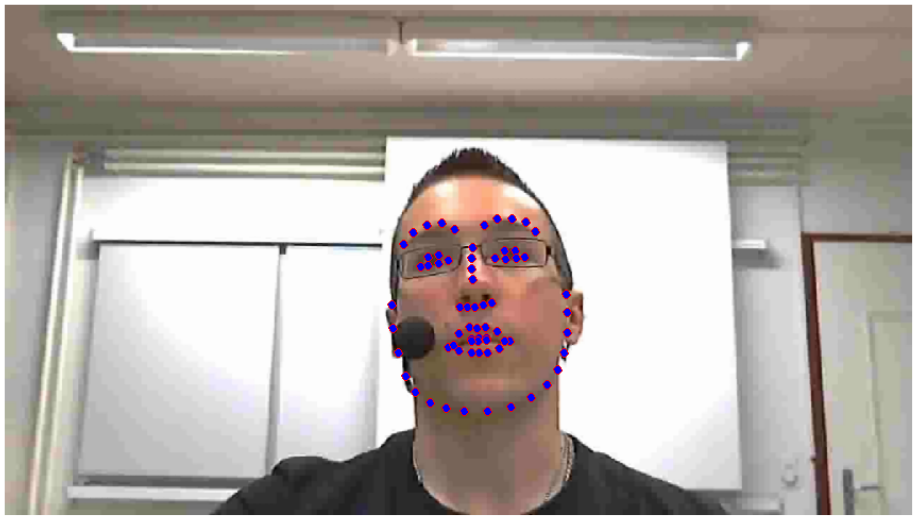


Figure 4.6: *Example of the output obtained from the landmark inference process applied to a single frame of a video sequence. It consists of $L = 68$ fiducial points.*

The input driven model is able to incorporate a displacement for the hidden state dynamics $\mathbf{c}^{(j)}(t+1) = (\mathbf{A}\mathbf{v}^{(j)}(t+1) - \widetilde{\mathbf{v}}^{(j)}(t+1))$, where $\widetilde{\mathbf{v}}^{(j)}(t+1) \sim P(\mathbf{v}^{(j)}(t+1) \mid \widetilde{\mathbf{a}}_V^{(j)}(t+1))$ is the current emission predicted by the sampled visceromotor action. Learning of visceromotor parameters $\Theta_V^{(j)} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \boldsymbol{\Sigma}_{\mathbf{v}}^{(j)}, \boldsymbol{\Sigma}_{\mathbf{r}}^{(j)})$, is performed via the variational Bayesian EM algorithm to optimize the lower bound of the log marginal likelihood

$$\ln P(\mathbf{r}_{1:T}^{(j)} \mid \mathbf{c}_{1:T}^{(j)}) \geq \mathcal{F}(Q(\mathbf{v}_{0:T}^{(j)}), Q(\Theta_V^{(j)}), \mathbf{r}_{1:T}^{(j)}), \tag{4.36}$$

with

$$\mathcal{F} = \int d\Theta_V^{(j)} d\mathbf{v}_{0:T}^{(j)} Q(\Theta_V^{(j)}) Q(\mathbf{v}_{0:T}^{(j)}) \ln \frac{P(\Theta_V^{(j)}, \mathbf{v}_{0:T}^{(j)}, \mathbf{r}_{1:T}^{(j)} \mid \mathbf{c}_{1:T}^{(j)})}{Q(\Theta_V^{(j)}) Q(\mathbf{v}_{0:T}^{(j)})} \tag{4.37}$$

and $Q(\Theta_V^{(j)}), Q(\mathbf{v}^{(j)})$ are the free distributions or variational priors (see Beal et al. (2004), for details).

At any time t the estimates of the mean $\boldsymbol{\mu}_{\mathbf{v}}^{(j)}(t)$ and covariance $\boldsymbol{\Sigma}_{\mathbf{v}}^{(j)}(t)$ of the hidden state $\mathbf{v}^{(j)}$ can be obtained via forward recursion (filtering),

$$\begin{aligned} P(\mathbf{v}^{(j)}(t) \mid \mathbf{r}_{1:t}^{(j)}, \mathbf{c}^{(j)}(t)) \propto P(\mathbf{r}^{(j)}(t) \mid \mathbf{v}^{(j)}(t)) \\ \int d\mathbf{v}^{(j)}(t-1) P(\mathbf{v}^{(j)}(t-1) \mid \mathbf{r}_{1:t-1}^{(j)}, \mathbf{c}^{(j)}(t-1)) \\ P(\mathbf{v}^{(j)}(t) \mid \mathbf{v}^{(j)}(t-1), \mathbf{c}^{(j)}(t)) \end{aligned} \tag{4.38}$$

The mean $\mu_{\mathbf{v}}^{(j)}(t)$ is provided at inference time in input to the j-th action state-space as the observation of the current visceromotor state $\mathbf{v}^{(j)}(t)$.

Facial mimicry and physiological signal generation

The quality of facial mimicry animation is not a crucial concern in our current research work, the model being agnostic to the adopted output. Here, for visualisation purposes, Eq. 4.17 is intended to drive both a synthetic avatar and the simulator of the well known iCub humanoid robot (see Fig. 4.7). In the first case, mimicry is obtained by simply warping the observer's facial image in frontal pose. Observer's biometric parameters $\mathbf{s}_{\mathcal{O}}$ and sampled $\widetilde{\mathbf{m}}(t)$ are used to generate observer's facial state $\mathbf{w}_{\mathcal{O}}(t)$ in a central pose. In Fig. 4.8 is shown the incremental generation of a facial mimicry, starting from the basic shape $\mathbf{w}_i$, applying the observer's shape parameters $\mathbf{s}_{\mathcal{O}}$, and adding the observed motor ones $\mathbf{m}(t)$. The latter is projected into the image space and based on the projected mesh, a piecewise affine transformation is applied to the observer's face (Dornaika and Davoine, 2008).

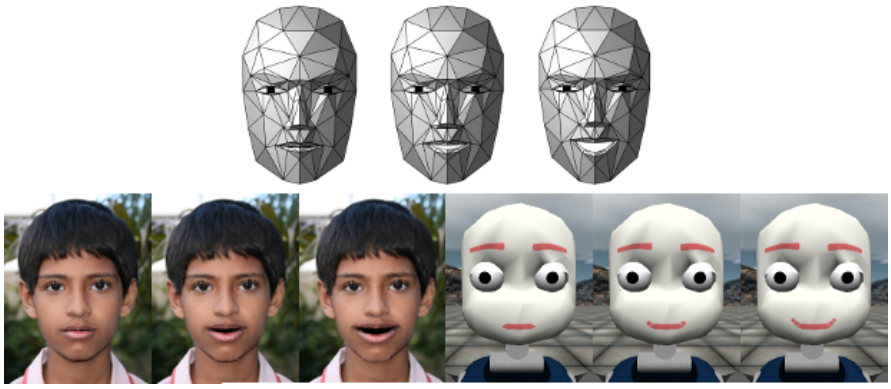


Figure 4.7: *Examples of facial mimicry. The top row shows the motor state-space dynamics $\mathbf{w}_{\mathcal{O}}(t)$ (Candide-3 model). Bottom row shows the corresponding image morphing and iCub simulator evolution.*

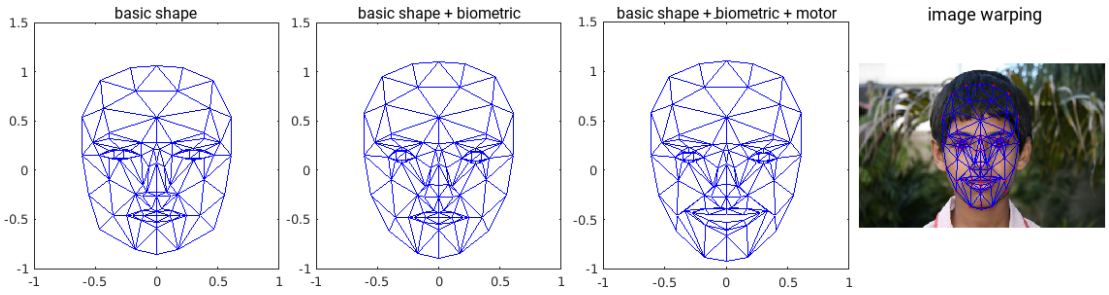


Figure 4.8: *The main steps of a facial mimicry process: starting from the basic shape, applying the observer's shape parameters, adding the action component and finally realise an image warping of the observer's neutral face image.*

As to the iCub Simulator, facial regions are modelled as subsystems handling precise face parts, named *ports*, corresponding to simulated LEDs. In particular, four subsystems are conceived: left eyebrow, right eyebrow, mouth and eyelids. For each of them an hexadecimal number defines the related port activation. Facial mimicry is obtained by converting AUs to these subsystem handles.

As to physiological sampling, by inverting the wavelet transform we reconstruct a chunk of the physiological responses of length equal to that given as model input. In this work we do not focus on the actual usage of such signals, so we simply plot their dynamics, but they could potentially be used to simulate the autonomic behaviour of a virtual agent or bio-inspired robot, as it happens in the 'sweating humanoid' Kengoro (Kozuki et al., 2016).

Modelling assumptions

In our model, we rely on the Markov assumption for the underlying stochastic processes and a discrete time approximation of their continuous dynamics. The Markov assumption is a subtle issue. Quoting (Van Kampen, 1992), "*When a physicist talks about a process, he normally refers to a certain phenomenon involving time. Concerning a process defined in this way it is meaningless to ask whether or not it is Markovian unless one specifies the variables to be used for its description. The art of the physicist is to find those variables that are needed to make the description (approximately) Markovian.*" More precisely, in our case we rely on the markovianity of the *hidden* variables governing the core affect dynamics. Beyond the scope of limiting the complexity of the model while dealing with nice mathematical properties, this choice has been proven to be a reasonable representation of the process when modelling empirical data. As remarkably shown by Kuppens et al. (2010), observable V/A trajectories from a single subject can indeed be modelled via an Ornstein-Uhlenbeck (OU, Uhlenbeck and Ornstein (1930a)) state-space markovian diffusion process. Though, we will not address this issue here, nothing prevents from extending the model, with a reasonable effort even at the implementation level, in order to account for the dynamics of the *observable* variables (physiological signals and the visible facial expression sequence) in terms of higher order autoregressive dynamics; one recent example is provided by Mattos et al. (2017).

As to discretisation, this certainly rise from the fact that, practically, we are dealing with sampled signals (video frames, physiological time series). However it is important to note that the discrete representation of the stochastic differential equations modelling the stochastic process evolution has a deeper connection to the proper integration of such equations. More formally (but see Mahnke et al. (2009)), given a stochastic process $X = \{X_t, 0 \leq t \leq T\}$ under the assumption that it is a Markov process with an infinitesimal generator, then its dynamics is specified in the form of the Itô SDE

$$dX(t) = f(X(t), t)dt + D(X(t), t)dW(t)$$

Such compact form should be more correctly interpreted in integral form as

$$X(t) - X(t_0) = \int_{t_0}^{t} f(X(\tau), \tau)d\tau + \int_{t_0}^{t} D(X(\tau), \tau)dW(\tau)$$

Note that, whilst the first integral is a conventional Riemann integral or a Riemann-Stieltjes integral if the function $f(X(\tau), \tau)$ contains steps, the second integral is a stochastic Stieltjes integral of the type $\int_{t_0}^{t} G(\tau)dW(\tau)$ whose evaluation requires some caution. To such end, the integral is partitioned into n time subintervals, which are separated by the points t_i with $t_0 \leq t_1 \leq t_2 \cdots \leq t_{n-1} \leq t$, and intermediate points τ_i are

defined within each of the subintervals $t_{i-1} \leq \tau_i \leq t_i$ and crucially the value of the integral depends on the position chosen for t_i within the subintervals. Semimartingales and in particular local martingales, are the most common stochastic processes that allow for straightforward application of Itô's formulation to stochastic calculus. Itô chooses $\tau_i = t_{i-1}$ and this leads to $\int_{t_0}^{t} G(\tau)dW(\tau) = \lim_{n\to\infty} \sum_{i=1}^{n} G(t_{i-1})(W(t_i) - W(t_{i-1}))$, where the limit is a mean square limit (the standard limit in Hilbert space theory). In the Itô approach, the starting point is the general Itô difference equation

$$\Delta X(t) = f(X(t), t)\Delta t + D(X(t), t)\Delta W(t)$$

where Δt and $\Delta W(t)$ are the time interval and the random increment, respectively. Then, equal time intervals are chosen to have $t_k = k\Delta t + t_0$, with $X_k = X(t_k)$, $\Delta X_{k-1} = X_k - X_{k-1}$, $\Delta W_{k-1} = W_k - W_{k-1}$ for the integration to find the general solution (Euler-Maruyama representation)

$$X_n = X_0 + \sum_{k=0}^{n-1} f(X_k, t_k)\Delta t + \sum_{k=0}^{n-1} D(X_k, t_k)\Delta W(t)$$

in the limit $n \to \infty$. In this sense the Euler-Maruyama approximation is not only a discrete time form of the continuous SDE, but a proper representation of its correct integral formulation. More practically, the choice of Δt in our case is basically dictated by the video frame sampling rate of the visible facial expression. The situation is quite more cumbersome in the case of physiological signals where a suitable "window of observation" must be chosen and synchronised with the video rate, and it may not coincide with the physiological signal sampling rate (which, in turn, is independent of the video sampling rate). These issues will be discussed more specifically in the experimental Section 5.2.

Summary

In the current chapter we presented a computational model involved during a face-to-face interaction, shared between an hypothetical *expresser* ($\mathcal{E}$) and an *observer* ($\mathcal{O}$). The description, mainly focused on the state of the observer agent, went through the definition of a series of random variables (RV) responsible of the perception and generation of facial expressions. This task has been first tackled in terms of Probabilistic Graphical Model (PGM) and then directly derived in a possible implementation. The latter is composed by different modules which are the results of so many implementation choices that are instrumental in our experiments, while not affecting the generality of the model. All the considered parts will be be put together in Chapter 5 where two sets of experiments will test their effectiveness.

CHAPTER 5

Experiments

THE focus of the experiments presented in this chapter is on assessing the hypothesis that the simulation-based mechanism together with the extra autonomic activity information available during learning can improve the analysis of facial expressions when only visual information is available. The experiments will also evaluate the goodness of the implementation choices, by addressing the task of signal generation after an affective interaction (a sort of mother and child interaction) that realises the learning stage. To sum up, two sets of experiments are taken into account:

1. the generation of both physiological signals and visible cues given a learnt core affect state space;

2. the inference of valence/arousal in an observer, on the basis of his autonomic activity information and the visual cues of an expresser.

Datasets

As outlined in Chapter 2, early research in affective computing was mostly focused on facial expression recognition and, by and large, datasets were built on a single modality, namely images or videos (Calvo and D'Mello, 2010; Sariyanidi et al., 2015). To truly capture the multidimensionality of emotion, in the last fifteen years the number of public repositories has grown larger, where behavioural data gathered in realistic and natural setting experiments have been recorded by multiple modalities (Picard et al., 2001; Calvo and D'Mello, 2010; Nicolaou et al., 2011; Gunes and Schuller, 2013; D'Mello and Kory, 2015; Poria et al., 2017).

To cope with the requirements introduced by the proposed model, the experiments have been conducted referring to two public available datasets, namely RECOLA (Ringeval

et al., 2013) and AMHUSE (Boccignone et al., 2017). In both these datasets, emotions have been considered in terms of continuous time and continuous valued dimensional affect, namely arousal and valence. Nowadays, indeed, research interest is moving more and more towards a continuous emotional space (Fontaine et al., 2007) and the measure of emotion is often shaped in terms of these two parameters, already introduced in Section 2.4.2.

The Audio Visual Emotion Recognition Challenge (AVEC (Ringeval et al., 2015b)) has been proposed, focusing on affect analysis as a regression problem. The RECOLA database, deployed for this challenge (Ringeval et al., 2013), addresses multimodal data recordings in the context of spontaneous collaborative and affective interactions in French. Participants were recorded in dyads during a video conference while completing a task requiring collaboration. The dataset includes multimodal data, i.e., audio, video, ECG and EDA as well as the result of continous annotations performed by 6 French-speaking raters (3 males and 3 females). The two affective dimensions (arousal and valence) were annotated separately and time-continuously, using a slider with values ranging from -1 to +1 and a step of 0.01. Results from the raw data show that the internal consistency between annotators, measured in terms of Cronbach's α (Cronbach, 1951), is acceptable for valence ($\alpha = 0.74$) and good for arousal ($\alpha = 0.80$).

In the same vein, we created the multimodal dataset AMHUSE (Boccignone et al., 2017), conceived following such research trend, but focusing on the positive emotion of amusement. AMHUSE dataset involved 36 subjects in order to record the reactions in presence of 3 amusing and 1 neutral video stimuli. The participants were 9 females and 27 males, with an age varying from 18 to 54 years old ($\mu = 26.7$ and $\sigma = 8.8$). Gathered data include RGB video and depth sequences along with physiological responses: electrodermal activity, blood volume pulse, temperature (cfr. Fig. 5.1). We also cope with the labelling problem, by proposing and exploiting DANTE (Dimensional ANnotation Tool for Emotions) to produce video annotations continuously and not on a frame-by-frame basis. The goal was to provide a suitable tool for continuous affect state modelling and benchmarking. The recorded videos were, therefore, annotated by a team of four annotators (3 males and 1 female) in terms of valence and arousal continuous dimensions. The frequency of annotations is 25 Hz with values ranging from -1 to 1 and a step of 0.001. Multiple analyses were performed on the annotations to assess the agreement between annotators. Namely, the Cronbach's α, the mean Pearson's correlation coefficient and the Concordance Correlation Coefficient (CCC) (Lawrence and Lin, 1989). Results indicate a good inter-rater reliability for the valence, whilst, as expected, a poorer value for arousal is observed. Indeed, it is well known that the level of arousal is more difficult to distinguish than valence, resulting in a lower agreement between the annotators. In particular, the valence's value of Cronbach's α is 0.84 and falls in the range $0.8 \leq \alpha < 0.9$, which is to be considered a good internal consistency, while is equal to 0.33 for the arousal. The mean correlation coefficient, instead, is equal to 0.31 for arousal and 0.74 for valence. The CCC reflects such trend, giving 0.09 for the arousal and 0.51 for the valence. Both the dataset and the annotation tool are made publicly available for research purposes.

Note that even if the context of this last dataset does not include the interaction with other humans/agents, the data are still suitable for the experiments. Indeed, the simulated interaction experiments as described later in Section 5.3 is done always fo-

cusing on the visual perception of a single side of the interaction. The generation of an hypothetical 'external response', in fact, would require the modelling of the interaction between high level cognitive and affective processes (precisely, Adolphs (2002b)'s stage 3) that are out of the scope of this work.

However, since we are mostly dealing with a low level stage of perception emotion understanding, we assume that such "inner kernel" of affect perception, mostly developed in early dyadic interaction (e.g., mother and infant interaction) is later adapted and exploited to ground perception of emotions along triadic interactions (the observer watching the expresser who, in turn, reacts to the perception of a third event).

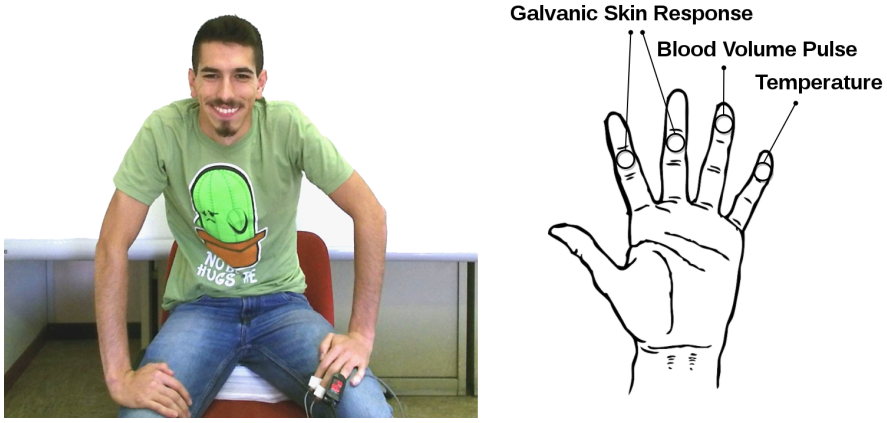


Figure 5.1: *Experimental setup (Left) and placement detail of the physiological sensors used during the acquisition process of AMHUSE dataset (Right).*

In Tables 5.1-5.2 a brief comparison of the considered dataset is reported. In particular, there are shown the corresponding subject cardinality, the recorded signals, the extracted visual features as well as the provided emotional annotations.

Table 5.1: *Recorded signals in each dataset. In brackets the number of complete data. ECG = electrocardiogram, BVP = blood volume pulse, EDA = electrodermal activity, SKT = skin temperature.*

Dataset	Subjects	Audio	RGB	Depth	ECG	BVP	EDA	SKT
RECOLA	46(18)	✓	✓	-	✓	-	✓	-
AMHUSE	36	-	✓	✓	-	✓	✓	✓

Table 5.2: *(Left) Extracted visual features provided in each dataset. FP = Fiducial Points, Pose = Head pose, AU = Action Units. (Right) Emotional annotations provided in each corpus. S = Self report, E = External. In brackets the number of external annotators. C = Continuous, VA = Valence, Arousal.*

Dataset	FP	Pose	AU
RECOLA	-	✓	✓
AMHUSE	✓	-	✓

Annotator	Type	Emotion space
E(6) + S	C	VA + 5 tags
E(4) + S	C	VA

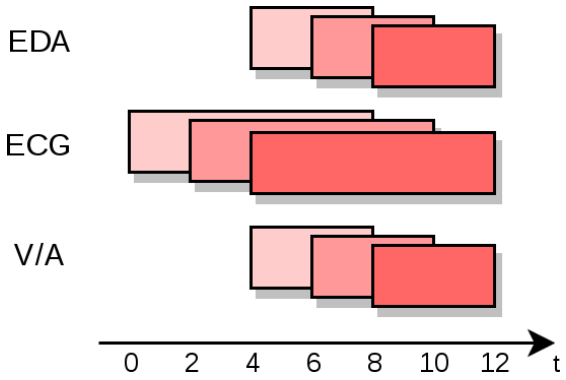


Figure 5.2: *Visualisation of the considered sampling windows for the electrodermal and electrocardiogram data present in RECOLA dataset, as well as for continuous emotional annotations.*

Physiological signal processing

The physiological signals considered in the model vary according to the availability in the dataset adopted in each experiment. The widest considered subset of such signals $u^{(j)}(t)$ includes: electrodermal activity (EDA), the skin temperature and the heart rate variability (HRV) derived from the ECG.

The processing of these signals starts with a step that aims at carving out additional properties from raw data and suppressing the unwanted aspects (e.g. noise, external trends, artefacts). Since these kind of signals are dynamic and exhibit time-varying statistics in both the time and frequency domains, in all experiments we extract the features $\mathbf{r}^{(j)}(t)$ (Eq. 4.35) using discrete wavelet transform (DWT). This approach allows for the analysis of non-stationary signals at multiple scales making use of an analysis window to extract signal segments. The length of such moving windows strictly depends on the nature of the signal and its fluctuations in the time domain, in order to permit the extraction of a discriminative behaviour. As pointed by Levenson (1988); Ringeval et al. (2015a), it does not exist yet in literature a prominent theory regarding the best length of temporal window to choose for a specific signal and emotion. The length of the analysis window can vary greatly according to the modality, and specifically for the ones activated by the autonomic nervous system it should be noted that they may remain activated for long time.

In Fig. 5.2 it is shown an example of the considered window for the EDA and ECG data of the RECOLA dataset. It worth notice that the 50% overlap between windows is commonly used to increase the number of windows and to mitigate the "loss" at the edges of the window, which can provide a more accurate estimate of the fluctuation function especially for the long-time-scale windows.

After pre-processing stage, including signal segmentation, we select empirically a suitable level of Daubechies 3 (db3), namely the last level for which at least one coefficient is correct, following

$$(N_w - 1) * 2^L < N$$

where N_w is the length of the decomposition filter associated with the chosen mother wavelet, L is the chosen level and N is the signal length. The constraint above results

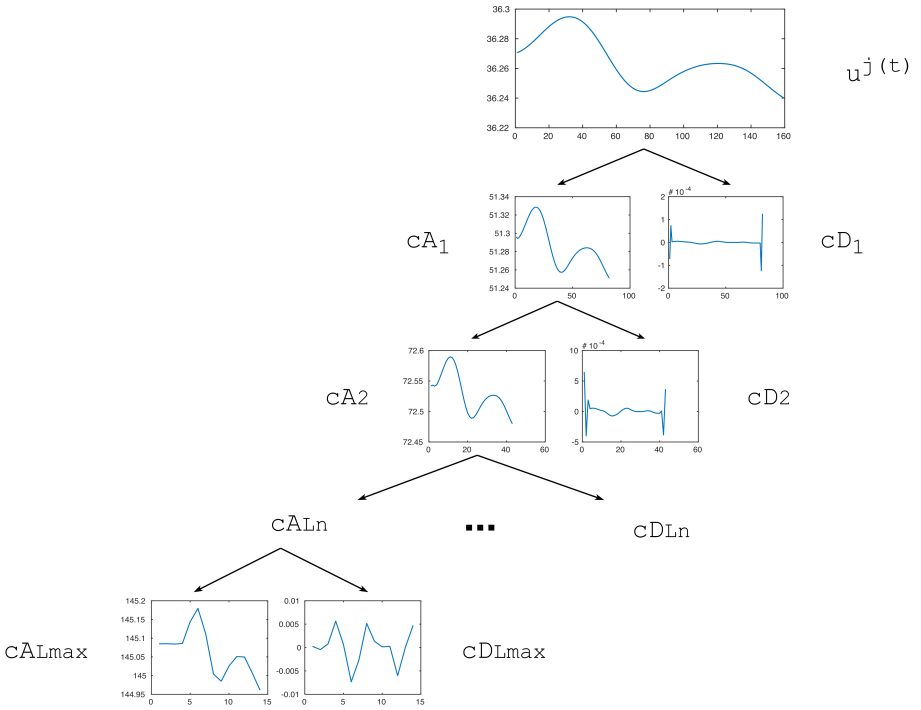


Figure 5.3: *The Daubechies 3 ($N_w = 6$) wavelet decomposition of a de-noised temperature signal chunk ($N = 160$) from AMHUSE dataset. In this case $L_{max} = 5$ and the considered features $r^{(j)}(t)$ become $cA_{L_{max}}$.*

in

$$L_{max} = \log_2(N/(N_w - 1))$$

and, after extracting the coefficients, we retain only the approximate ones as feature vector for the physiological data, such that $r^{(j)}(t) = cA_{L_{max}}$ (cfr. Fig. 5.3).

Electrodermal activity (EDA)

The electrodermal activity is a measure of the electrical skin resistance in presence of sweat produced by the body. More precisely, when a high condition of sweating occurs, the electrical skin resistance drops down. A dryer skin produces essentially higher resistance. Emotions with a prominent presence of positive or negative arousal, such as excitement, stress or fear can induce fluctuations of skin conductivity (Lang et al., 1993; Nakasone et al., 2005). A typical signal of this nature presents two main additive components: a slowly changing tonic part, referred as skin conductance level (SCL) and a phasic skin conductance response (SCR) characterized by rapidly changing peaks associated with short-term stimulus. In order to quantify the SCR amplitude, a decomposition process over the original EDA signal is needed. The adopted approach relies on the assumption that SCRs are caused by discrete episodes of sudomotor bursts that can be approximated by an appropriate impulse response function (IRF). In particular, its dynamic can be modelled by a two-compartment diffusion model, namely

Figure 5.4: *Visualisation of the Ledalab software used to preprocess the electrodermal signal and to apply the Continuous Decomposition Analysis (CDA) for phasic skin conductance extraction.*

the 'poral valve model' (Edelberg, 1993) where the sweat is released to the first compartment (sweat duct), floats to the second compartment (corneum) and is eliminated by evaporation from this. The Bateman bi-exponential function proposed by Alexander et al. (2005) well describes the diffusion process, and is defined as:

$$b(t) = c\left(e^{\frac{-t}{\tau_1}} - e^{\frac{-t}{\tau_2}}\right), \tag{5.1}$$

where τ_1 measures the steepness of rise and τ_2 its decay, while c is a constant term for the gain. The deconvolution of original EDA signal with the IRF above permits to extract the SCR components. A straightforward implementation of this process is provided by Ledalab, shown in Figure 5.4, a specific software for the analysis of skin conductance data. The implementation of this feature, in particular, relies on the work introduced by Benedek and Kaernbach (2010) and known as Continuous Decomposition Analysis (CDA). This is preceded by a classical signal preprocessing for noise reduction, including a low-pass Butterworth filter and an adaptive Gaussian smoothing, as well as an optimization step to evaluate the parameter values ($\tau_{1,2}$) that depend on inter-individual differences in skin characteristics.

Skin temperature (SKT)

Skin temperature, present only in the AMHUSE dataset, is acquired via a body temperature sensor placed on the little finger of the non-dominant hand, and recorded in terms of degrees in Celsius scale ($^{\circ}C$). Although the studies in this regard are not so many, it is proven that body temperature changes according to different emotional states. Plutchik (1956) reports that "*embarrassment, resentment, conflict, depression,*

anxiety, guilt, and "sudden elation" were all associated with drops of finger temperature ; erotic excitement without conflict, reassurance, and relaxation were associated with rises of skin temperature". Variations in the skin temperature are the main effects of the changes in blood flow caused by 'vascular resistance'. This is the resistance that must be overcome to push blood through the circulatory system and create blood flow. This is modulated by the effect of contraction of the muscles mediated by the sympathetic nervous system. Thus, it is evident that the skin temperature variation directly reflects autonomic nervous system activity and is another effective indicator of emotional status.

Electrocardiography (ECG)

Electrocardiography is the process of recording electrical activity of the heart, typically involving electrodes displaced on the skin. Its tracing consists of a sequence of well known patterns, including a P wave (atrial depolarization), a QRS complex (ventricular depolarization) and a T wave (ventricular repolarization), the main components of a single cardiac cycle, namely an heartbeat. The time distance between two successive R peaks is referred as RR Interval (RRI). This feature, and in particular the observation of the trend in the number of R peaks, represents the basis of most of the analysis carried out on the electrocardiogram signal. Indeed, the amount of complete heartbeats in a specific time window, referred as heart rate (HR), is closely related to emotional arousal and linearly depends on the activity of the sympathetic and parasympathetic nervous systems. Raw ECG tracing, anyway, requires standard preprocessing to filter out noises and respiration trends. The first ones, as in the case of EDA, are typically induced by power line interferences in the recording instrumentation, as well as loss of contact between the electrodes and the skin or motion artefacts. The respiration, on the other side, introduces a baseline wander in the signal that may causes problems in the detection of peaks. Such wander are characterised by a low frequency trend, that can be easily removed adopting an high pass filter or a median filtering. The main steps of the ECG preprocessing can be, therefore, summarised in three main consecutive steps:

1. de-trend and de-noise the raw ECG signal;
2. detect subsequent RR peaks looking at local maxima;
3. measure the time distance between two consecutive R peaks.

This procedure is summarised in Figure 5.5, where a sample gathered from RECOLA dataset is being processed. At the end of the aforementioned basic steps, the amplitude of the sample is computed as the inverse of the time difference between consecutive R peaks and is placed at the instant of the second R peak.

In the case of AMHUSE dataset, the recorded signal is the result of a blood volume pulse (BVP) sensor placed on a finger. This is a non invasive mean to obtain an indirect measure of the heart rate, via the arterial oxygen saturation of hemoglobin. This information, obtained via photoplethysmography (PPG), consists in sending two lights at different wavelengths, namely $660nm$ (red light spectrum) and $940nm$ (infrared light spectrum). The first wavelength is absorbed by deoxyhemoglobin (Hb) and the second by hemoglobin (HbO_2), which together affect the blood stream. By considering the

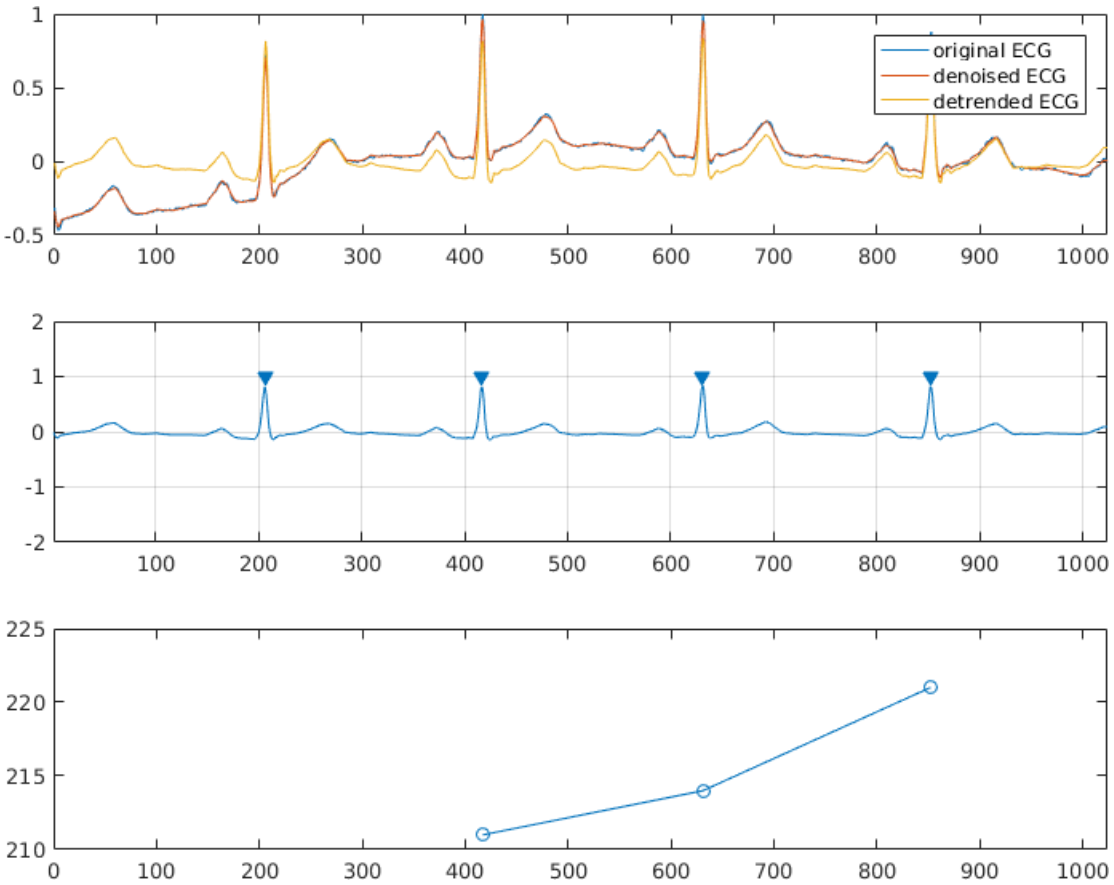


Figure 5.5: *Visualisation of the three main steps adopted for electrocardiographic (ECG) signal from the RECOLA dataset. It consists of a de-trend and de-noise phase, followed by peaks detection and calculation of RR distance for heart rate variability (HRV) extraction.*

absorption levels it is possible to calculate the heart rate with alternating vasodilatation and vasoconstriction. This signal is strictly correlated with the heart rate measured via ECG and increases in presence of pleasant stimuli Selvaraj et al. (2008). The value provided by PPG, indeed, corresponds to the so called "instantaneous heart-rate" (Electrophysiology, 1996), namely the number of times the heart would beat in one minute if the duration of successive cardiac cycles were constant. Heart rate values are provided as the number of contractions of the heart per minute (BPM) and directly relates to the RR distance, by following the basic formula from the textbooks of physiology and medicine (Braunwald et al., 1998; Hall, 2010): $RR[ms] = 60[sec] * 1000/BPM$.

Interaction experiments

Both the interaction experiments presented here do require a preliminary learning phase of the model, one per each involved subject present in the two datasets. This stage, thinking back to Section 4.2, corresponds to a supervised learning, where the facial deformations $\mathbf{m}(t)$ and autonomic states $\mathbf{v}^{(j)}(t)$, observed from a specific agent, are placed in the down-most layer ($h = 1$) of a 3-layer deep GP model (cfr. Eqs. 4.24,4.25), treated as different 'modalities'. Such modalities share the same latent space while

keeping private some of its dimensions, resulting at the $h = 2$ level in the $\mathbf{a}_M(t)$ and $\mathbf{a}_V^{(j)}(t)$ state-spaces, respectively. In this setting, the valence/arousal annotations, obtained as the result of the Evaluator Weighted Estimator (Grimm and Kroschel, 2005) of the multiple annotators, perform the role of control variables $\mathbf{c}(t)$ placed as inputs of the top layer ($h = 3$). The dimensionality of such architectural setting depends upon the considered dataset (see Tab. 5.3 for full specification). In case of AMHUSE dataset it is the following: $\mathbf{r}^{(j)}$, $p = 14$; $\mathbf{v}^{(j)}$, $k = d = 7$; $\mathbf{m}$, $N_{\mathbf{m}} = 11$; $\mathbf{m}$ and $\mathbf{v}^{(j)}$ are preprocessed through a standard PPCA stage (Tipping and Bishop, 1999) so that the input deep GP layer is partitioned in three subspaces, recurring to a Manifold Relevance Determination (Damianou et al., 2012), each having dimension $D = 4$; action state-spaces $\mathbf{a}_M(t)$ and $\mathbf{a}_V^{(j)}(t)$ have dimension $Q^{(2)} = 2 \times 3$, mirroring the $Q^{(3)} = 2$ dimensional core affect state-space. The latter is chosen akin to Russell's core affect Russell (1980). To make the point and have a pictorial idea of all the RVs and latent space involved in this work, in Figure 5.6 it is shown a visualisation of the principal components and their relation in a hierarchical manner.

Table 5.3: *Dimensionality of the features considered in each state-space of the proposed model for both the considered datasets: RECOLA with $j = \{1, 2\}$ and AMHUSE with $j = \{1, 2, 3\}$. Indices j of physiological values correspond to $\{EDA, HRV, SKT\}$. First three rows represent the input/output values for each of the deep GP layers, indexed with h.*

$\mathbf{c}$	2		2			h = 3
$\mathbf{e}$	6		6			h = 2
$\mathbf{a}_m$	4		4			h = 1
$\mathbf{a}_V^{(j)}$	4	4	4	4	4	
$\mathbf{m}$	11		11			
$\mathbf{v}^{(j)}$	10	8	7	7	7	
$\mathbf{w}$	113		113			
$\mathbf{r}^{(j)}$	20	15	14	14	14	
$\mathbf{y}$	68		68			
$\mathbf{u}^{(j)}$	1000	8000	160	160	160	
Dataset	**RECOLA**		**AMHUSE**			

Prior to the following experiments, it is worth remarking that each adopted core affect state space is learnt considering the physiological signals and visible cues of one subject at a time.

Experiment I

The aim of the first experiment is to assess the generative capability of the system. To such end, in Figure 5.7 we report the comparison between original multimodal data and their generation via our system on one subject's session randomly chosen. The experiment starts from a known sequence of valence/arousal in the top layer, which is generatively propagated to the bottom ones. For the sake of comparison, all the generated chunks of values $\mathbf{v}^{(j)}(t)$ and $\mathbf{m}(t)$ are brought to their original 1-dimensional representation and concatenated to each other to recreate the original signal. In the

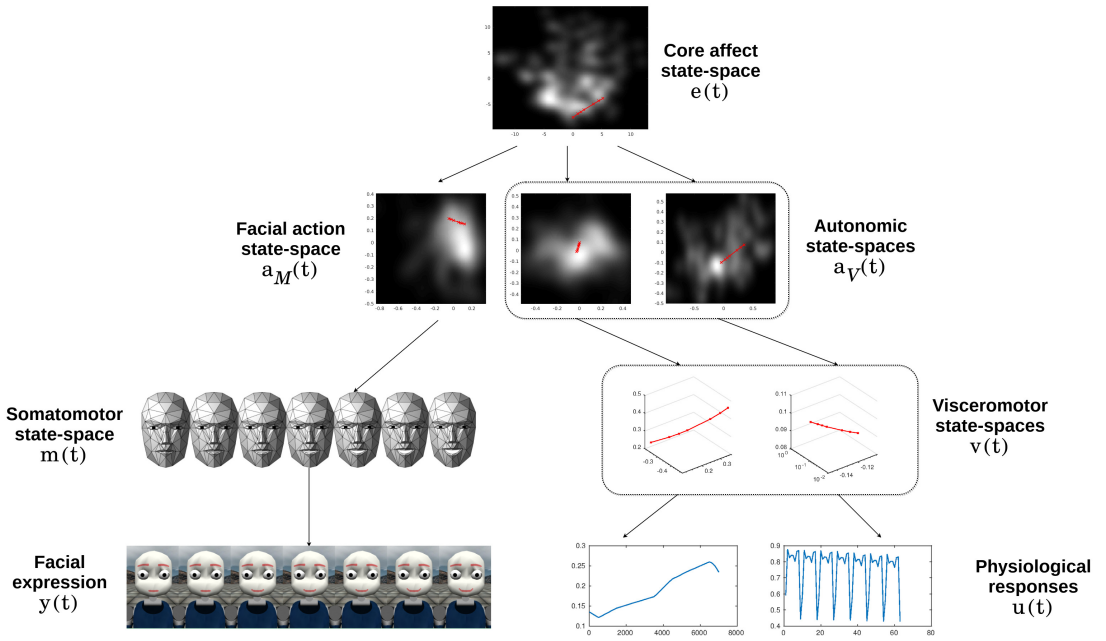
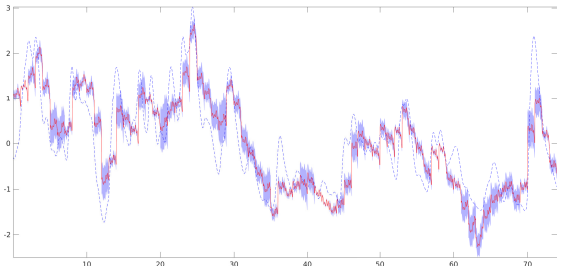


Figure 5.6: *Visualisation of the model among each of its main components (cfr. Fig. 3.12). Note that for clarity the $r^{(j)}(t)$ and $w(t)$ are omitted and $v(t)$ state-space shows only 3 of the k dimensions.*

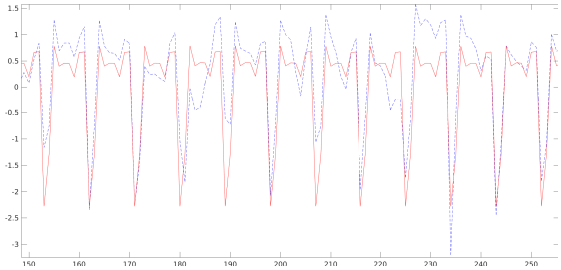
former case, the feature vector $\mathbf{r}^{(j)}(t)$ is obtained through Eq. 4.35 and, thanks to the orthogonal property of the considered wavelet transformation, we are able to generate back the physiological responses $u^{(j)}(t)$ via inverse discrete wavelet transform (IDWT). In the latter case, generated facial deformation controls $\mathbf{m}(t)$ are simply reshaped according to the AU cardinality, $N_m = 11$, and plotted against time as the corresponding ground truth.

To give a quantitative evaluation of the system generative capability, we compute the mean square error (μ_{MSE}) and mean Pearson's correlation coefficient (μ_r), at the 0.05 significance level, between the ground truth and the predicted sequences for each of the considered signals, obtained as the result of 10 sampling processes applied to all the learned models. In Figure 5.7 we show the results obtained with a model learned on the 'P19' subject's data, brought from RECOLA dataset. In particular, for the EDA signal (Fig. 5.7a) we obtained a $\mu^{(1)}_{MSE_{19}} = 0.2462$ and a $\mu^{(1)}_{r_{19}} = 0.8769$ ($p < 0.001$). A similar result is achieved also for the HRV (Fig. 5.7b), where $\mu^{(2)}_{MSE_{19}} = 0.2696$ and $\mu^{(2)}_{r_{19}} = 0.8648$ ($p < 0.001$). In both cases the correlation is statistically significant. Finally, for the action units activation values (Fig. 5.7c), we obtained $\mu^{\mathbf{m}}_{MSE_{19}} = 1.2438$ and $\mu^{\mathbf{m}}_{r_{19}} = 0.3969$ as a result of evaluation over the 11 considered AUs, where AU_k, $k = \{2, 4, 5, 7, 9, 10, 15, 20, 23, 26, 45\}$. The correlation coefficients were respectively $r_k = \{0.46, 0.22, 0.23, 0.07, 0.36, 0.41, 0.14, 0.27, 0.46, 0.88, 0.57\}$. In all cases the p-value was under significance level (0.05), apart for AU7 and AU15 where $p = 0.58, 0.24$, respectively.

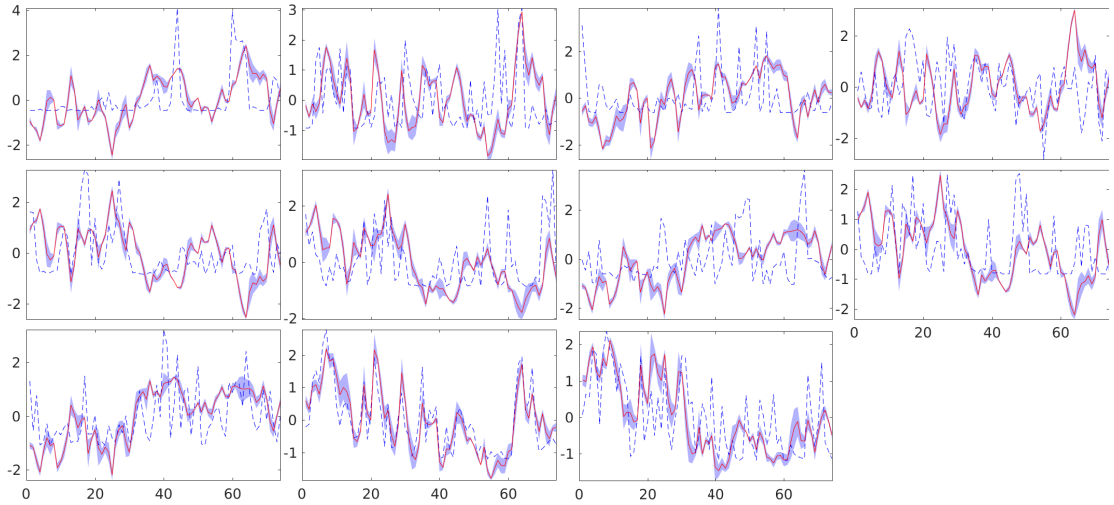
A comparison of all the obtained results, considering RECOLA and AMHUSE datasets, is reported in Table 5.4, where the best and the mean results gathered in both cases are put in evidence, and the p-value was under significance level (0.05). Fig-



(a) *Reconstructed EDA signal.*



(b) *Chunk of a reconstructed HRV signal.*



(c) *Reconstructed Action Unit activation values. From top-left to bottom-right:* AU_k , $k = \{2, 4, 5, 7, 9, 10, 15, 20, 23, 26, 45\}$.

Figure 5.7: *Results of Experiment I. Generation of a session of physiological signals* $u^{(j)}(t)$ *(a), (b) and facial actions* $\mathbf{m}(t)$ *(c) (red) compared to the ground truth (dashed blue). In shaded light blue the* 95% *prediction confidence interval.*

ure 5.8, instead, shows how the data nature impacts the obtained results. This could happen because of the presence of noisy or corrupted data in a specific session, or even a biased emotional annotation made by the annotators. This phenomenon can be mitigated by working on two fronts: the adoption of a larger dataset both intra- and inter-subject (remember that each model is learnt relying solely on the data of one subject at a time); and by the involvement of physiological experts. These could be engaged to annotate the physiological data in terms of continuous emotional values, instead of relying on the shown facial expression only.

Experiment II

The goal here is to evaluate the aptness of the multimodal simulation-based mechanism to provide the observer a core affect dynamics, which mirrors that of novel expressers. In this case, the observer can only rely on the visual information displayed by the expresser, whilst autonomic activity is at this point "innate", i.e. learnt previously from an expresser different from the new ones.

Table 5.4: *Results of Experiment I, in terms of mean square error (mse) and Pearson's correlation coefficient (cc). It shows the mean (μ), the standard deviation (σ), the minimum and the maximum values for each of the considered physiological signal and both the adopted datasets. In bold the best results achieved in each trial.*

		μ	σ	min	max	μ	σ	min	max
EDA	mse	1.36	0.73	**0.25**	2.52	1.19	0.36	**0.52**	1.62
	cc	0.32	0.36	-0.26	**0.88**	0.40	0.18	0.19	**0.74**
HRV	mse	0.66	0.20	**0.27**	0.90	0.71	0.72	**0.05**	2.35
	cc	0.67	0.10	0.55	**0.86**	0.64	0.36	-0.17	**0.97**
SKT	mse	-	-	-	-	1.14	0.54	**0.30**	1.95
	cc	-	-	-	-	0.43	0.27	0.02	**0.85**
		RECOLA				**AMHUSE**			

This second experiment has been conducted adopting a one vs. all comparison. At learning stage, a model for each subject of the considered datasets is learnt. At testing stage, one learnt model is referred to as observer, while all the others are exploited as expressers interacting with the chosen observer. The interaction process is driven on the expresser's side by random sampling of core affect trajectories (Gaussian random walk with drift) from which via top-down forward sampling, visible and physiological cues are generated. As in every face-to-face interaction in the real world, the observer can only rely upon expresser's visible cues inferred from the facial expression. Along the interaction, observer's affective dynamics unfolds as described in Algorithm 1.

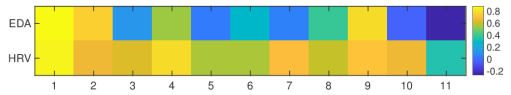
In order to assess the effectiveness of the different visible and hidden cues in determining an observer core affect state (that is predictive of that of the expresser), four (or five) different settings were adopted. In particular, we simulated the prediction process by relying:

1. only on the observer's somatomotor route (SM),
2. combining SM and the electrodermal route (VM_{EDA})
3. combining SM and heart-rate route (VM_{HRV}),
4. considering SM and all available visceromotor routes $\text{VM}_{(j)}$,

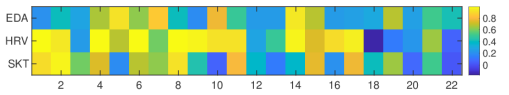
and, only in case of the AMHUSE dataset, also

5. combining SM and the skin temperature route (VM_{SKT}).

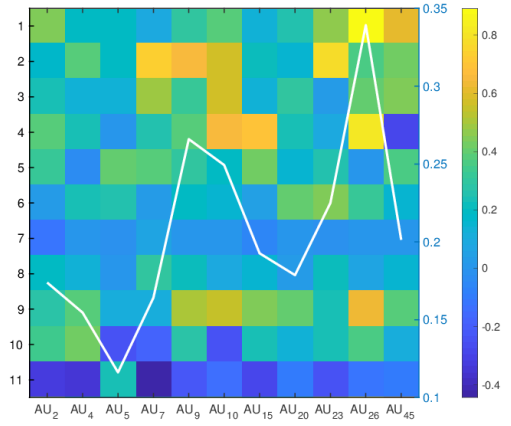
The results, shown in Figure 5.9, provide evidence of the importance of physiological internal cues in the prediction of other's internal core affect state. In particular, for the RECOLA dataset (cfr. Fig. 5.9a) it is shown that for arousal, the root mean square error (RMSE) value between expresser's and observer's trajectories improves from 0.987 of the first setting to 0.73 of the last one. A similar behaviour can be noticed also for the valence, from 1.276 for the first setting to 0.987 for the "complete" setting. Adopting the AMHUSE dataset (cfr. Fig. 5.9b) the trend is confirmed, in particular it can be noticed how the adoption of the somatomotor-only route (SM) results



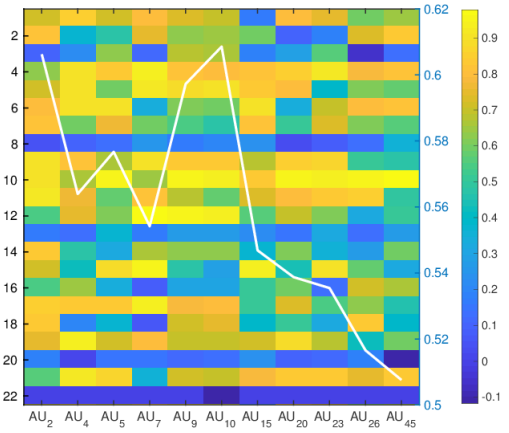
(a) *Physiological data reconstruction from RECOLA dataset.*



(b) *Physiological data reconstruction from AMHUSE dataset.*



(c) *Somatomotor data reconstruction from RECOLA dataset.*



(d) *Somatomotor data reconstruction from AMHUSE dataset.*

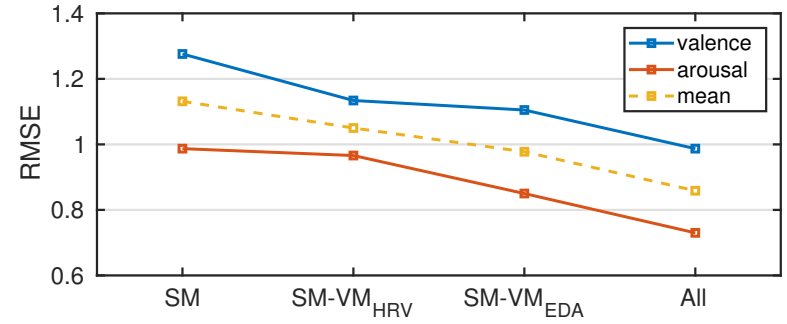
Figure 5.8: *Results of Experiment I. Distribution of the Pearson's correlation coefficients (CC) between the ground truth and the predicted value over each of the considered signals and learned models. The CC value is represented as a coloured square, going from blue (low correlation) to yellow (high correlation). On the x-axis of 5.8a and 5.8b are reported the index of the learned models for both the datasets, while on the y-axis the considered physiological data. On the x-axis of 5.8c and 5.8d are reported the considered Action Units (AU) of the somatomotor state-space, while on the left y-axis the index of the learned models for both the datasets. The white line (right y-axis) corresponds to the mean CC value for each AU over all the learned models. This visualization highlights the variety and the dependence from the data on the goodness of the results.*

in a better value for the valence, rather than for the arousal. This result confirms the importance of the facial expression in the valence emotional dimension. A result that appears mitigated in the RECOLA dataset, because of the acquisition setting and the subject's pose which disadvantages the face recording.

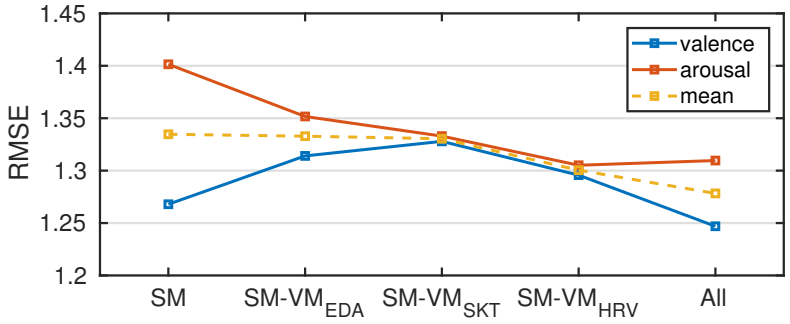
Discussion

The experiments presented above have a twofold value: assess the adequacy of the adopted implementation choices and evaluate, from an observer's side, the contribution of his internal autonomic information during a simulation-based mechanism.

The first point can be appraised relying on the Experiment I that, despite the good results in the reconstruction process highlighted in Table 5.4, also shows the fragmentation of its behaviour along the different learned models (cfr. Fig. 5.8). We should recall, indeed, that each considered model shares the same architectural choices (cfr. Tab. 5.3) and is trained over a single subject's set of samples. The presence of noisy



(a) *RECOLA dataset.*



(b) *AMHUSE dataset.*

Figure 5.9: *Results of Experiment II. Namely the root mean square error (RMSE) between the expresser's emotional value and the observer's inference, for each of the considered settings: only somatomotor route (SM), somatomotor and electrodermal routes (SM-VM$_{EDA}$), somatomotor and heart-rate routes (SM-VM$_{HRV}$) and all somato- and viscero- motor routes. Only for the AMHUSE dataset is present also the combination of somatomotor and skin temperature routes (SM-VM$_{SKT}$). The dashed line represents the mean RMSE of the combined prediction (valence/arousal).*

or corrupted data in such samples directly affects the model's expressiveness and consequently the results. Moreover, it should be noticed that the considered emotional annotations provided as constraint in the top layer are the result of a manual annotation done relying solely on the shown facial expression. This leaves some problems open. When analysing the continuous dimensions of emotion, there is the necessity of continuous annotations, but as summarized in Sariyanidi et al. (2015): "*Labelling data is a challenging and laborious task, particularly for spontaneously displayed expressions and emotions*." Moreover, combining multiple annotations is even harder (Metallinou and Narayanan, 2013). In this respect, we tried to mitigate this problem developing DANTE (Boccignone et al., 2017) that is friendly and intuitive and allow the user to seamlessly annotate valence and arousal separately, during the video reproduction of the subject's recorded behaviour. Anyway this solution does not face the issue of the annotation of non visible cues. The involvement of physiological experts, in fact, could comprise the annotation of autonomic data in terms of continuous emotional values, only looking at the recorded physiological traces and identifying behavioural patterns.

The second experiment confirms the proposed hypothesis, or rather the importance

of physiological internal cues in the prediction of other's internal core affect state. In general, in Figure 5.9 it is shown how the involvement of all the observer's somato- and viscero- motor routes results in a lower RMSE of the inferred emotional couple state (valence/arousal). Interestingly enough, looking at the detail of the single emotional components of Figure 5.9b it can be noticed that, as expected, the setting that contemplates only the somatomotor route outperforms all the other evaluations of the valence dimension, while presents a symmetric behaviour for the arousal. This result appears weaker on the RECOLA dataset (Fig. 5.9a), where the dataset acquisition setting and the pose assumed by the participants strongly affect the face visibility in the recorded videos (cfr. Fig. 4.6).

As a final remark, we have adopted some implementation choices that are instrumental in our experiments, while not affecting the generality of the model. For instance we have exploited the Candide-3D model Ahlberg (2010) to support the internal motor state-space dynamics. Though developed for the purpose of model-based image coding and facial animation, it exhibits many appealing features (such as the straightforward representation of SUs and AUs). There are of course limitations, mostly due to a lack of exact correspondence between the Action Units set used by the detector and the ones considered in the Candide model. For example, AU_{12} (Lip Corner Puller) is not accounted for by our Candide implementation, thus it is never activated for any expression. Interestingly enough, associated AUs can be activated as surrogates, for instance AU_{14} (Dimpler, that forms in the cheeks when one smiles) in place of AU_{12}. This effect can be easily noticed for the "happy" expression.

In this concern, there are viable possibilities such as replacing Candide with detailed 3D facial muscle models Eskil and Benli (2014). The latter option could be somehow mandatory, when precise and realistic visible mimicry is requested. Nevertheless, albeit some simulationist approaches claim it as a necessary condition Goldman and Sripada (2005), others retain internal, or *as if* simulation a more plausible mechanism Goldman and Sripada (2005); Adolphs (2002b); Damasio (1999).

Summary

In this chapter we effectively realised the implementation choices described in Chapter 4. This went through a first presentation of the considered datasets, highlighting the main components for each of them. It followed with a description of the essential steps adopted during the signal preprocessing phase, distinguishing between the peculiar characteristics of each signal. In the last sections two experiments were presented covering both the potentials of the model: inference and generation, along with a brief discussion of the obtained results.

CHAPTER 6

Theoretical implications

In this chapter we discuss the proposed model by elaborating on some key theoretical aspects, and outline some connections with current research.

For what concerns the theoretical model, we further expand on the view of the theoretical model as a structured *ensemble* of stochastic processes.

As to implementation model, we discuss how, coping with the difficult inference problems - arising from the complex entanglement of the processes involved -, brings about the predictive coding account, under suitable optimisation choices (Variational Bayes optimisation on the free energy of the system),

Predictive coding has a long history, originating with the insights of Von Helmholtz (1867) and reaching recent prominence in the "Bayesian brain" hypothesis (Knill and Pouget, 2004). The idea is that, in order to support adaptive responses, the brain must discover information about the likely causes of sensory signals (in our case either due to external and internal perception) without direct access to these causes, using only information in the flux of sensory signals themselves. According to predictive coding, this is accomplished via probabilistic inference on the causes of sensory signals, computed according to Bayesian principles. Estimating the probable causes of data (the posterior) given observed conditional probabilities (likelihoods) and prior beliefs about probable causes entails, a predictive or generative model of the sensory data.

The generalisation to a hierarchical context implies that posteriors at one level form the priors at one level lower, thus enabling priors to be induced from the data stream itself (empirical Bayes).

Such discussion paves the way, together with the multilevel analysis framework, for a comprehensive comparison of the proposal we have put forward in this dissertation to other works addressing the computational modelling of affect.

The theoretical model as an entanglement of stochastic processes

Our basic assumption, **Assumption 0** in Chapter 3, posits that the core affect dynamics results from a complex, open system, and is therefore subject to stochastic variability resulting from the many internal and external events that influence core affect. In particular, **Assumption 5** introduced the idea that a dynamic input-state-output model could account for such dynamics.

Subsequent assumptions, aimed at identifying the interacting systems as those state-spaces involved by the somatomotor and visceromotor routes; the theoretical model introduced in Chapter 4 structurally shaped such interactions in the form of a Probabilistic Graphical Model.

Indeed, all state-space denoting variables actually are stochastic processes, even though, for notational simplicity and presentation convenience, we have treated them as one realisation of the process. Let us now give a more precise shape to the stochastic view of the theoretical model and to related implications.

On a measurable state-space $(\Omega, \mathcal{A}, \mathbb{P})$ the following are given:

- a family of probability measures $\mathcal{M} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ depending on a parameter θ, with density P_θ;
- a pair of stochastic processes $X = \{X_t, 0 \leq t \leq T\}$ and $Y = \{Y_t, 0 \leq t \leq T\}$ taking values in $\mathbb{R}_x$ and $\mathbb{R}_y$, respectively.

Suppose X, under $\mathbb{P}_\theta$, is a Markov process with an infinitesimal generator, then we can write the state-space equations of a dynamical stochastic system in the following form of Itô SDE (to be interpreted as an Ito stochastic integral):

$$dX_t = f(X_t, C_t)dt + D^{1/2}dW_t, \tag{6.1}$$

$$dY_t = g(X_t, C_t)dt + R^{1/2}dV_t, \tag{6.2}$$

where $W = \{W_t, 0 \leq t \leq T\}$ and $V = \{V_t, 0 \leq t \leq T\}$ are respectively independent standard processes (e.g. Wiener), of the same dimension of X and Y respectively. D, R are diffusion coefficients. The latter could be in general a function of the states, i.e. $D = D(X_t), R = R(X_t)$

The variable C also is defined for wider generality as the stochastic processes $C = \{X_t, 0 \leq t \leq T\}$, though in specific cases can be deterministic, stochastic, or both. It represents the system control, which is also variously referred to in the literature as input, cause or source. This can be shaped in many ways, for example as a function of both X and Y (e.g. to introduce feedback) or an exogenous input (e.g., the labelling sequence provided along a supervised learning stage).

Also, we denote f and g the generic (vector or scalar valued) nonlinear, potentially time-varying functions, i.e. mappings of the kind $T \times L^2(\Omega, \mathcal{A}, \mathbb{P}) \mapsto L^2(\Omega, \mathcal{A}, P)$ to a (Lebegue square-integrable) Hilbert space $L^2(\Omega, \mathcal{A}, P)$ with finite second-order moments.

In fact, Eqs. 6.1 and 6.2 can be easily recognised as diffusion processes, f and g being their respective drifts (Van Kampen, 2011).

We can think of this processes as the limit of the discrete-time processes

$$X_{t+\Delta t} - X_t = f(X_t, C_t)\Delta t + D^{1/2}\sqrt{\Delta t}\epsilon_{X_t}, \tag{6.3}$$

$$Y_{t+\Delta t} - Y_t = g(X_t, C_t)\Delta t + R^{1/2}\sqrt{\Delta t}\epsilon_{Y_t}, \tag{6.4}$$

Equations 6.3 and 6.4 are known as the Euler-Maruyama approximation of Eqs. 6.1 and 6.2.

Assume that Ω is the canonical state-space $\Gamma([0,T];\mathbb{R}_{x+y})$, in which case X and Y are the canonical processes on $\Gamma([0,T];\mathbb{R}_x)$ and $\Gamma([0,T];\mathbb{R}_y)$, respectively, and P_θ is the probability law of (X,Y). In such case X is the state process, which is not directly observed; rather, the information about its evolution is obtained through the noisy observed process Y.

Then, Eqs. 6.1 and 6.2 define a generalised input-output state-space system (SSM) where the states X_t mediate the influence of the input on the output and endow the system with memory. The state and observation perturbations or fluctuations are provided by noise terms ϵ_X, ϵ_Y, which can be defined via the stochastic integrals $W_t = \int_0^t \epsilon_{X_s} ds$, $V_t = \int_0^t \epsilon_{Y_s} ds$. In the case of W, V being Wiener processes, ϵ_X, ϵ_Y represent Gaussian additive noise, and have the same dimension of X, Y, respectively. If errors are iid Gaussian random variables, then the specific scaling of the white noise with Δt gives rise to the nondifferentiable trajectories of sample paths characteristic for a diffusion process.

The classic input-output SSM can be recovered from Eqs. 6.1 and 6.2, under the independence assumption $(Y_t \perp Y_{t-1} \mid X_t)$:

$$dX_t = f(X_t, C_t)dt + D^{1/2}dW_t, \tag{6.5}$$

$$Y_t = g(X_t, C_t) + R^{1/2}\epsilon_{Y_s}, \tag{6.6}$$

Obviously, considering Eqs. 6.3 and 6.4, under the conditional independence assumption $(Y_{t+\Delta t} \perp Y_t \mid X_{t+\Delta t})$, then Y_t only depends on X_t, C_t and we recover the discrete time input-output SSM. In such case, the stochastic difference equations (3.1) and (3.2) can be easily obtained.

Indeed, such stochastic model grounds, at the implementation modelling level, the components realising the main sub-graphs of the PGM: Eqs. 4.24 and 4.25 used in the hidden and visible layers of the Deep GP adopted in the implementation model for instantiating the dynamics of the central affect state-spaces; Eqs. 4.26 and 4.31 governing the somatomotor state-space; eventually, Eqs. 4.34 and 4.35, that we have adopted for instantiating the visceromotor state-space dynamics.

Equations 6.1 and 6.2 above formalise the generative process: when the dynamics unfolds, the process Y generates a σ-algebra.

Denote $\mathcal{Y}_t = Y_{0:t} = \{Y_0, Y_1, \cdots, Y_t\}$ a filtration. We can define then an innovation process $E = \{E_t, 0 \leq t \leq T\}$ or *prediction error*

$$E_t = Y_t - \int_0^t \mathbb{E}_\theta[g(X_s, C_s) \mid Y_{0:s}, C_{0:s}]ds \tag{6.7}$$

which will be central when we want to invert the model.

The framework defined via Eqs. 6.1 and 6.2 ensures that the probability measures in $\mathcal{M}$ are mutually absolutely continuous. Then, the connection of the stochastic process

to the Bayesian setting can be made via the Radon-Nikodym Theorem (Stuart, 2010). If we let θ_0 be the reference set of parameter and write $\mathbb{P}_{\theta_0}$ as $\mathbb{P}_0$ (prior measure, with associated prior density P_0), the Radon-Nikodym derivative of $\mathbb{P}_\theta$ with respect to $\mathbb{P}_0$ provides the complete data likelihood.

$$\frac{d\mathbb{P}_\theta}{d\mathbb{P}_0} = \frac{P_\theta(Y_t \mid C_t)}{P_\theta(Y_{0:T} \mid C_{0:T})} \tag{6.8}$$

Let $\mathbb{P}_\theta^Y$ denote the restriction of $\mathbb{P}_\theta$ to the σ algebra generated by process Y, then the likelihood function for estimating the parameters θ on the basis of a given observation path $Y = \{Y_t, 0 \leq t \leq T\}$ can be expressed as (Wang and Titterington, 2004)

$$\mathcal{L}(\theta \mid Y_{0:T}) = E_0 \left[\frac{d\mathbb{P}_\theta^Y}{d\mathbb{P}_0} \mid Y_{0:T}, C_{0:T} \right]. \tag{6.9}$$

where E_0 denotes the expectation under $\mathbb{P}_0$

This provides the necessary link to the Bayesian view of the dynamic stochastic process, which is in turn instantiated in terms of a Probabilistic Graphical Model.

But before discussing this point, it is worth remarking that yet is intriguing to look at the dynamics of the core affect as the dynamics of a stochastic process, which emerges due to the entanglement, at a deeper level, of stochastic processes occurring in the somatomotor and visceromotor routes.

Indeed, the study of affect or emotion dynamics entails investigating the patterns and underlying processes that describe people's fluctuations and changes in emotion and its components across time, notwithstanding the fact that most emotion research has focused on emotion or affect as a state and on identifying its antecedents and consequences (Kuppens et al., 2010)

Oravecz et al. (2011) have developed a specific dynamic system model, namely a model for temporal fluctuations in the core affect state over time, with individual differences for the crucial parameters. The core of the model can be described in terms of two equations. They focus on a particular case of the system described by Eqs. 6.1, 6.2, where state equation (6.1) is assumed to be the Ornstein-Uhlenbeck (OU) process (Uhlenbeck and Ornstein, 1930a).

$$dX_t = \beta(C - X_t)dt + D^{1/2}dW_t, \tag{6.10}$$

$$Y_t = X_t + R^{1/2}\epsilon_{Y_s}, \tag{6.11}$$

where $\beta > 0$ and control C in the simplest case is assumed to be a constant, i.e. $C = const$, or time-varying in the most general case. The instantaneous change in X_t, that is, dX_t, depends on how far the current state X_t is from the point C. This control parameter is called a steady state or attractor: as a straightforward example in the one dimensional case, if X_t is below C (i.e., $C - X_t < 0$), the first derivative is positive, and consequently X_t will increase; the opposite holds when X_t is above. The parameter β controls the magnitude of the "attraction" effect: if β is large ($\beta \gg 1$), the difference between the actual state and C tends to be magnified; therefore a faster change will occur in the direction of C; with small β), the change becomes substantially slower. Based on this property, the parameter is often called the dampening force or

centralising tendency. The stochastic innovation term dW_t incorporates the multiple smaller and larger impacts that the core affect system undergoes at a given moment.

Interestingly, C acts as a set point that reflects the baseline functioning of the system, an affective "home base", which reflect the affective comfort zone of an individual, signalling that everything is normal. The attractor keeps the system in balance by pulling core affect back to its home base, creating an emergent coherence around it. It is surmised, the attractor strength reflects the regulatory processes that are installed to keep a person's core affect in check.

To sum up, these three key processes - affective home base, variability, and attractor strength -are largely responsible for producing the myriad ways people can display changes and fluctuations in their core affect throughout daily life (Kuppens et al., 2010).

The model has been empirically evaluated in two extensive experience-sampling studies on people's core affective experiences. The findings have shown that it is capable of adequately capturing the observed dynamics in core affect across both large and shorter time scales and illuminate how the key processes are related to personality and emotion dispositions. More precisely it was capable of replicating the shape of individuals' core affect trajectories, how often they are in particular feeling states across time, and the dynamical forces that impinge on their feelings when in different feeling states. In conclusion, the model accounts for individual differences in temporal patterns and trajectories observed in people's affective experiences (Kuppens et al., 2010).

The model by Kuppens et al. (2010) constitutes a theoretical model of the core affect dynamics, under the assumption that such state-space results from a complex, open system. In this perspective, the model proposed in this dissertation is an attempt to make a step further in such direction, by grounding core affect dynamics as the result of an open system interaction with visuomotor and visceromotor state-spaces. In turn, each state-space within these routes undergoes a stochastic diffusion. Control variables and outputs account for the information inflow/outflow between subsystems.

Clearly, the higher complexity of our setting calls for an implementation suitable to support adequate inferential steps, which will be discussed in the following.

A hierarchical predictive view of the implementation model

Inverting the model, from a Bayesian standpoint, boils down to an inferential step. When we want to make inference that can be either on hidden states or on parameters θ, or both and relies upon computing the posterior $P(X, \theta \mid Y, C)$, from which $P(\theta \mid Y, C)$ and $P(X \mid Y, C)$ can be obtained via marginalisation (hereafter, we drop random variable indexes from the density P for notational simplicity).

In fact, in the most general case, the joint posterior distribution, can be written

$$\begin{aligned} P(X_t, \theta_t \mid Y_{0:t}, C_{0:t}) &= P(X_t \mid \theta_t, Y_{0:t}, C_{0:t}) P(\theta_t \mid Y_{0:t}, C_{0:t}) \\ &= P(\theta_t \mid X_t, Y_{0:t}, C_{0:t}) P(X_t \mid Y_{0:t}, C_{0:t}) \end{aligned} \tag{6.12}$$

It is easy to show, by using Bayes' rule under Markovian assumption, that the Bayesian recursive estimation of $P(\theta_t \mid Y_{0:t}, C_{0:t})$ boils down to

$$P(\theta_t \mid Y_{0:t}, C_{0:t}) = \frac{P(Y_{0:t} \mid \theta_t, C_{0:t}) P(\theta_t \mid Y_{0:t-1}, C_{0:t-1})}{P(Y_t \mid Y_{0:t-1}, C_{0:t-1})}, \tag{6.13}$$

with parameters prior given by the Chapman-Kolmogorov equation

$$P(\theta_t \mid Y_{0:t-1}, C_{0:t-1}) = \int P(\theta_t \mid \theta_{t-1}) P(\theta_{t-1} \mid Y_{0:t-1}, C_{0:t-1}) d\theta_{t-1}. \tag{6.14}$$

Substitution in Eq. 6.12 gives:

$$\begin{aligned} P(X_t, \theta_t \mid Y_{0:t}, C_{0:t}) &= P(X_t \mid \theta_t, Y_{0:t}, C_{0:t}) P(\theta_t \mid Y_{0:t}, C_{0:t}) \\ &= P(X_t \mid \theta_t, Y_{0:t}, C_{0:t}) \frac{P(Y_{0:t} \mid \theta_t, C_{0:t}) \int P(\theta_t \mid \theta_{t-1}) P(\theta_{t-1} \mid Y_{0:t-1}, C_{0:t-1}) d\theta_{t-1}}{P(Y_t \mid Y_{0:t-1}, C_{0:t-1})} \end{aligned} \tag{6.15}$$

It is clear that this kind of problem entails a dual or triple estimation, depending on whether the hyperparameters related to the distributions are known. In a full Bayesian analysis is computationally complex because complicated multiple integrations are involved. Involved integrals have no general analytic solution, particularly when the generative model is nonlinear, thus some algorithmic approximation should be devised.

There are actually two roads that can be pursued: free-form, Monte Carlo based approximation and variational, deterministic approximations. As to the first option, Markov chain Monte Carlo for numerical integration helps to side-step this problem, but it is clearly time-consuming, samples of parameter values are required to be stored and there are risks to be run as to whether or not convergence has occurred. In engineering and machine learning, free-form densities are usually approximated by the sample density of a large number of "particles" that populate state-space. In statistics the problem of Bayesian inference for both the state and parameters, within partially observed, non-linear diffusion processes has been tackled using Markov Chain Monte Carlo (MCMC) approaches based on data augmentation, Monte Carlo exact simulation methods, or Langevin / hybrid Monte Carlo methods (Bishop, 2006a; MacKay, 2004). Within the signal processing community solutions to the so called Zakai equation based on particle filters, a variety of extensions to the Kalman filter/smoother and mean field analysis of the SDE together with moment closure methods have also been proposed (Archambeau et al., 2008).

More recently, a deterministic approximate approach to the intractable Bayesian inference problem, the variational Bayesian approximation (VB), has been introduced (see Beal and Ghahramani, 2003; MacKay, 2004 for an insightful discussion). Rather than use sampling, the main idea behind variational inference is to use optimisation. In a nutshell:

1. posit a family of approximate densities, namely a set of densities over the latent variables (either states and/or parameters); then,
2. try to find the member of that family that minimises the distance to the exact posterior.

Variational Bayes draws together variational ideas from the analysis of intractable latent variable models and from Bayesian inference. This framework facilitates analytical calculation of posterior distributions over the hidden variables, parameters and structures. They are computed via an iterative algorithm, VBEM (Beal and Ghahramani, 2003).

Variational approximations rely on bound approximation, by adopting approximating posteriors. Further, if a fixed-form approximation is adopted, this choice allows one to represent the density in terms of a small number of quantities, namely its sufficient statistics.

As to the neural, realisation level plausibility (Marr, 1982) of each approach, this is currently matter of debate.

Sanborn and Chater (2016) argue that sampling provides a natural and scalable implementation of Bayesian models. In this view "Bayesian brains" need not represent or calculate probabilities at all and are, indeed, poorly adapted to do so. Instead, the brain is a Bayesian sampler. Only with infinite samples does a Bayesian sampler conform to the laws of probability; with finite samples it systematically generates classic probabilistic reasoning errors, including the unpacking effect, base-rate neglect, and the conjunction fallacy. A key insight is that, although explicitly representing and working with a probability distribution is hard, drawing samples from that distribution is relatively easy. Sampling does not require knowledge of the whole distribution. It can work merely with a local sense of relative posterior probabilities.

In a different vein, Friston (2008) suggests that free-form approximations and their related sampling schemes are not really viable in a neuronal context. The dimensionality of the representational problems entailed by neuronal computations probably precludes particle-based (i.e., free-form) representations: face analysis, a paradigmatic example in perceptual inference. Faces can be represented in a perceptual space of about thirty dimensions (i.e., faces have about thirty discriminable attributes). To populate a thirty-dimensional space we would need at least 2^{30} particles, where each particle could correspond to the activity of thirty neurons (note that the conditional mean can be encoded with a single particle). The brain has about 2^{11} neurons at its disposal (Friston, 2008), hence a fixed-form assumption should be mandatory for the brain.

Indeed, in the implementation model we have opted for deterministic approximations of the different components that at the most general level can be described by the variational approximation, which is based on the free energy theorem (cfr. Appendix B for a detailed derivation).

Consider the joint distribution of all processes we are dealing with $P(X_{0:t}, \theta_{0:t}, Y_{0:t}, C_{0:t})$.

In what follows, to simplify notation, we denote

- $\mathbf{Z} = \{X_{0:t}, \theta_{0:t}\}$: the set of hidden states and parameters;
- $\mathbf{O} = \{I_{0:t}, C_{0:t}\}$: the set of observable measurements and controls;

Also, the upper case Roman characters such as, e.g., $\mathbf{O}$ will be used to denote random variables (either scalar or vector belonging to $\mathbf{O}$) while the lower case Roman characters) will denote a single observation as is common in statistics literature.

The idea in variational Bayes (VB) approximation relies on the introduction of a (simple) distribution, $Q_{\mathbf{Z}}(\mathbf{z})$, which is arbitrary (also with respect to the Lebegue measure), to approximate the complicated conditional distribution $P_{\mathbf{Z}|\mathbf{O}}$. In general, the well-known Kullback-Leibler divergence (see Cover and Thomas, 1991) gives us a suitable measure of the discrepancy between $Q_{\mathbf{Z}}$ and $P_{\mathbf{Z}|\mathbf{O}}$. Note that the Kullback-Leibler divergence is not symmetric, and thus a pseudo metric.

Thus, the goal is to find an approximation $Q_{\mathbf{Z}} \in \mathcal{Q}$ which minimizes KL $\left(Q_{\mathbf{Z}} \| P_{\mathbf{Z}|\mathbf{O}}\right)$. This is more conveniently achieved by introducing the functional (see Def. B.1, Ap-

pendix B)

$$\begin{aligned}\mathcal{F}(Q_{\mathbf{Z}}) &:= \int_{\mathbf{Z}} Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z})}{Q_{\mathbf{Z}}(\mathbf{z})}\, d\mathbf{z} \\ &= \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})}{Q_{\mathbf{Z}}(\mathbf{Z})}\right],\end{aligned}$$

which is called the *free energy* (MacKay, 2004) in analogy with statistical mechanics.

Then it can be shown that the following fundamental relationship holds

$$\log P_{\mathbf{O}}(\mathbf{o}) = \mathcal{F}(Q_{\mathbf{Z}}) + \mathrm{KL}\left(Q_{\mathbf{Z}} \| P_{\mathbf{Z}|\mathbf{O}}\right). \tag{6.16}$$

It is useful to note that, due to the fact that $\mathrm{KL}\left(Q_{\mathbf{Z}} \| P_{\mathbf{Z}|\mathbf{O}}\right) \geq 0$ almost everywhere (a.e.), $\log P_{\mathbf{O}}(\mathbf{o}) \geq \mathcal{F}(Q_{\mathbf{Z}})$, and thus $e^{\mathcal{F}(Q_{\mathbf{Z}})}$ provides a lower bound for the marginal density of $\mathbf{O}$.

Note that the left hand side does not vary with $\mathbf{z}$; moreover, we are considering an experiment where we observe $\mathbf{o}$, so the quantity is fixed. It is generally referred to as the log marginal likelihood or the log evidence. Intuitively, since $\log P_{\mathbf{O}}(\mathbf{o})$ i does not vary with $Q_{\mathbf{Z}}$, it is clear that the functionals $\mathcal{F}$ and KL are inversely related. Therefore, a minimization of KL amounts to a maximization of $\mathcal{F}$, i.e., it is possible to determine the optimal approximating distribution $Q^*_{\mathbf{Z}}$ as

$$Q^*_{\mathbf{Z}} := \arg\min_{Q_{\mathbf{Z}} \in Q} \mathrm{KL}\left(Q_{\mathbf{Z}} \| P_{\mathbf{Z}|\mathbf{O}}\right) = \arg\max_{Q_{\mathbf{Z}} \in Q} \mathcal{F}(Q_{\mathbf{Z}}), \tag{6.17}$$

where Q denotes any set of valid probability densities (recall that the choice of $Q_{\mathbf{Z}}(\mathbf{z})$ is entirely arbitrary so long as it is a probability density with respect to the Lebesgue measure). We will refer to $Q^*_{\mathbf{Z}}$ as the Q-VB approximation.

Finding $Q^*_{\mathbf{Z}} \in \mathcal{Q}$ when $\mathcal{Q}$ is any probability density is in general a difficult task. To make our analysis more tractable we can impose an independence structure on $\mathbf{Z}$; that is, we will only consider $Q_{\mathbf{Z}}$'s which come from the set

$$\mathcal{Q} := \left\{ Q_{\mathbf{Z}}(\mathbf{z}) : Q_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{m} Q_{Z_i}(z_i) \right\}, \tag{6.18}$$

where $Q_{Z_i}(z_i)$ is the probability density of Z_i, the ith element of $\mathbf{Z}$. This is also known as the mean-field approximation. Sometimes even this task proves difficult and more restrictions are imposed to make Q even smaller. Such techniques are referred to as restricted variational Bayes (R-VB) techniques. For example, we could add the additional requirement that each $Q_{Z_i}(z_i)$ is in the exponential family. For the rest of this discussion, we will take Q^* as defined in Equation 6.17 for our Q defined in (6.18).

Under the mean-field approximation, the free energy theorem (cfr. Appendix B, for a detailed proof) grants that the solution that holds for the distribution $Q^*_{Z_j}$ - the jth factor of $Q^*_{\mathbf{Z}}$ is:

$$Q^*_{Z_j}(z_j) \propto \exp\left\{\mathcal{U}_j(Z_j)\right\}, \tag{6.19}$$

with $\mathcal{U}_j(Z_j) = \mathbb{E}_{\prod_{\substack{i=1 \\ i \neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]$ being the expected internal energy of r.v. Z_j, and where in general $\mathbf{Z} = \{Z_1, \ldots, Z_{j-1}, z_j, Z_{j+1}, \ldots, Z_m\}$.

Taking stock of this result, the *VBEM algorithm* maximises the free energy $\mathcal{F}(Q_{\mathbf{Z}})$ with respect to each factor $Q_{Z_j}(z_j), \quad j = 1, \cdots, m$ by iteratively applying Equation 6.19, while holding other factors $Q_{Z_i}(z_i), \quad i \neq j$ fixed (Beal and Ghahramani, 2003).

For the case we are discussing, $\mathbf{Z} = \{X_{0:t}, \theta_{0:t}\}$, thus $Q_{\mathbf{Z}}(\mathbf{Z}) = Q_X(X_{0:t})Q_\theta(\theta_{0:t})$. Then, the VBEM algorithms boils down to iterate:

VB-E (expectation) step hold $Q_\theta^{(\ell)}(\theta_{0:t})$ fixed, and infer $Q_X^{(\ell+1)}(X_{0:t})$ via Eq. 6.19;

VB-M (maximisation) step hold $Q_X^{(\ell+1)}(X_{0:t})$ fixed and infer $Q_\theta^{(\ell+1)}(\theta_{0:t})$ via Eq. 6.19.

Readers familiar with the Expectation-Maximisation (EM) algorithm by Dempster et al. (1977) may note the similarity between this iterative algorithm and EM. This relationship is described in detail in Beal and Ghahramani (2003).

One straightforward example of this procedure is Equation 4.37, related to the inference in the visceromotor state-space, which is obtained by setting, in that case, $Q_{\mathbf{Z}} = Q(\mathbf{v}_{0:T}^{(j)})Q(\Theta_V^{(j)})$.

A key observation is that the optimisation of the free energy in dynamical systems involves the optimisation of the prediction error or innovation (Eq. 6.7) entailed by the optimisation of the expected internal energy $\mathcal{U}_j(Z_j)$.

One interesting example has been provided by Daunizeau et al. (2009), for the case of the classic input-output SSM. This as previously mentioned, can be recovered from Eqs. 6.1 and 6.2, under the independence assumption $(Y_t \perp Y_{t-1} \mid X_t)$, rewritten more compactly, with some abuse of notation:

$$\dot{X}_t = f(X_t, C_t) + \epsilon_{X_t}, \tag{6.20}$$
$$Y_t = g(X_t, C_t) + \epsilon_{Y_t}, \tag{6.21}$$

where the dot notation stands for the time derivative d/dt.

In this case, under Gaussian assumptions on the state and observation noises and under the Euler-Maruyama discretisation scheme (Eqs. 6.3, 6.4), the discrete-time variant of the state-space model, by setting unitary time step $\Delta t = 1$ yields the Gaussian likelihood and transition densities

$$X_{t+1} \sim P_X(X_{t+1} \mid X_t, C_t, \theta_X) = \mathcal{N}(f(X_t, C_t), D^{-1}), \tag{6.22}$$
$$Y_t \sim P_Y(Y_t \mid X_t, C_t, \theta_Y) = \mathcal{N}(g(X_t, C_t), R^{-1}) \tag{6.23}$$

When the free energy $\mathcal{F}$ is derived (see Kilner et al., 2007; Friston, 2008), it can be shown that at its heart $\mathcal{F}$ optimisation relies on computing prediction errors on hidden states dynamics $\dot{X}_t$, observations Y_t and parameters θ_X, θ_Y.

To sum up, and coming back to the general case of Equation 6.15, inference on the joint posterior $P(X_t, \theta_t \mid Y_{0:t}, C_{0:t})$, can be decomposed as the optimisation of the first term with respect to X and the second with respect to Θ, which in turn involves optimisation with respect to prediction errors either for hidden state inference and parameter learning.

This is consistent with the idea that (Candy, 2016) in the general case of joint state/-parameter inference, the joint sequential optimisation of Eq. 6.15 can be seen as made of two distinct yet coupled processes: a parameter estimator and a state estimator. The

parameter estimator provides estimates that are corrected by the corresponding prediction error during each recursion. These estimates are then provided to the state estimator in order to update the model parameters used in the estimator. After both state and parameter estimates are calculated, a new measurement is processed and the procedure continues[1]. The joint optimisation process can thus be represented as the dual process outlined in Figure 6.1.

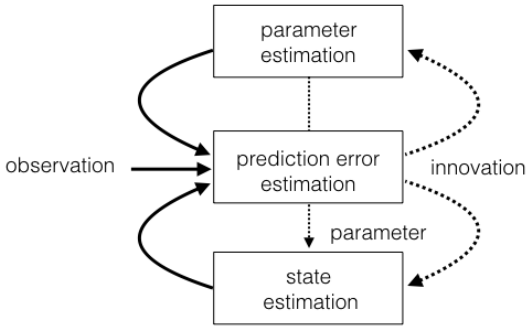


Figure 6.1: *Simplified structure illustrating the coupling between parameter and state estimation processes through the innovation and measurement sequences*

Further, it has been shown that it is possible to generalise the input-output SSM (Eqs., 6.20, 6.21) into a hierarchical form spanning on $l = 1, \cdots, L$ levels:

$$\begin{aligned}
\dot{C}_t^{(L)} &= g(C_t^{(L+1)}) + \epsilon_{C_t}^{(L+1)}, \\
&\vdots \\
\dot{X}_t^{(l)} &= f(X_t^{(l)}, C_t^{(l)}) + \epsilon_{X_t}^{(l)}, \\
\dot{C}_t^{(l-1)} &= g(X_t^{(l)}, C_t^{(l)}) + \epsilon_{C_t}^{(l)}, \\
&\vdots \\
\dot{X}_t^{(1)} &= f(X_t^{(1)}, C_t^{(1)}) + \epsilon_{X_t}^{(1)}, \\
\dot{Y}_t &= g(X_t, C_t) + \epsilon_{Y_t}
\end{aligned} \tag{6.24}$$

Note that in the hierarchical form the controls $C_t^{(L)}, C_t^{(L-1)}, \cdots C_t^{(1)}$ are used to link levels: $\dot{C}_t^{(l)}$ constrains as a top-down signal either $\dot{C}_t^{(l-1)}$ $\dot{X}_t^{(l-1)}$; at the same time $\dot{C}_t^{(l-1)}$ accounts for the emission of $\dot{X}_t^{(l-1)}$. The hidden states $X_t^{(L)}, X_t^{(L-1)}, \cdots X_t^{(1)}$ provide the necessary dynamics over time. This is in particular true when we assume time conditional independencies between controls, i.e., $(C_t^{(l)} \perp C_{t-1}^{(l)})$, and between observations, $(Y_t \perp Y_{t-1} \mid X_t)$

The conditional independence of the fluctuations at different hierarchical levels means that the hierarchy has a Markov property over levels. Thus in probabilistic terms

[1] this dual process system can be considered to be a form of identifier, since system identification is typically concerned with the estimation of a model and its associated parameters from noisy measurement data (Candy, 2016)

we can write

$$P_\theta(\dot{X}_t, \dot{C}_t) = P_{\theta^{(L)}}(\dot{C}_t^{(L)}) \prod_{l=1}^{L-1} P_{\theta^{(l)}}(\dot{X}_t^{(l)}, \mid \dot{C}_t^{(l)}) P_{\theta^{(l)}}(\dot{C}_t^{(l)} \mid \dot{X}_t^{(l+1)}, \dot{C}_t^{(l+1)}) \qquad (6.25)$$

where the prior on controls $P_{\theta^L}(\dot{C}^L)$ is restricted to the uppermost level L.

This provides substance to our opening surmise (cfr. Chapter 3) that the layers of the dynamic input-state-output models organised as in Figure 3.12, can be seen as form of hierarchical predictive coding (Wolpert et al., 2003).

Indeed, it has been suggested (Friston and Stephan, 2007; Friston, 2008; Daunizeau et al., 2009) that predictive coding may capture a general principle of cortical functional organization. It fluently explains a broad range of evidence (though a key prediction, that of distinct "state" and "error" neurons in different cortical laminae, remains to be established) and has attractive computational properties. For instance, the generalisations of forward/inverse models in motor control to imitation and social interactions that have been put forward (Wolpert et al., 2003) are formally distinct from and more complicated than predictive coding. In predictive coding there is no separate inverse model or controller. A forward model is simply inverted by suppressing the prediction error generated by the forward model.

The precision of prediction error signals plays a key role in these models on the grounds that hierarchical models of perception require optimisation of the relative precision of top-down predictions and bottom-up evidence (Friston and Stephan, 2007; Friston, 2008; Daunizeau et al., 2009).

Taking stock of the above discussion, the model we have presented in this Thesis - namely the theoretical level together with the implementation level -, can be viewed in the perspective of a hierarchical predictive scheme, as outlined in Figure 6.2

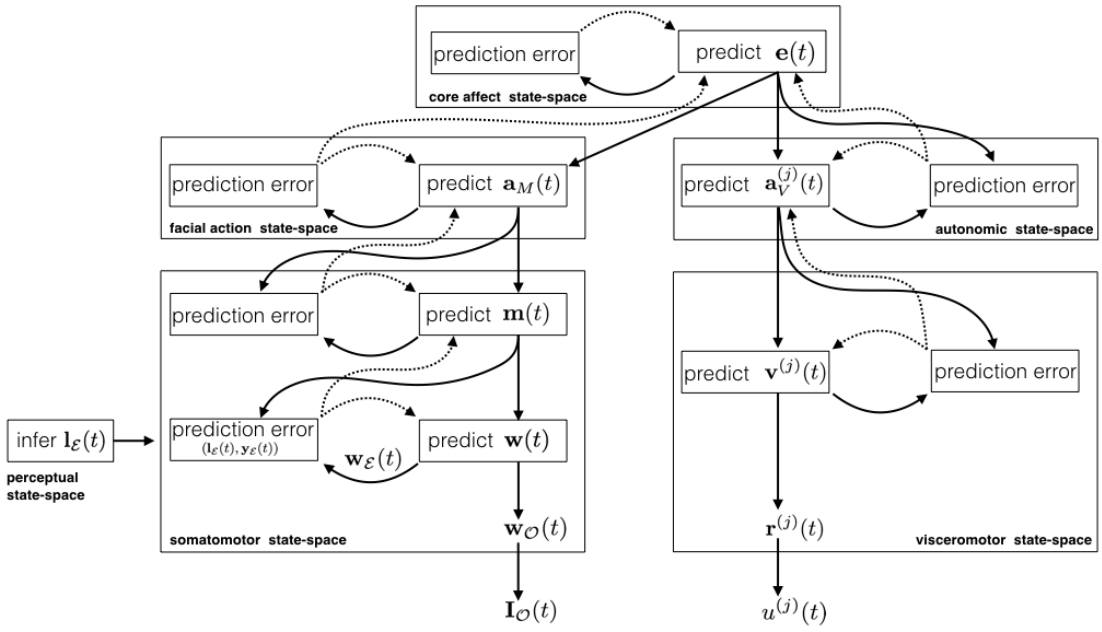


Figure 6.2: *A hierarchical predictive view of the model*

Within this scheme the most likely cause (core affect state) of an observed affective facial action can be inferred by minimising the prediction error at all levels of the hierarchy that are engaged during action observation. This jointly entails the processes involved in both the somatomotor and visceromotor routes.

Such perspective has been mostly applied to exteroceptive perception (e.g., vision, audition), and action. However it has been argued by Seth (2013) that subjective feeling states arise from actively-inferred generative (predictive) models of the causes of interoceptive afferents. Such predictive, inferential account of interoception (the sense of the internal physiological condition of the body) has been named "interoceptive inference".

Similar to what we have been surmising in setting up the model, Seth (2013) has hypothesised that interoceptive predictive coding relies upon the engagement of an extended autonomic neural substrate with emphasis on the anterior insular cortex as a comparator. Cogently, the AIC is ideally placed both to detect and to cause changes in physiological condition, and to integrate interoceptive and exteroceptive signals. Functionally, it instantiates interoceptive representations accessible to conscious awareness and is associated with processes that involve visceral representation, interoception, and emotional awareness relevant to selfhood (Seth, 2013).

According to (Seth, 2013), starting with a high-level, cognitive desired or inferred physiological state (which is itself subject to update based on higher-level motivational and goal-directed factors), generative models are engaged which predict interoceptive and exteroceptive signals via corollary discharge. Applying active inference, prediction errors are transcribed into actions via engagement of classical reflex arcs (motor control) and autonomic reflexes (autonomic control). The resulting prediction error signals are used to update the (functionally coupled) generative models and the inferred/desired state of the organism. At high hierarchical levels these generative models merge into a single multimodal model. Interoceptive predictions are proposed to be generated, compared, and updated within a salience network anchored on the anterior insular and anterior cingulate cortices (the emotional, limbic motor areas that engage brainstem regions as targets for visceromotor control and relays of afferent interoceptive signals.These areas together engage subcortical regions, such as the periaqueductal gray matter and parabrachial nucleus, as targets for visceromotor control and relays for viscerosensory afferents, as well as many other areas related to self and emotion, including the amygdala, nucleus accumbens, and orbitofrontal cortex.

Sympathetic and parasympathetic outflow from the anterior insula are in the form of interoceptive predictions that enslave autonomic reflexes (e.g., heart/respiratory rate, smooth muscle behaviour), just as proprioceptive predictions enslave classical motor reflexes in predictive coding formulations of motor control. Eventually, subjective feeling states (emotional experiences) arise from such active interoceptive inference.

Interestingly, the precision of prediction error of Seth's proposal plays a key role in these models on the grounds that hierarchical models of perception require optimisation of the relative precision of top-down predictions and bottom-up evidence. This process corresponds to modulating the gain of error units at each level, implemented by neuromodulatory systems. While for exteroception this may involve cholingeric neurotransmission via attention; for interoception, proprioception, and value-learning, prediction error precision is suggested to be encoded by dopamine.

Where are we now? A retrospective survey of the state-of-the-art

In a nutshell, to answer the Where are we now? question, we are standing in between models that aim at explaining in full details, i.e., at the biological neural level, high level cognitive functions, and theoretical models that bear no relations with the brain, trying to bridge the gap. From an epistemological perspective perspective we are endorsing a sort of "cautionary" realism.

As Sprevak (2016) put it, there are two options for interpreting Bayesian models One option is instrumentalism, where Bayesian machinery should be understood, not as referring to hidden neural entities and processes, but as a formal device (Sprevak, 2016) to describe human behavioural patterns concisely and to make predictions without evidence that the mechanisms that generate that behaviour are Bayesian. The alternative option is realism: Bayesian models are interpreted as picking out real entities and processes in the human brain (the "Bayesian brain hypothesis"Knill and Pouget (2004)). It is a controversial issue but is matter of current debate. Indeed, in the context of influential overarching current theories of cognitive and brain function, a multi-level approach, given its direct application to cognitive phenomena, can provide a potential missing link in a "trinity of models" of the brain and behaviour from the lowest levels of organisation (small-scale networks) all the way to its highest levels (cognition and behaviour): Bayesian models, Friston's theory of cortical responses based on the free-energy principle, and attractor-basin dynamics Sanborn and Chater (2017).

Nonetheless, the theoretical insights previously discussed together with the multi-level analysis we have set up, has paved the way for comparing our work with other computational proposals, that we have touched in Chapter 2.

Machine learning-based approaches. As said before, for these approaches affect detection basically is a pattern recognition problem (D'Mello and Kory, 2015). They cover a wide range of different algorithms, include the processing of single or multiple modalities combined together, they do address emotion from either a discrete and a dimensional perspective. The meta-analysis by D'Mello and Kory (2015) provides evidence for these merits. Remarkably, facial expression analysis recognition plays an important role Sariyanidi et al., 2015; D'Mello and Kory, 2015; Vinciarelli et al., 2012; Calvo and D'Mello, 2010.

However, the general assumption that understanding affective states from facial expressions can be accomplished through the computer vision and pattern recognition "pipeline" (Sariyanidi et al., 2015) is at best questionable.

It is possible to state that the "pipeline" paradigm stems from early inferential/appraisal theories of emotion, namely visual perception of emotional stimuli is followed by cognitive or appraisal/interpretative processing of the stimulus, which in turn triggers affective responses and feelings. However, it is anyway a loose link. Hardly, any of the work presented in such realm makes a claim on psychological or neurobiological plausibility of what is proposed. For instance, in facial expression analysis, Ekman's work is likely to be the most cited. This is rather instrumental: exploitation of the the FACS code is "easily" amenable to the possibility of a straightforward computational implementation; indeed, many of such works do not even attempt at going beyond the AU detection stage (Sariyanidi et al., 2015; Calvo and D'Mello, 2010); AU detection *per se* has become research goal.

This trend is likely to happen for historical reasons: many researchers now working

on facial expression analysis share a background in face detection and recognition techniques as developed in the computer vision community, and do not have indeed any real interest in the actual problem of emotions. As a consequence, machine learning-based processing of affect has inherited the current attitude that pervades computer vision, namely, considering the vision problem as a "big data" issue.

The focus, in such perspective, is performance on benchmarks. Though sophisticated from a computer science standpoint, models proposed for affective computing bear no relation with the modelling idea that motivates our work. They are basically machine learning models and as such are models devised for solving the classic machine learning problems: classification, regression, dimensionality reduction, clustering. More than often - this being true specially for the current deep learning wave - models that have been used for other purposes are simply re-trained on affective datasets.

If we go along with the definition that computational modelling of emotion refers to attempts to develop and validate computational models of human emotion mechanisms (Reisenzein et al., 2013), then, from the multilevel perspective,works in this realm do neither address the computational theory of emotion nor the implementation modelling issues.

AI-based approaches The state of affairs is different as to these approaches. In this field, there are proposals that do address theoretical explanations. In the review by Reisenzein et al. (2013), the authors claim that to realise its potential, the exchange between the two disciplines, as well as the intradisciplinary coordination, should be further improved. They make three proposals for how this could be achieved: 1) systematising and classifying the assumptions of psychological emotion theories; 2) formalising emotion theories in implementation-independent formal languages (set theory, agent logics); and 3) modelling emotions using general cognitive architectures (such as Soar and ACT-R), general agent architectures (such as the BDI architecture) or general-purpose affective agent architectures.

This is a claim which involves both the theoretical modelling and the implementation modelling explanation level. They review current work considering two alternatives:

1. formalising emotion theories in a precise but implementation-independent formal language; and

2. implementing emotion theories in a (comparatively) general-purpose cognitive or agent architecture.

As to choice 1), they provide an in-depth review of papers that exploit set theory as a language for formalising psychological emotion theories. Set-theoretical formalism is well suited to represent computational theories (as well as psychological theories interpreted as information-processing theories), because a common way of describing computational processes is as mathematical functions that map input to output information. Clearly, different from the Simulation-theory perspective assumed in the work presented in this thesis, the Theory-theory approach is the paradigm, supported by an appraisal framework. One clear example is the work by Broekens et al. (2008) who breaks up the emotion generation process assumed by cognitive appraisal theories into three linked functions: first, perceptual processes map external objects into percepts

(or mental objects); second, appraisal processes map mental objects into different appraisal dimension values; third, mediating processes map appraisals into components of emotional reactions (e.g., feelings, expressions, actions). In such perspective,

> the set-theoretical formalisation of emotion theories [...] ties in naturally with a still more rigorous set-theoretical formalism for theory reconstruction, developed as part of the so-called structuralist program in the philosophy of science [...] Essentially, this formalism consists of set theory (including whatever parts of mathematics are needed) enriched with a set of precisely defined special concepts that describe common, recurring components of empirical theories. According to structuralism, empirical theories are best viewed, from the systematic perspective, as set-theoretic structures composed of so-called theory-elements (Reisenzein et al., 2013).

One particularly interesting alternative for formalising emotion theories (intentional-level, cognitive emotion theories) - is to represent them in a logic language. Most of these so-called agent logics belong to the class of belief-desire-intention (BDI) logics, that describe autonomous agents on the intentional level in terms of beliefs, desires (goals), intentions (roughly, the goals that the agent has committed to) and possibly other related attitudes, such as capabilities.

Clearly, developing emotion theories in BDI logics is a fairly different approach from the one pursued here. Logical BDI models are formalised theories of qualitative rational decision making. The rationality assumptions incorporated in the usual axioms for belief, desire and intention imply that the logical BDI models can only be regarded as strong idealisations of human belief-formation and decision making. This is clearly in contrast with the view we endorse here, which fundamentally stems from Damasio's radical critique to such a Cartesian stance (Damasio, 2005).

Nevertheless, probabilistic approaches such as that by Conati (2002) share with us at least the effort in devising a theoretical model which is well grounded in probabilistic structures informed by psychological theoretical constraints.

As an alternative, the use of (suitably extended) general-purpose cognitive and agent architectures to model emotion theories, are suitable models of the basic design-level structure of cognitive systems or agents, respectively. These are concerned with the description of their domain-independent, basic computational structures (e.g., memories, representational systems) and processes (e.g., a repeated sense-think/ decide-act cycle): " an architecture appears as the hardwired part of an agent that cannot be changed (at least not easily)" (Reisenzein et al., 2013).

All these proposals do address the implementation model explanation, though, different from us, the ultimate neural realisation level is seldom considered.

Interestingly, the role of architectures in computational modelling extends beyond these practical benefits: architectures can also play important theoretical roles. Sloman (1992) even argues that an understanding of emotions can only be attained by considering the role of emotions in cognitive architectures: "A proper analysis of the concept of an affective state or process must be based on a more general theory of the coarse-grained architecture of mind"

In other terms, this is equivalent to state that the implementation model actually *is* the model. This view is akin to some argument that has raised a fierce debate in the philosophy of science (see the discussion by Frigg and Reiss, 2009). Philosophy of science

has traditionally placed scientific theories in a central role, and has reduced the problem of mediating between theories and the world to formal considerations. The argument is that many applications of scientific theories, however, involve complex mathematical models whose constitutive equations are analytically unsolvable (Winsberg, 1999). The study of these applications often consists in developing representations of the underlying physics on a computer, and using the techniques of computer simulation in order to learn about the behaviour of these systems. In many instances, these computer simulations are not to be considered simple number-crunching techniques. They involve a complex chain of inferences that serve to transform theoretical structures into specific concrete knowledge of physical systems (Winsberg, 1999). It is beyond the scope of this dissertation to further weigh up this controversy, but clearly this is a different perspective from ours.

For instance, BDI logical explanation ad the theoretical level make no assumptions about how beliefs, desires and intentions, and the mental operations performed with them, are computationally implemented. BDI architectures, in contrast, address these questions. As Reisenzein et al. (2013) put it: "They are pieces of software that implement the abstract principles of BDI agents".

They basically execute a fixed perception-deliberation-action: 1) perception of events; 2) updating of the belief and desire (goal) base (adding new beliefs and desires, dropping false beliefs and satisfied or impossible desires); 3) generating options (plans) (deliberating); 4. choosing the most suitable plan (generating an intention); and 5) executing intentions (acting) (Reisenzein et al., 2013). In this respect, affective agent architectures such as FAtiMA (Hudlicka, 2004) and MAMID Hudlicka (2008) owe credits to the EMA model (Marsella and Gratch, 2009) - deeply rooted in symbolic reasoning - , and, in turn, to classic appraisal theory. Indeed, the valence and intensity of emotions are computed in MAMID as linear combinations of several eliciting factors, a paradigmatic example of an emergent model (see Figure 2.4, Chapter 2, and related discussion).

As a final comment, the AI-oriented approaches - either when they address the theoretical modelling level or the computational/realisation modelling level -, overlook by and large the issue of embodiment and the neurobiological basis of affect construction. To sum up, all the above issues entailed by the AI-oriented research program, make our model to stand apart.

Robotic models Roboticists' view is the closest one to our perspective and motivations (Asada, 2015; Wiese et al., 2017). They do address embodiment and enactment, though the "bodyware" puts severe constraints (degree of freedom, real-time, etc.) and limitations. Resulting actual working implementations on real hardware, show profound differences between simulators and human brains/bodies (Wiese et al., 2017).

Results so far achieved are often provided in terms of software/hardware architectures, hence implementation models, often lacking of a clear theoretical model guiding the design.

Even when loosely inspired by neuroscience, more than often models boil down to implementation models, that beside specific details, are just fable variations over the same leitmotiv. For instance, Lim and Okuno (2015) propose a cross-modal associative a long-term memory for encoding feelings, formalised in the shape of a Gaussian Mixture Model (GMM) representation, whose parameters are learned via the expectation-

maximization algorithm. Barros and Wermter (2016) address the same issue but the associative module is provided in the form of a self-organising layer, namely a Self-Organizing Map (SOM) to learn emotion concepts. SOMs are neural networks trained in an unsupervised fashion to create a topological grid that represents the input stimuli (each neuron is trained to be a prototype of the input stimuli). So, after training, each neuron will have a strong emotion representation and neurons which are neighbours are related to similar expressions.

These two solutions might seem, *prima facie*, very different. However, their comparison is strictly related to the level we are considering. At the computational theory level they basically involve the same model: emotion associative memory is conceived as a discrete latent variable model representation solving an (unsupervised) clustering problem. Indeed, it has been mathematically shown by Heskes (2001) that SOMs are nothing but regularised Mixture Models. This result also shows that the difference at the implementation level is minimal, apart from the regularisation term.

In other cases (Barros and Wermter, 2016; Churamani et al., 2017) the direct mapping (via deep neural networks) from the latent space of affective expressions learnt in a bottom-up, feed-forward sweep to facial gestures, synthetic speech etc., is very much reminiscent of the pipeline adopted by machine learning approaches

Indeed the use of deep architectures, can lead to efficient implementation models capable to handle the multimodal nature of emotion (Kim et al., 2013), though, for the reasons we have discussed, these can be hardly assumed as models *tout court*.

The same drawback holds even in cases where deep generative architectures have been proposed (Horii et al., 2016) to actualise mental simulation for inferring the emotion of others from ambiguous multimodal signals. Indeed, Horii et al. (2016) proposal is the most related to the work we have presented in this Thesis, at least at the implementation modelling level. Their theoretical model basically boils down to the generative description of the RBM machine. Having said that, they rely on an acted dataset (IEMOCAP) where facial expressions and hand movements have been recorded with a motion-capture system, thus avoiding cumbersome issues related to actual processing of visual cues. Also, physiological signals, a fundamental aspect of emotional unfolding, are not taken into account.

In other proposals, such as that by Vitale et al. (2014) there is a clear attempt to address the computational theory level so to frame a simulationist approach. However in that case there is no actual account for constraints from neuroscience. As to the implementation level a shallow GP-LVM latent space is used to design a latent affect space. Motor representation is not explicitly addressed and the latent space is assumed as the affective space *tout court*. Only static images are considered as perceptual input so that visual processing is severely limited. Visuomotor mapping is instantiated as a projection to the GP-LVM latent space, which is achieved through a simple variant of PCA.

Notwithstanding such limitations, the work described in (Vitale et al., 2014) has provided early insights for the work we have developed here.

Summary

In this chapter we have discussed some key issues at the heart of the proposed model, both at the theoretical and at the implementation modelling level. These provide a broader perspective on our work, while altogether laying out a scaffolding for a retrospective consideration of work in the field of computational modelling of affect. Eventually, they are likely to pave the way for future research work.

CHAPTER 7

Conclusions

THE presented work concerned with the understanding of affective signals from others, both in human-human and human-agent interactions. In particular, we proposed and formalised a novel, probabilistic model for dynamic affective facial expression processing relying on a mechanism, which involves both facial gesture and autonomic, physiological simulation. This has been done approaching the large body of evidence from affective neuroscience showing that when we observe emotional facial expressions, we react with congruent facial mimicry. The work went through the definition of a multilevel analysis framework, which aimed at devising a computational model informed and constrained by knowledge that we have available both at the psychological and at the neuroscience explanation levels. The work sees its realisation in two interaction experiments that allowed to test the goodness of modelling choices.

This chapter briefly summarises the main contributions and the results obtained in this dissertation.

Summary of Contributions

The thesis is widely multidisciplinary and can be ideally divided into two main parts: the first one, more theoretically, presented the context of the problem and the main psychological, neurobiological and computational models of affect; the second part detailed the mathematical and probabilistic foundations behind the proposed model, along with the signal processing techniques adopted and the results obtained in interaction experiments. These relied also on datasets collected along experiments with human subjects. In summary, the main contributions are the following:

- Chapter 2 provides a broad review of the different contributions to the conundrum of affective modelling. The attempt here was to devise some general guidelines

to better frame the subsequent work, with a particular focus on affective facial expression analysis.

- Chapter 3, *a novel multilevel modelling framework* is defined, starting from Marr's analysis of levels of explanation in cognitive science, which is revised at the light of current research on Bayesian methods; this identifies the theoretical model and the implementation model as the two different but intertwined levels of explanation that constitute a computational model; the theoretical model is shaped in the form of a Probabilistic Graphical Model.

- In the same chapter, motivated by a large body of work in affective neuroscience, we outlined the *architecture of a distributed neural system* for the perception of dynamic facial expressions of emotion. This relies upon the interaction of visceromotor and somatomotor routes as mediated by a core affect component. It can be seen as an extension of Adolph's original neurobiological model, which embeds current insights on the existence of mirroring mechanisms in both the visceromotor and somatomotor activities for the enactment of an emotional response in the perceiver, via internal simulation.

- On such basis, we abstracted *a novel functional architecture* of an agent perceiving affective facial expressions when engaged in a dyadic social interaction with another agent; at the most general level, under suitable assumption, this can be considered as a hierarchical, structured entanglement of stochastic processes.

- Chapter 4 presents the main result: *a novel computational model for the perception of affective facial expressions via simulation*. In the multilevel framework, the model articulates in a theoretical model and an implementation model. The first is formalised as a Bayesian generative model shaped in the form of a Probabilistic Graphical Model, where the graph structure is constrained by the functional architecture posited in Chapter 3. The implementation model is designed by exploiting the representational compositionality of the probabilistic graph, so to address the most suitable instantiation of each sub-graph/component.

- In particular, the central affect components have been instantiated in terms of a Deep Gaussian Process model; this choice has shown to provide a suitable implementation of the close exchange of information that at the neural level occurs between orbitofrontal cortex, the amygdala and the anterior insula.

- A *novel algorithm for simulation-based perception* of affective expressions has been proposed: the implementation model of the PGM, when put into work, relies on forward (generative) and backward (inferential) dynamics; this endows the model with internal simulation capability, that is exploited both for core affect construction (learning stage) and enactment (perceptual/mimicry stage).

- Chapter 5, instead, presents an effective realisation of the implementation model and experiments. To such end, a *novel multimodal affective dataset*, AMHUSE, experimentally acquired by us (and now publicly available for the research community) is put in comparison with another well-known, public multimodal dataset. The chapter follows with a description of the essential steps adopted during data

processing and the construction of the models adopted in two different interaction experiments.

- Chapter 6 discusses, in a broader perspective, the theoretical implications of our work and, accordingly, frames the proposed model within current research.

APPENDIX $\mathcal{A}$

Probabilistic Graphical Models

In this appendix we briefly recall the so called *probabilistic graphical models* which allow to enormously simplify complex joint distributions using conditional independence property in order to achieve factorizations directly by inspection of the graph, and without having to perform any analytical manipulations.

First of all, let recall that conditional independence properties play an important role in using probabilistic models for pattern recognition by simplifying both the structure of a model and the computations needed to perform inference and learning under that model. Furthermore, it is more frequent the case in which two events are independent given an additional event with respect to the case where events are independent *tout court*.

Focusing on random variables, let X, Y and Z be three variables such that the conditional distribution of X given Y and Z does not depend on the values of Y. We say that X is *conditionally independent* of Y given Z if

$$P(X|Y,Z) = P(X|Z).$$

The same can be expressed by considering the joint distribution of X and Y conditioned on Z, i.e.

$$P(X,Y|Z) = P(X|Y,Z)P(Y|Z) = P(X|Z)P(Y|Z).$$

The definition of conditional independence requires that the above factorization holds for all possible values of Z; to denote this property, we use the shorthand notation

$$(X \perp Y \mid Z).$$

Note that this property can be easily extended to sets of random variables $\mathbf{X}$,$\mathbf{Y}$ and $\mathbf{Z}$, in this case we say that $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$ in a distribution P if the latter satisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$.

Appendix A. Probabilistic Graphical Models

A *probabilistic graphical models* is a pair $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ of sets called nodes and edges, respectively. The nodes denote random variables $\mathcal{V} = \{X_1, \ldots, X_n\}$, while the edge set collects directed edge $X_i \to X_j$ between pair of nodes $X_i, X_j \in \mathcal{V}$. We denote by $X_{pa(i)}$ the parents of node X_i in the graph, and by $X_{pred(i)}$ the variables in the graph that are not descendants of X_i. We say that $X_1, \ldots, X_k$ form a path if $X_i \to X_{i+1}$, for all $i = 1, \ldots, k-1$. A cycle in $\mathcal{G}$ is a directed path $X_1, \ldots, X_k$ where $X_1 = X_k$. A graph is acyclic if it contains no cycles. Naturally, to avoid cycles in our graph we cannot have both $X_i \to X_j$ and $X_j \to X_i$.

A *directed acyclic graph* (DAG) is a key concept to define a coherent probabilistic model, as DAGs are the basic graphical representation that underlies Bayesian networks. A formal definition of the semantics of a Bayesian network structure is given in the following.

Definition A.1. A *Bayesian network* (BN) structure $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a DAG encoding for each node X_i the conditional independence assumptions of its nondescendants given its parents:

$$\forall X_i \in \mathcal{V} : \ (X_i \perp \{X_{pred(i) \setminus pa(i)}\} \mid X_{pa(i)}).$$

In other words, $\mathcal{G}$ encodes a set of conditional independence assumptions, called the *local independence*, and denoted by $\mathcal{I}_l(\mathcal{G})$.

However, a BN graph could be defined also in terms of a joint distribution P representable as a set of conditional probability distributions (CPDs) associated with the graph $\mathcal{G}$. Specifically,

Definition A.2. Let P be a distribution over $\mathcal{X}$. We define $\mathcal{I}(P)$ to be the set of *independence assertions* of the form $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ that holds in P.

Given this definition, we can derive that $\mathcal{I}_l(\mathcal{G}) \subseteq \mathcal{I}(P)$, and we say that $\mathcal{G}$ is a *I-map* (independency map) for P. More broadly:

Definition A.3. Let $\mathcal{K}$ be any graph object associated with a set of independencies $\mathcal{I}(\mathcal{K})$. We say that $\mathcal{K}$ is an *I-map* for a set of independencies $\mathcal{I}$ if $\mathcal{I}(\mathcal{K}) \subseteq \mathcal{I}$.

We can now say that $\mathcal{G}$ is an I-map for P if $\mathcal{G}$ is an I-map for $\mathcal{I}(P)$. Let note that, the direction of the inclusion requires that any independence that $\mathcal{G}$ asserts must also holds in P, but not the *vice versa*, that is P could have independencies not reflected in $\mathcal{G}$.

These key concepts allow the compact factorized representation, fundamental for the BN manipulation. Precisely,

Definition A.4. Let $\mathcal{G}$ be a BN graph over the variables $X_1, \ldots, X_n$. We say that a distribution P over the same space factorises according to $\mathcal{G}$ if P can be expressed as a product:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{pa(i)}). \tag{A.1}$$

The individual factors $P(X_i \mid X_{pa(i)})$ are the CPDs or local probabilistic models, and the whole equation is called the *chain rule for BNs*.

Definition A.5. A *BN* is a pair $\mathcal{B} = (\mathcal{G}, P)$ where P factorizes over $\mathcal{G}$, and where P is specified as a set of CPDs associated with $\mathcal{G}$'s nodes. The distribution P is often annotated as $P_{\mathcal{B}}$.

The conditional independence assumptions implied by a BN structure $\mathcal{G}$ allow us to factorize a distribution P for which $\mathcal{G}$ is an I-map into small CPDs as stated in the following theorem (see Koller and Friedman (2009) for the demonstration).

Theorem A.1. *Let $\mathcal{G}$ be a BN structure over a set of RVs $\mathcal{X}$, and let P be a joint distribution over the same space. If $\mathcal{G}$ is an I-map for P, then P factorizes according to $\mathcal{G}$.*

Theorem A.1 proves the factorization of P according to $\mathcal{G}$, but also the converse holds: factorization according to $\mathcal{G}$ implies the associated conditional independencies.

Theorem A.2. *Let $\mathcal{G}$ be a BN structure over a set of random variables $\mathcal{X}$ and let P be a joint distribution over the same space. If P factorizes according to $\mathcal{G}$, then $\mathcal{G}$ is an I-map for P.*

We now move to understand when we can guarantee that an independence $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ holds in a distribution associated with a BN structure $\mathcal{G}$.

Definition A.6. Let $\mathcal{G}$ be a BN structure, and $X_1 \rightleftharpoons \cdots \rightleftharpoons X_n$ a trail in $\mathcal{G}$. Let $\mathbf{Z}$ be a subset of *observed variables*. The trail $X_1 \rightleftharpoons \cdots \rightleftharpoons X_n$ is *active* given $\mathbf{Z}$ if

- Whenever we have a *v*-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants are in $\mathbf{Z}$;
- no other node along the trail is in $\mathbf{Z}$.

Graphs where there are more than one trail between two nodes, give rise to the notion of *d-separation*, standing for directed separation, which provides us with a notion of separation between nodes in a directed graph:

Definition A.7. Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three sets of nodes in $\mathcal{G}$. We say that $\mathbf{X}$ and $\mathbf{Y}$ are *d-separated* given $\mathbf{Z}$, denoted d-sep$_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$, if there is no active trail between any node $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given $\mathbf{Z}$. We use $\mathcal{I}(\mathcal{G})$ to denote the set of independencies that correspond to d-separation: $\mathcal{I}(\mathcal{G}) = \{(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) : \text{d-sep}_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})\}$.

This set is also called the set of *global Markov independencies*.

A first property we want to ensure for d-separation as a method for determining independence is *soundness*: if we find that two nodes X and Y are d-separated given some $\mathbf{Z}$, then we are guaranteed that they are, in fact, conditionally independent given $\mathbf{Z}$. To prove this it holds

Theorem A.3. *If a distribution P factorizes according to $\mathcal{G}$, then $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$.*

In other words, any independence reported by d-separation is satisfied by the underlying distribution. Also the complementary property, the *completeness*, is desirable. This holds if d-separation detects all possible independencies, that is, given two variables X and Y independents given $\mathbf{Z}$, then they are d-separated. To formalize this property, we first introduce the notion of faithful distribution:

Definition A.8. A distribution P is *faithful* to $\mathcal{G}$ if, whenever $(X \perp Y \mid \mathbf{Z}) \in \mathcal{I}(P)$, then d-sep$_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$.

In other words, any independence in P is reflected in the d-separation properties of the graph. We can now introduce this result:

Theorem A.4. *For almost all distributions P that factorize over $\mathcal{G}$, that is, for all distributions except for a set of measure zero in the space of CPD parametrizations, we have that $\mathcal{I}(P) = \mathcal{I}(\mathcal{G})$.*

This shows that there exists a single distribution that is faithful to the graph, that is, where all of the dependencies in the graph hold simultaneously. Second, not only does this property hold for a single distribution, but it also holds for almost all distributions that factorize over $\mathcal{G}$.

These results state that for almost all parametrizations P of the graph $\mathcal{G}$ (that is, for almost all possible choices of CPDs for the variables), the d-separation test precisely characterizes the independencies that hold for P.

Aiming at finding a graph $\mathcal{G}$ that precisely captures the independencies in a given distribution P, we define the *perfect map*:

Definition A.9. We say that a graph $\mathcal{K}$ is a perfect map (P-map) for a set of independencies $\mathcal{I}$ if we have that $\mathcal{I}(\mathcal{K}) = \mathcal{I}$. We say that $\mathcal{K}$ is a *perfect map* for P if $\mathcal{I}(\mathcal{K}) = \mathcal{I}(P)$.

In many domains, we wish to represent distributions over systems whose state changes over time. In these cases, we wish to construct a single, compact model that captures the properties of the system dynamics, and produces distributions over different trajectories.

Our focus is on modelling dynamic settings, where we reason about how the state of the world evolves over time. We can model such settings in terms of a *system state* whose value at time t is a snapshot of the relevant attributes (hidden or observed) of the system at that time. We assume that the system state is represented, as usual, as an assignment of values to some set of random variables $\mathcal{X}$. We use $X_i^{(t)}$ to represent the instantiation of the variable X_i at time t. For a set of variables $\mathbf{X} \subseteq \mathcal{X}$, we use $\mathbf{X}^{(t_1:t_2)}, (t_1 < t_2)$ to denote the set of variables $\mathbf{X}^{(t)} : t \in [t_1, t_2]$. An assignment of values to each variable $X_i^{(t)}$ for each relevant time t correspond to a trajectory in our probability space. Our goal therefore is to represent a joint distribution over such trajectories. Clearly, the space of possible trajectories is a very complex probability space, so representing such a distribution can be very difficult. We therefore make a series of simplifying assumptions that help make this representational problem more tractable.

The first simplification concerns the discretization of the timeline into a set of *time slices*: measurements of the system state taken at intervals that are regularly spaced with a predetermined time granularity Δ. Thus, we can now restrict our set of random variables to $\mathcal{X}^{(0)}, \mathcal{X}^{(1)}, ...$, where $\mathcal{X}^{(t)}$ are the ground random variables that represent the system state at time $t \cdot \Delta$. This assumption simplifies our problem from representing distributions over a continuum of random variables to representing distributions over countably many random variables, sampled at discrete intervals.

Let consider a distribution over trajectories sampled over a prefix of time $t = 1, ..., T$, $P(\mathcal{X}^{(0)}, \mathcal{X}^{(1)}, ...\mathcal{X}^{(T)})$, abbreviated as $P(\mathcal{X}^{(0:T)})$. We can reparametrize the distribution using the chain rule for probabilities, in a direction consistent with time:

$$P(\mathcal{X}^{(0:T)}) = P(\mathcal{X}^{(0)}) \prod_{t=0}^{T-1} P(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(0:t)}). \tag{A.2}$$

A considerably simplification of this formulation is obtained adopting the Markov assumption, that is that the future is conditionally independent of the past given the present:

Definition A.10. We say that a dynamic system over the template variables $\mathcal{X}$ satisfies the Markov assumption if, for $t \geq 0$,

$$(\mathcal{X}^{(t+1)} \perp \mathcal{X}^{0:(t-1)} \mid \mathcal{X}^{(t)}).$$

Such system is called *Markovian*.

The Markov assumption allows to simplify the distribution in Eq. A.2 as:

$$P(\mathcal{X}^{(0:T)}) = P(\mathcal{X}^{(0)}) \prod_{t=0}^{T-1} P(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{t}).$$

A last simplification assumption concerns the system stationarity:

Definition A.11. We say that a Markovian dynamic system is *stationary* (also called *time invariant* or *homogeneous*) if $P(\mathcal{X}^{(t+1)} \mid P(\mathcal{X}^{t}))$ is the same at all t. In this case we can represent the process using a transition model $P(\mathcal{X}' \mid \mathcal{X})$, so that, for any $t \geq 0$,

$$P(\mathcal{X}^{(t+1)} = \xi' \mid \mathcal{X}^{t} = \xi) = P(\mathcal{X}' = \xi' \mid \mathcal{X} = \xi).$$

Definition A.12. A *2-time-slice Bayesian network* (2-TBN) for a process over $\mathcal{X}$ is a conditional Bayesian network over $\mathcal{X}'$ given $\mathcal{X}_I$, where $\mathcal{X}_I \subseteq \mathcal{X}$ is a set of interface variables.

Remembering that, in a conditional Bayesian network, only the variables $\mathcal{X}'$ have parents or CPDs. The interface variables $\mathcal{X}_I$ are those variables whose values at time t have a direct effect on the variables at time $t + 1$. Thus, only the variables in $\mathcal{X}_I$ can be parents of variables in $\mathcal{X}'$. Overall, the 2-TBN represents the conditional distribution:

$$P(\mathcal{X}' \mid \mathcal{X}) = P(\mathcal{X}' \mid \mathcal{X}_I) = \prod_{i=1}^{n} P(\mathcal{X}'_i \mid X'_{pa(i)}).$$

Definition A.13. A *dynamic Bayesian network* (DBN) is a pair $\langle \mathcal{B}_0, \mathcal{B}_{\rightarrow} \rangle$, where $\mathcal{B}_0$ is a Bayesian network over $\mathcal{X}^{(0)}$, representing the initial distribution over states, and $\mathcal{B}_{\rightarrow}$ is a 2-TBN for the process. For any desired time span $T \geq 0$, the distribution over $\mathcal{X}^{(0:T)}$ is defined as a unrolled Bayesian network, where, for any $i = 1, ..., n$:

- the structure and CPDs of $\mathcal{X}_i^{(0)}$ are the same as those for $\mathcal{X}_i$ in $\mathcal{B}_0$,

- the structure and CPD of $\mathcal{X}_i^{(t)}$ for $t > 0$ are the same as those for $\mathcal{X}_i'$ $\mathcal{B}_{\rightarrow}$.

Thus, we can view a DBN as a compact representation from which we can generate an infinite set of Bayesian networks (one for every $T > 0$).

APPENDIX $\mathcal{B}$

The Free Energy Theorem

Use

- $\mathbf{O} = \{O_1, \ldots, O_{j-1}, O_j, O_{j+1}, \ldots, O_m\}$, a collection of *observable* random variables;
- $\mathbf{Z} = \{Z_1, \ldots, Z_{j-1}, Z_j, Z_{j+1}, \ldots, Z_n\}$, a collection of *hidden* or latent random variables

Let $P_{\mathbf{O},\mathbf{Z}}(\mathbf{o}, \mathbf{z})$ be the joint distribution of $\{\mathbf{O}, \mathbf{Z}\}$ and $Q_{\mathbf{Z}}(\mathbf{z})$ an arbitrary probability distribution or density with respect to the Lebesgue measure.

Definition B.1 (Free energy)**.**

$$\begin{aligned}
\mathcal{F}\left(Q_{\mathbf{Z}}\right) &:= \int_{\mathbf{Z}} Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{o}, \mathbf{z})}{Q_{\mathbf{Z}}(\mathbf{z})}\, d\mathbf{z} \\
&:= \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{O}, \mathbf{Z})}{Q_{\mathbf{Z}}(\mathbf{Z})}\right] \\
&= \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O}, \mathbf{Z})\right] - \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log Q_{\mathbf{Z}}(\mathbf{Z})\right] \\
&:= \mathcal{U}(\mathbf{O}) + \mathcal{H}(\mathbf{Z})
\end{aligned}$$

where $\mathcal{U}(\mathbf{Z})$ is the *internal energy* and $\mathcal{H}(\mathbf{Z})$ the Shannon or *canonical entropy* of the collection of r.v. $\mathbf{Z}$.

The following fundamental relation concerning the log evidence of observations, namely, $\log P_{\mathbf{O}}(\mathbf{o})$ and $\mathcal{F}\left(Q_{\mathbf{Z}}\right)$ holds:

Prop. B.1.

$$\log P_{\mathbf{O}}(\mathbf{o}) = \mathcal{F}\left(Q_{\mathbf{Z}}\right) + KL\left(Q_{\mathbf{Z}} \| P_{\mathbf{Z}|\mathbf{O}}\right), \tag{B.1}$$

where KL $\left(Q_{\mathbf{Z}} \| P_{\mathbf{Z}|\mathbf{O}}\right) = \int_{\mathbf{Z}} Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{Q_{\mathbf{Z}}(\mathbf{z})}{P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o})}\, d\mathbf{z}$ *is the relative entropy or Kullback-Leibler divergence (Cover and Thomas, 1991) between* $Q_{\mathbf{Z}}$ *and the posterior distribution* $P_{\mathbf{Z}|\mathbf{O}}$.

Proof. Using the conditional probability definition and taking logs,

$$\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z}) = \log P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o}) + \log P_{\mathbf{O}}(\mathbf{o}) \tag{B.2}$$

which rearranges to

$$\log P_{\mathbf{O}}(\mathbf{o}) = \log P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z}) - \log P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o}). \tag{B.3}$$

Then Equation B.3 grants that

$$\log P_{\mathbf{O}}(\mathbf{o}) = \log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z})}{Q_{\mathbf{Z}}(\mathbf{z})} - \log \frac{P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o})}{Q_{\mathbf{Z}}(\mathbf{z})}, \tag{B.4}$$

and multiplying both sides by $Q_{\mathbf{Z}}(\mathbf{z})$ we obtain

$$Q_{\mathbf{Z}}(\mathbf{z}) \log P_{\mathbf{O}}(\mathbf{o}) = Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z})}{Q_{\mathbf{Z}}(\mathbf{z})} - Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o})}{Q_{\mathbf{Z}}(\mathbf{z})}. \tag{B.5}$$

By integrating with respect to $\mathbf{z}$:

$$\int_{\mathbf{Z}} Q_{\mathbf{Z}}(\mathbf{z}) \log P_{\mathbf{O}}(\mathbf{o})\, d\mathbf{z} = \int_{\mathbf{Z}} Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z})}{Q_{\mathbf{Z}}(\mathbf{z})}\, d\mathbf{z} - \int_{\mathbf{Z}} Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o})}{Q_{\mathbf{Z}}(\mathbf{z})}\, d\mathbf{z}. \tag{B.6}$$

$Q_{\mathbf{Z}}(\mathbf{z})$ is an arbitrary, but normalised distribution, i.e., $\int_{\mathbf{z}} Q_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} = 1$, and using Def.B.1, we obtain (B.1) □

Recall that the basic property of the relative entropy is stated by the following (Cover and Thomas, 1991, Theorem 8.6.1)

Theorem B.2 (Cover and Thomas, 1991, Theorem 8.6.1)**.**

$$KL(Q\|P) \geq 0 \tag{B.7}$$

with equality iff $Q = P$ *almost everywhere (a.e.)*

Then, the free energy $\mathcal{F}(Q_{\mathbf{Z}})$ is a *lower* bound on the log evidence of observations $\log P_{\mathbf{O}}(\mathbf{o})$:

Prop. B.3.

$$\log P_{\mathbf{O}}(\mathbf{o}) \geq \mathcal{F}(Q_{\mathbf{Z}}) \tag{B.8}$$

Proof. Follows directly from Eqs. B.1 and B.7. □

Definition B.2 (Mean field approximation)**.** Let $Q_{Z_i}(z_i)$ be the probability distribution of Z_i, the ith element of $\mathbf{Z}$. Then

$$\mathcal{Q} := \left\{ Q_{\mathbf{Z}}(\mathbf{z}) : Q_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{m} Q_{Z_i}(z_i) \right\}, \tag{B.9}$$

is called the mean-field approximation of distribution Q.

Remark. Clearly, the following trivially holds:

$$\mathbb{E}_{Q_{\mathbf{Z}}}[\mathbf{Z}] = \prod_{i=1}^{m} \mathbb{E}_{Q_{Z_i}}[Z_i]\,. \tag{B.10}$$

Lemma B.4. *Under the assumption that Q is factorised according to the mean-field approximation $\mathcal{Q}$ (Def. B.2), then*

$$\mathcal{F}(Q_{\mathbf{Z}}) = -KL\left(Q_{Z_j} \| \exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\ i\neq j}}^{m} Q_{Z_i}}[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})]\right\}\right) - \sum_{\substack{k=1\\ k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}[\log Q_{Z_k}(Z_k)] \tag{B.11}$$

Proof.

$$\begin{aligned}
&\mathcal{F}(Q_{\mathbf{Z}}) = \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})}{Q_{\mathbf{Z}}(\mathbf{Z})}\right] \\
&= \mathbb{E}_{Q_{\mathbf{Z}}}[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z}) - \log Q_{\mathbf{Z}}(\mathbf{Z})] \\
&= \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z}) - \log \prod_{k=1}^{m} Q_{Z_k}(Z_k)\right] \\
&= \mathbb{E}_{Q_{\mathbf{Z}}}[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})] - \sum_{k=1}^{m} \mathbb{E}_{Q_{\mathbf{Z}}}[\log Q_{Z_k}(Z_k)] \\
&= \mathbb{E}_{Q_{Z_j}}\left[\mathbb{E}_{\prod_{\substack{i=1\\ i\neq j}}^{m} Q_{Z_i}}[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})]\right] - \sum_{k=1}^{m} \mathbb{E}_{Q_{Z_k}}[\log Q_{Z_k}(Z_k)] \\
&= \mathbb{E}_{Q_{Z_j}}\left[\mathbb{E}_{\prod_{\substack{i=1\\ i\neq j}}^{m} Q_{Z_i}}[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})]\right] - \mathbb{E}_{Q_{Z_j}}\left[\log Q_{Z_j}(Z_j)\right] - \sum_{\substack{k=1\\ k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}[\log Q_{Z_k}(Z_k)]\,.
\end{aligned}$$

Using the log-exp transformation:

$$
\begin{aligned}
&= \mathbb{E}_{Q_{Z_j}}\left[\log\left(\exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right\}\right)\right] - \mathbb{E}_{Q_{Z_j}}\left[\log Q_{Z_j}(Z_j)\right] \\
&\qquad - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right] \\
&= \mathbb{E}_{Q_{Z_j}}\left[\log\left(\exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right\}\right) - \log Q_{Z_j}(Z_j)\right] \\
&\qquad - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right] \\
&= \mathbb{E}_{Q_{Z_j}}\left[\log \frac{\exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right\}}{Q_{Z_j}(Z_j)}\right] - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right] \\
&= -\mathbb{E}_{Q_{Z_j}}\left[\log \frac{Q_{Z_j}(Z_j)}{\exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right\}}\right] - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right] \\
&= -\mathrm{KL}\left(Q_{Z_j} \| \exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right\}\right) - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right].
\end{aligned}
$$

□

Definition B.3.

$$\mathcal{U}_j(Z_j) = \mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right] \tag{B.12}$$

is the expected internal energy of r.v. Z_j

Theorem B.5 (Free energy theorem)**.** *The free-energy is maximised with respect to* $Q^*_{Z_j}(z_j)$ *when*

$$Q^*_{Z_j}(z_j) \propto \exp\left\{\mathcal{U}_j(Z_j)\right\} \tag{B.13}$$

Proof. We want to find the optimal approximating distribution

$$Q^*_{Z_j}(z_j) = \arg\max_{Q_{Z_j}\in Q} \mathcal{F}\left(Q_{\mathbf{Z}}\right)$$

Use Eq. B.11 to write:

$$\max_{Q_{Z_j}\in Q} \mathcal{F}\left(Q_{\mathbf{Z}}\right) = \max_{Q_{Z_j}\in Q}\left\{-\mathrm{KL}\left(Q_{Z_j} \| \exp\left\{\mathcal{U}_j(Z_j)\right\}\right) - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right]\right\} \tag{B.14}$$

The term $-\sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right]$ does not depend on Q_{Z_j}, further, Eq. B.7 grants that $\mathrm{KL}\left(Q_{Z_j} \| \exp\left\{\mathcal{U}_j(Z_j)\right\}\right) \geq 0$.

Thus, to maximize $\mathcal{F}(Q_{\mathbf{Z}})$ we need to minimize the KL term. Up to constant terms:

$$\max_{Q_{Z_j} \in Q} \mathcal{F}(Q_{\mathbf{Z}}) \propto \min_{Q_{Z_j} \in Q} \left\{ \mathrm{KL}\left(Q_{Z_j} \| \exp\left\{\mathcal{U}_j(Z_j)\right\}\right)\right\} \tag{B.15}$$

The KL term in the last equation is minimised precisely when the two terms are equivalent a.e., thus the optimal distribution for which $Q^*_{Z_j}(z_j) = \arg\max_{Q_{Z_j} \in Q} \mathcal{F}(Q_{\mathbf{Z}})$ is

$$Q^*_{Z_j}(z_j) = \exp\left\{\mathcal{U}_j(Z_j)\right\} + const = \arg\max_{Q_{Z_j} \in Q} \mathcal{F}(Q_{\mathbf{Z}}) \tag{B.16}$$

□

APPENDIX C

Gaussian Processes

In traditional supervised learning, we are given a training dataset of n observations $D = \{(x_i, y_i) \mid i = 1, \ldots, n\}$ which collects input-output pairs (x_i, y_i) where the input $x_i \in \mathcal{R}^q$ and the output y_i denotes a scalar output. For sake of notation, the column vector inputs are stacked in a $q \times n$ matrix X, and the outputs are collected in the vector y, thus writing $D = (X, y)$. The goal is to learn input-output mappings f between empirical data, such that

$$y_i = f(x_i). \tag{C.1}$$

In this perspective Gaussian Processes (GPs) can be adopted as a non-parametric approach, in that it finds a distribution over all the possible functions f that are consistent with the observed data.The idea is to define a prior over functions which can be converted into a posterior once we have seen some data. Since it seems difficult to define such a distribution, what we effectively need is to define a distribution over the function's values at the finite set of points $\{x_i\}_{i=1}^n$.

Definition C.1 (Gaussian Process)**.** A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Formally, a GP assumes that $[f(x_1), \ldots, f(x_n)]^T$ are jointly Gaussian with mean function $\mu(x)$ and covariance function $\Sigma(x)$ given by $\Sigma_{ij} = k(x_i, x_j)$, (with k positive definite kernel function), i.e.

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \ldots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \ldots & k(x_n, x_n) \end{bmatrix} \right). \tag{C.2}$$

In a more compact notation we can write

$$f(x) \sim \mathcal{GP}(\mu(x), k(x, x')),$$

where the mean function $\mu(x)$ and the covariance function $k(x, x')$ are defined as

$$\begin{aligned} \mu(x) &= \mathbb{E}[f(x)] \\ k(x, x') &= \mathbb{E}[(f(x) - \mu(x))(f(x') - \mu(x'))] \end{aligned}$$

At a first glance Gaussian processes presents two main strengths: the representative power, being distributions over generic functions and the tractability of the distributions involved in the process. Indeed, the marginalisation property, or *consistency* requirement (Rasmussen and Williams, 2006), of normal distributions permits to consider a finite set of function instantiations, marginalising over all the other infinite unseen inputs. This simply means that if a GP specifies the distribution in (C.1), then the same GP also specifies the single draw $f(x_1) \sim \mathcal{N}(\mu(x_1), k(x_1, x_1))$, where $\Sigma_{11} = k(x_1, x_1)$ is the relevant submatrix of Σ.

Moreover, when modelling a stationary situation, the mean function is typically (Damianou, 2015; Murphy, 2012; Rasmussen and Williams, 2006) selected as $\mu(x) = 0$, that permits to describe the Gaussian process only relying on the second order statistic, namely the covariance function. The key idea in using the kernel functions as covariance is that if x and x' are close in the input space, then we expect the output of the function at those points to be similar or close, too.

For example, a typical squared exponential (SE) (or radial basis function (RBF)) kernel function is defined as

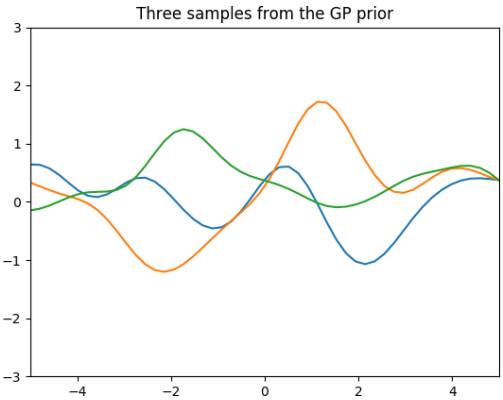
$$k_{\text{SE}}(x, x') = \sigma_{\text{SE}}^2 \exp\left\{-\frac{1}{2}\left(\frac{x - x'}{\lambda}\right)^2\right\} \tag{C.3}$$

where the lengthscale $\lambda > 0$ and the amplitude $\sigma_{\text{SE}} > 0$ are the hyperparameters. The former describes how smooth the function is, while the latter determines the variation of function values from their mean. In Figure C.1a are shown three latent functions randomly drawn from a zero-mean GP *prior* with SE kernel function and $\lambda = 0.8, \sigma_{SE} = 0.5$, that in probabilistic terms gives (omitting the conditioning on hyperparameters) the conditional density $P(f|X) = \mathcal{N}(f|0, K)$.

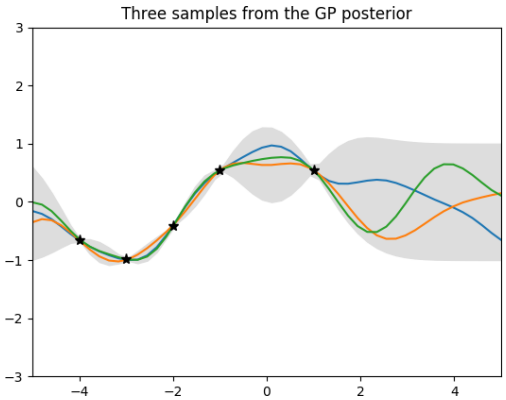
As said above, we are primarily interested in capturing the knowledge that the training data provides rather then drawing random functions from the prior, by just transforming the prior into a posterior.

Let us first consider the simple case where the observations are noise free, that is when we have the dataset $D = \{(x_i, f_i) \mid i = 1, \ldots, n\}$. By denoting with X_* and f_* the input-output test and by using standard rules for conditioning Gaussians (see (Bishop, 2006b)), the posterior on test has the following form

$$P(f_*|X_*, X, f) = \mathcal{N}(f_*|\mu_*, \Sigma_*) \tag{C.4}$$



(a) *Three samples from the GP prior with SE kernel function. Hyperparameters:* $\lambda = 0.8$ *and* $\sigma_{SE} = 0.5$.



(b) *Three samples from the GP posterior obtained observing five data points.*

Figure C.1: *Example of a Gaussian Process fitting observed data generated from a cosine function. Each function sample is presented in three colours; the observed points are plotted with black asterisks and two times the standard deviation is grey shaded.*

where

$$\mu_* = \mu(X_*) + K_*^T K^{-1}(f - \mu(X)) = K_*^T K^{-1} f$$
$$\Sigma_* = K_{**} - K_*^T K^{-1} K_*$$

Being $K_* = k(X, X_*)$ the cross-covariance between training and test inputs and $K_{**} = k(X_*, X_*)$ the covariance between test inputs. Therefore, in the noise free case, GP yields $f(x)$ with no uncertainty acting as an interpolator of the training data (as shown in Fig. C.1b).

Let us now consider the more realistic case in which we are asked to make predictions using noisy observations, meaning that we do not have access to function values themselves, but only to noisy versions thereof, i.e. when

$$y_i = f(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2). \tag{C.5}$$

Assuming additive independent identically distributed Gaussian noise ϵ with variance σ_ϵ^2, the prior on the noisy observations becomes

$$k_{\text{SE}}(x, x') = \sigma_{\text{SE}}^2 \exp\left\{-\frac{1}{2}\left(\frac{x - x'}{\lambda}\right)^2\right\} + \sigma_\epsilon^2 \mathbf{I}, \tag{C.6}$$

reason why the (C.4) may be rewritten as

$$P(f_*|X_*, X, y) = \mathcal{N}(f_*|\mu_*, \Sigma_*) \tag{C.7}$$

where (assuming that the mean is zero for notational simplicity)

$$\mu_* = K_*^T K_y^{-1} y$$
$$\Sigma_* = K_{**} - K_*^T K_y^{-1} K_*$$

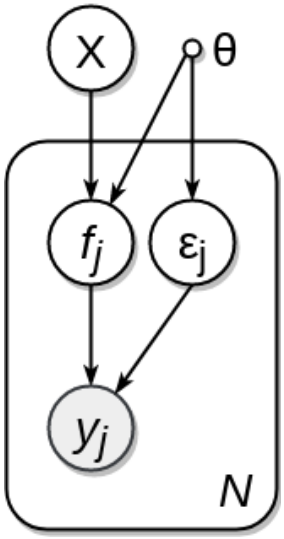


Figure C.2: *Visualisation of a Gaussian Process Latent Variable Model, expressed as a PGM. The grey circles represent the observed variables, while the white circles the latent ones.*

being K_y the covariance of the observed noisy responses.

A direct extension of the classical supervised GP is the Gaussian Process Latent Variable Model (GPLVM) (Lawrence, 2004, 2005), a probabilistic model for non-linear dimensionality reduction and unsupervised learning that can be seen as a non-linear generalisation of probabilistic PCA (Tipping and Bishop, 1999). In this model the inputs are unobserved and are treated as latent variables, while only the output are given by applying multiple GPs. Moreover, due to the difficulty to integrate out the latent variables, they are optimized by using variational algorithms based on mean field approximations, as in some Bayesian extensions of PPCA (Tipping and Bishop, 1999).

In this scenario, we assume that data are represented through multi-dimensional variables which are collected in matrices where rows correspond to instances and columns correspond to dimensions (or features). Thus, the output $Y \in \mathcal{R}^{n \times p}$ can be approximated by a lower dimensional matrix $X \in \mathcal{R}^{n \times q}$ through a vector valued non-linear function mapping the i-th row x_i of X in correspondent row y_i of Y, such that

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}) \tag{C.8}$$

A GPLVM provides a solution to this approximation by employing p independent GPs as prior for latent mapping $f(X) = (f_1(X), \ldots, f_p(X))$ such that

$$f_j(X) \sim \mathcal{GP}(0, k_f(X, X')), \quad j = 1, \ldots, p.$$

A GPLVM can be represented via a PGM, as shown in Fig. C.2.

In this setting, the generative procedure shown in (C.2) can be rewritten as

$$y_{ij} = f_j(x_i) + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}),$$

providing the basic independence assumptions which allows to write the likelihood of

the data given the inputs X by marginalizing out the function evaluations F as

$$\begin{aligned} P(Y|X) &= \int P(Y|F)P(F|X)dF \\ &= \int \prod_{i=1}^{n}\prod_{j=1}^{p} P(y_{ij}|f_{ij}) \prod_{j}^{p} P(f_j|X)dF \\ &= \prod_{j=1}^{p} \mathcal{N}(y_j|0, K_{ff} + \sigma_\epsilon^2 \mathbf{I}) \end{aligned}$$

The above partial tractability of the model gives rise to a straightforward MAP training procedure where the latent inputs X are selected according to

$$X_{\text{MAP}} = \arg\max_{X} \; P(Y|X)P(X) \tag{C.9}$$

This is the approach suggested by Lawrence (2005). The same optimization holds also for the hyperparameters (not mentioned here for sake of simplicity).

The optimisation process of a GPLVM made adopting a MAP approach, however, hides two main drawbacks as pointed out in Damianou (2015). First, the sensitiveness to overfitting caused by the fact that MAP is a point estimation and tends to maximise with respect to X and the hyperparameters. Second, this approach cannot be used for automatic selection of the latent dimensionality. An increase on the latent dimensions, indeed, results in an increase of the optimal size.

For these reasons, in Titsias and Lawrence (2010) it is presented a variational Bayesian approach to marginalise the latent variables X, optimising the resulting lower bound on the marginal likelihood with respect to the hyperparameters.

Following the typical Bayesian methods, the goal is to compute the logarithm of the marginal likelihood of the data

$$\begin{aligned} \log P(Y) &= \log \int P(Y, X)dX \\ &= \log \int P(Y|X)P(X)dX. \end{aligned}$$

Anyway the non-linear presence of X in the covariance matrix term makes the previous likelihood intractable. In this case, the standard variational Bayesian approach (Beal, 2003) comes to give a slight hand, refer to Chap. 6 for a wider derivation. The basic idea is to consider a variational distribution $Q(X)$ that approximates the true posterior distribution $P(X|Y)$ and has a Gaussian form over the latent variables

$$Q(X) = \prod_{i=1}^{q} \mathcal{N}(x_i|m_i, S_i).$$

Based on Jensen's inequality (Jensen, 1906), we can obtain the standard variational

lower bound on the log marginal likelihood

$$\begin{aligned}\log P(Y) &= \log \int P(Y|X)P(X)dX \\ &= \log \int Q(X)\frac{P(Y|X)P(X)}{Q(X)}dX \\ &\geq \int Q(X)\log\frac{P(Y|X)P(X)}{Q(X)}dX \\ &= \int Q(X)\log P(Y|X)dX - \int Q(X)\log\frac{Q(X)}{P(X)}dX \\ &= \mathcal{F}(Q(X)) - \mathrm{KL}(Q(X)\|P(X)) \end{aligned} \tag{C.10}$$

where

$$\begin{aligned}\mathcal{F}(Q(X)) &= \sum_{j=1}^{p}\int Q(X)\log P(y_j|X)dX \\ &= \sum_{j=1}^{p}\mathcal{F}_j(Q(X))\end{aligned}$$

But, even in this case, the term $\mathcal{F}_j(Q(X))$ is computationally intractable because it involves an analytically intractable integration.

A possible solution comes from Lawrence (2007) and results in the introduction of m auxiliary inducing variables $U \in \mathbf{R}^{m\times p}$ that constitute the latent function evaluations at a set of pseudo-inputs $Z \in \mathbf{R}^{m\times q}$. The augmented full joint probability takes then the form

$$\begin{aligned}P(Y,F,U,X,Z) &= P(Y|F)P(F|U,X,Z)P(U|Z)P(X) \\ &= \prod_{j=1}^{p}P(y_j|f_j)p(f_j|u_j,X,Z)P(u_j|Z)P(X)\end{aligned} \tag{C.11}$$

where

$$p(f_j|u_j,X,Z) = \mathcal{N}(f_j|K_{fu}K_{uu}^{-1}u_j, K_{ff} - K_{fu}K_{uu}^{-1}K_{uf})$$

is the conditional prior and

$$P(u_j|Z) = \mathcal{N}(u_j|o, K_{uu})$$

is the marginal GP prior over the inducing variables. Here, K_{uu} denotes the covariance matrix constructed by evaluating the covariance function on the inducing points and K_{uf} is the cross-covariance between the inducing and the latent points. The above explained processes follows the graphical representation in Fig. C.3.

The lower bound of the intractable term of $\mathcal{F}_j(Q(X))$ results in

$$\begin{aligned}\log P(y_j \mid X) \geq \int \phi(u_j)\log\frac{P(u_j)\mathcal{N}(y_j|\alpha_j,\beta^{-1}\mathbf{I}_N)}{\phi(u_j)}du_d \\ -\frac{\beta}{2}Tr(K_{ff} - K_{fu}K_{uu}^{-1}K_{uf})\end{aligned} \tag{C.12}$$

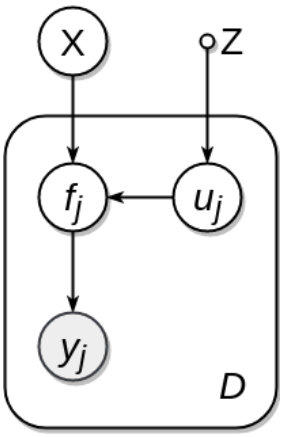


Figure C.3: *Visualisation of a Variational Gaussian Process Latent Variable Model, expressed as a PGM. The grey circle represent the observed variable, while the white circles the latent ones.*

with $\alpha_j = K_{fu}K_{uu}^{-1}u_j$ and $\phi(u_j)$ is a variational distribution over the inducing variables u_j. Replacing the computed lower bound (Eq. C.12) in the functional $\mathcal{F}_j(Q(X))$ we have

$$\mathcal{F}_j(Q(X)) \geq \int Q(X) \left[\int \phi(u_j) \log \frac{P(u_j)\mathcal{N}(y_j \mid \alpha_j, \beta^{-1}\mathbf{I}_N)}{\phi(u_j)} du_d \right. \\ \left. -\frac{\beta}{2}Tr(K_{ff}) + \frac{\beta}{2}Tr(K_{uu}^{-1}K_{uf}K_{fu}) \right] dX \quad \text{(C.13)}$$

Performing firstly the integration with respect to X and maximizing the resulting lower bound with respect to $\phi(u_j)$ permits to obtain the closed form of the lower bound

$$\mathcal{F}_j(Q(X)) \geq \log \left[\frac{\beta^{\frac{N}{2}}|K_{uu}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}|\beta\mathbf{\Psi}_2 + K_{uu}|^{\frac{1}{2}}} e^{-\frac{1}{2}y_j^T W y_j} \right] \\ - \frac{\beta\psi_0}{2} + \frac{\beta}{2}Tr(K_{uu}^{-1}\mathbf{\Psi}_2) \quad \text{(C.14)}$$

where $W = \beta\mathbf{I}_N - \beta^2\mathbf{\Psi}_1(\beta\mathbf{\Psi}_2 + K_{uu})^{-1}\mathbf{\Psi}_1^T$, while

$$\psi_0 = \sum_{i=1}^{N} \int k(x_i, x_i)\mathcal{N}(x_i \mid \mu_i, S_i)dx_n$$

$\mathbf{\Psi}_1$ is a matrix $N \times M$

$$(\mathbf{\Psi}_1)_{nm} = \int k(x_n, z_m)\mathcal{N}(x_n \mid \mu_n, S_n)dx_n$$

and $\mathbf{\Psi}_2$ is an $M \times M$ matrix such that $\mathbf{\Psi}_2 = \sum_{n=1}^{N} \mathbf{\Psi}_2^n$ and

$$(\mathbf{\Psi}_2^n)_{mm'} = \int k(x_n, z_m)k(z_{m'}, x_n)\mathcal{N}(x_n \mid \mu_n, S_n)dx_n$$

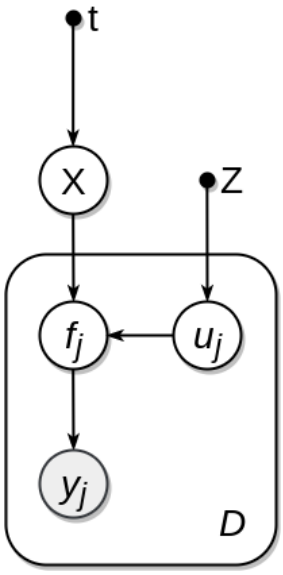


Figure C.4: *Visualisation of a variational GPLVM dynamical model, expressed as a PGM. The grey circle represent the observed variable, while the white circles the latent ones.*

As pointed in Wang et al. (2006, 2008); Damianou et al. (2011) when dealing with time-series data, as the ones considered in this thesis, the learning of latent space could benefit from the introduction of a dynamical prior. The 'spatial constraint' introduced in the classical GP, in fact, here is extended also to time. In particular we wish that two points x and x' that are temporally close, are nearby also in the latent space, resulting in a non-linear but smooth mapping.

This results in the so called GP-LVM dynamical model, where the observed data y_i are considered drawn from a multivariate time series $\{y_i, t_i\}_{i=1}^n$ together with the time $t_i \in \mathcal{R}_+$ at which they are observed. In particular, a *temporal* latent function $\mathbf{x}(t) = (x_1(t), \dots, x_Q(t))$ with $\mathbf{x}(t) \in \mathcal{R}^q$ governs the generation of latent variables, such that the individual components are taken from a Gaussian process with covariance function $k_x(t, t')$

$$x_j(t) \sim \mathcal{GP}(0, k_x(t, t')), \quad j = 1, \dots, q.$$

The covariance function k_x, in this case, determines the properties of each temporal latent function $x_j(t)$ and, based on the temporal behaviour of data, it could be assumed Markovian (e.g. Ornstein-Uhlenbeck (Uhlenbeck and Ornstein, 1930b)) or non-Markovian (e.g. squared-exponential) process.

According to the updated PGM shown in Fig. C.4, the full joint probability in Eq. C.11 should be rewritten as

$$P(Y, F, U, X, Z \mid t) = \prod_{j=1}^{p} P(y_j|f_j)P(f_j|u_j, X, Z)P(u_j|Z)P(X|t)$$

and the variational lower bound in Eq. C.10 reflects the same behaviour

$$\log P(Y) = \mathcal{F}(Q(X)) - \mathrm{KL}(Q(X)\|P(X|t)).$$

The first term captures the relation between observations and data point through the variational distribution, as happened in Eq. C.10, while the second one only involves the prior over data. Therefore the solution follows the one presented above for the variational GPLVM.

Bibliography

Abbott, L. and Kepler, T. B. (1990). Model neurons: from hodgkin-huxley to hopfield. In *Statistical mechanics of neural networks*, pages 5–18. Springer.

Adams, A. and Robinson, P. (2011). An android head for social-emotional intervention for children with autism spectrum conditions. In *Affective Computing and Intelligent Interaction*, pages 183–190. Springer.

Adolphs, R. (2002a). Neural systems for recognizing emotion. *Current opinion in neurobiology*, 12(2):169–177.

Adolphs, R. (2002b). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and cognitive neuroscience reviews*, 1(1):21–62.

Ahlberg, J. (2010). CANDIDE-3 — An updated parameterized face. Technical Report LiTH-ISY-R-2326, Linköping University, Department of Electrical Engineering, Linköping, Sweden.

Alexander, D. M., Trengove, C., Johnston, P., Cooper, T., August, J., and Gordon, E. (2005). Separating individual skin conductance responses in a short interstimulus-interval paradigm. *Journal of neuroscience methods*, 146(1):116–123.

AlZoubi, O., Calvo, R. A., and Stevens, R. H. (2009). Classification of eeg for affect recognition: An adaptive approach. In *Australasian Conference on Artificial Intelligence*, volume 5866, pages 52–61. Springer.

Anderson, D. J. and Adolphs, R. (2014). A framework for studying emotions across species. *Cell*, 157(1):187–200.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Angela, J. Y. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692.

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-taylor, J. S. (2008). Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems*, pages 17–24.

Arnold, M. B. (1960). *Emotion and personality.* Columbia University Press.

Asada, M. (2015). Towards artificial empathy. *International Journal of Social Robotics*, 7(1):19–33.

Avenanti, A., Bueti, D., Galati, G., and Aglioti, S. M. (2005). Transcranial magnetic stimulation highlights the sensorimotor side of empathy for pain. *Nature neuroscience*, 8(7):955.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *Proc. of the IEEE Int. Conf. on Computer Vision Workshops*, pages 354–361.

Bard, P. (1929). The central representation of the sympathetic system: as indicated by certain physiologic observations. *Archives of Neurology & Psychiatry*, 22(2):230–246.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. MIT press.

Barrett, L. F. and Russell, J. A. (2014). *The psychological construction of emotion*. Guilford Publications.

Barrett, L. F. and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews. Neuroscience*, 16(7):419.

Barros, P. and Wermter, S. (2016). Developing crossmodal expression recognition based on a deep neural model. *Adaptive behavior*, 24(5):373–396.

Beal, M. and Ghahramani, Z. (2003). The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, volume 7, pages 453–464. Oxford University Press.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University of London.

Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D. L. (2004). A bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–356.

Bechara, A. and Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and economic behavior*, 52(2):336–372.

Bechara, A., Damasio, H., and Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3):295–307.

Becker-Asano, C. (2008). *WASABI: Affect simulation for agents with believable interactivity*, volume 319. IOS Press.

Becker-Asano, C., Ogawa, K., Nishio, S., and Ishiguro, H. (2010). Exploring the uncanny valley with geminoid hi-1 in a real-world application. In *Proceedings of IADIS International conference interfaces and human computer interaction*, pages 121–128.

Benedek, M. and Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, 190(1):80–91.

Billard, A. G., Calinon, S., and Dillmann, R. (2016). Learning from humans. In Siciliano, B. and Khatib, O., editors, *Handbook of Robotics*, chapter 74, pages 1995–2014. Springer, Secaucus, NJ, USA, 2nd edition edition.

Bishop, C. M. (2006a). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Bishop, C. M. (2006b). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Boccignone, G., Conte, D., Cuculo, V., and Lanzarotti, R. (2017). AMHUSE: A Multimodal dataset for HUmour SEnsing. In *19th ACM International Conference on Multimodal Interaction (ICMI)*.

Boccignone, G. and Cordeschi, R. (2007). Bayesian models and simulations in cognitive science. In *Models and Simulations 2*, pages 1–16. Tilburg Center for Logic and Philosophy of Science.

Boccignone, G. and Cordeschi, R. (2015). Coping with levels of explanation in the behavioral sciences. *Frontiers in Psychology*, 6:213.

Bonini, L. (2017). The extended mirror neuron network. *The Neuroscientist*, 23(1):56–67.

Braunwald, E., Fauci, A., Isselbacher, K., et al. (1998). *Harrison's Principles of Internal Medicine*. McGraw-Hill Companies, USA.

Breazeal, C. (2003a). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1):119–155.

Breazeal, C. (2003b). Toward sociable robots. *Robotics and autonomous systems*, 42(3):167–175.

Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., and Blumberg, B. (2006). Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Learning*, 11(1-2).

Breazeal, C., Edsinger, A., Fitzpatrick, P., and Scassellati, B. (2001). Active vision for sociable robots. *IEEE Transactions on systems, man, and cybernetics-part A: Systems and Humans*, 31(5):443–453.

Breazeal, C. and Scassellati, B. (1999). How to build robots that make friends and influence people. In *Proceedings IEEE/RSJ International Conference onIntelligent Robots and Systems, IROS'99*, volume 2, pages 858–863. IEEE.

Broekens, J., Degroot, D., and Kosters, W. A. (2008). Formal models of appraisal: Theory, specification, and computational model. *Cognitive Systems Research*, 9(3):173–197.

Brooks, R. A., Breazeal, C., Marjanovic, M., Scassellati, B., and Williamson, M. M. (1999). The cog project: Building a humanoid robot. *Lecture Notes in Computer Science*, pages 52–87.

Bui, T. D. (2004). *Creating emotions and facial expressions for embodied agents*. PhD thesis, Taaluitgeverij Neslia Paniculata.

Calvo, R. and D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37.

Candy, J. V. (2016). *Bayesian signal processing: classical, modern, and particle filtering methods*, volume 54. John Wiley & Sons.

Cannon, W. B. (1927). The james-lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology*, 39(1/4):106–124.

Carr, L., Iacoboni, M., Dubeau, M.-C., Mazziotta, J. C., and Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the national Academy of Sciences*, 100(9):5497–5502.

Cauda, F., Costa, T., Torta, D. M., Sacco, K., D'agata, F., Duca, S., Geminiani, G., Fox, P. T., and Vercelli, A. (2012). Meta-analytic clustering of the insular cortex: characterizing the meta-analytic connectivity of the insula when involved in active tasks. *Neuroimage*, 62(1):343–355.

Ceruti, C., Cuculo, V., D'Amelio, A., Grossi, G., and Lanzarotti, R. (2017). Taking the hidden route: deep mapping of affect via 3D neural networks. In *Automatic affect analysis and synthesis (3AS)*.

Chakrabarti, B., Bullmore, E., and Baron-Cohen, S. (2006). Empathizing with basic emotions: common and discrete neural substrates. *Social neuroscience*, 1(3-4):364–384.

Chater, N., Tenenbaum, J., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291.

Churamani, N., Kerzel, M., Strahl, E., Barros, P., and Wermter, S. (2017). Teaching emotion expressions to a human companion robot using deep neural architectures. In *International Joint Conference on Neural Networks (IJCNN), 2017*, pages 627–634. IEEE.

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied artificial intelligence*, 16(7-8):555–575.

Cordeschi, R. (2002). *The Discovery of the Artificial: Behavior, Mind and Machines Before and Beyond Cybernetics*. Kluwer Academic Publishers.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. Wiley and Sons, New York, N.Y.

Craig, A. (2003). Interoception: the sense of the physiological condition of the body. *Current opinion in neurobiology*, 13(4):500–505.

Craig, A. D. and Craig, A. (2009). How do you feel–now? the anterior insula and human awareness. *Nature reviews neuroscience*, 10(1).

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Cuculo, V., Lanzarotti, R., and Boccignone, G. (2014). Using sparse coding for landmark localization in facial expressions. In *5th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6.

Dalgleish, T. and Power, M. (2005). *Handbook of cognition and emotion*. John Wiley & Sons.

Damasio, A. R. (1997). Towards a neuropathology of emotion and mood. *Nature*, 386(6627):769.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt.

Damasio, A. R. (2005). *Descartes' Error: Emotion, Reason, and the Human Brain*. Penguin.

Damasio, A. R. (2012). *Self comes to mind: Constructing the conscious brain*. Vintage.

Damianou, A. (2015). *Deep Gaussian processes and variational propagation of uncertainty*. PhD thesis, University of Sheffield.

Damianou, A., Ek, C., Titsias, M., and Lawrence, N. (2012). Manifold relevance determination. *arXiv preprint arXiv:1206.4610*.

Damianou, A. and Lawrence, N. (2013). Deep gaussian processes. In *Proceedings of the 16th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, volume 31 of *JMLR Workshop and Conference Proceedings*, pages 207–215, Scottsdale, AZ, USA. JMLR.

Damianou, A., Titsias, M. K., and Lawrence, N. D. (2011). Variational gaussian process dynamical systems. In *Advances in Neural Information Processing Systems*, pages 2510–2518.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. (2016). Variational inference for latent variables and uncertain inputs in gaussian processes. *Journal of Machine Learning Research (JMLR)*, 17(1):1425–1486.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. Appleton, New York.

Daunizeau, J., Friston, K. J., and Kiebel, S. J. (2009). Variational bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D: Nonlinear Phenomena*, 238(21):2089–2118.

Demiris, Y., Aziz-Zadeh, L., and Bonaiuto, J. (2014). Information processing in the mirror neuron system in primates and machines. *Neuroinformatics*, 12(1):63–91.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Dennett, D. (1987). *The Intentional Stance*. MIT Press, Cambridge, MA.

Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., De Luca, V., and Jaggi, M. (2016). Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. *Proceedings of SemEval*, pages 1124–1128.

Desimone, R. and Ungerleider, L. G. (1986). Multiple visual areas in the caudal superior temporal sulcus of the macaque. *Journal of Comparative Neurology*, 248(2):164–189.

Devillers, L. and Vidrascu, L. (2006). Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Ninth International Conference on Spoken Language Processing*.

Diano, M., Tamietto, M., Celeghin, A., Weiskrantz, L., Tatu, M.-K., Bagnis, A., Duca, S., Geminiani, G., Cauda, F., and Costa, T. (2017). Dynamic changes in amygdala psychophysiological connectivity reveal distinct neural networks for facial expressions of basic emotions. *Scientific Reports*, 7:45260.

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

Dornaika, F. and Davoine, F. (2008). Simultaneous facial action tracking and expression recognition in the presence of head motion. *International Journal of Computer Vision*, 76(3):257–281.

Dubois, J. and Adolphs, R. (2015). Neuropsychology: how many emotions are there? *Current Biology*, 25(15):R669–R672.

Edelberg, R. (1993). Electrodermal mechanisms: A critique of the two-effector hypothesis and a proposed replacement. *Progress in electrodermal research*, pages 7–29.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48(4):384.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Ekman, P. and Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

Electrophysiology, T. F. o. t. E. S. o. C. t. N. A. S. (1996). Heart rate variability : Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065.

Ellsworth, P. C. and Scherer, K. R. (2003). Appraisal processes in emotion. *Handbook of affective sciences*, 572:V595.

Eskil, M. T. and Benli, K. S. (2014). Facial expression recognition based on anatomy. *Computer Vision and Image Understanding*, 119:1–14.

Fanello, S. R., Ciliberto, C., Noceti, N., Metta, G., and Odone, F. (2017). Visual recognition for humanoid robots. *Robotics and Autonomous Systems*, 91:151–168.

Ferrari, P. F., Visalberghi, E., Paukner, A., Fogassi, L., Ruggiero, A., and Suomi, S. J. (2006). Neonatal imitation in rhesus macaques. *PLoS Biol*, 4(9):e302.

Fogassi, L. and Ferrari, P. F. (2011). Mirror systems. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(1):22–38.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.

Frigg, R. and Reiss, J. (2009). The philosophy of simulation: hot new issues or same old stew? *Synthese*, 169(3):593–613.

Friston, K. (2008). Hierarchical models in the brain. *PLoS computational biology*, 4(11):e1000211.

Friston, K. and Stephan, K. (2007). Free-energy and the brain. *Synthese*, 159:417–458. Published online.

Gallego, G. and Yezzi, A. (2015). A compact formula for the derivative of a 3-d rotation in exponential coordinates. *Journal of Mathematical Imaging and Vision*, 51(3):378–384.

Gallese, V., Keysers, C., and Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in cognitive sciences*, 8(9):396–403.

Gebhard, P. (2005). Alma: a layered model of affect. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 29–36. ACM.

Giere, R. (1999). Using models to represent reality. In Magnani, L., Nersessian, N., and Thagard, P., editors, *Model-Based Reasoning in Scientific Discovery*, pages 41 – 57, New York. Kluwer/Plenum.

Goldman, A. I. and Sripada, C. S. (2005). Simulationist models of face-based emotion recognition. *Cognition*, 94(3):193–213.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gothard, K. M. (2014). The amygdalo-motor pathways and the control of facial expressions. *Frontiers in neuroscience*, 8.

Grimm, M. and Kroschel, K. (2005). Evaluation of natural emotions using self assessment manikins. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 381–385. IEEE.

Gunes, H. and Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136.

Haag, A., Goronzy, S., Schaich, P., and Williams, J. (2004). Emotion recognition using bio-sensors: First steps towards an automatic system. In *Tutorial and research workshop on affective dialogue systems*, pages 36–48. Springer.

Hall, J. E. (2010). *Guyton and Hall Textbook of Medical Physiology, 12th Edition*. Saunders, 12 edition.

Hari, R. and Kujala, M. V. (2009). Brain basis of human social interaction: from concepts to brain imaging. *Physiological reviews*, 89(2):453–479.

Harnad, S. and Scherzer, P. (2008). First, scale up to the robotic turing test, then worry about feeling. *Artificial Intelligence in Medicine*, 44(2):83–89.

Haxby, J. V., Hoffman, E. A., and Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6):223–233.

Hegel, F., Spexard, T., Wrede, B., Horstmann, G., and Vogt, T. (2006). Playing a different imitation game: Interaction with an empathic android robot. In *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, pages 56–61. IEEE.

Heraz, A. and Frasson, C. (2007). Predicting the three major dimensions of the learner's emotions from brainwaves. *International Journal of Computer Science*, 2(3):187–193.

Heskes, T. (2001). Self-organizing maps, vector quantization, and mixture modeling. *IEEE Transactions on Neural Networks*, 12(6):1299–1305.

Holstege, G. (2016). How the emotional motor system controls the pelvic organs. *Sexual medicine reviews*, 4(4):303–328.

Horii, T., Nagai, Y., and Asada, M. (2016). Imitation of human expressions based on emotion estimation by mental simulation. *Paladyn, Journal of Behavioral Robotics*, 7(1).

Hudlicka, E. (2004). Two sides of appraisal: Implementing appraisal and its consequences within a cognitive architecture. In *Proceedings of the AAAI Spring Symposium 2004, Architectures for Modeling Emotion*.

Hudlicka, E. (2008). Modeling the mechanisms of emotion effects on cognition. In *AAAI Fall Symposium: Biologically inspired cognitive architectures*, pages 82–86.

Hudlicka, E. (2011). Guidelines for designing computational models of emotions. *Int. J. Synth. Emot.*, 2(1):26–79.

Iacoboni, M. (2005). Neural mechanisms of imitation. *Current opinion in neurobiology*, 15(6):632–637.

Iacoboni, M. (2009). Neurobiology of imitation. *Current opinion in neurobiology*, 19(6):661–665.

Jack, R. E., Garrod, O. G., and Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2):187–192.

James, W. (1884). What is an emotion? *Mind*, pages 188–205.

Janak, P. H. and Tye, K. M. (2015). From circuits to behaviour in the amygdala. *Nature*, 517(7534):284.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain sciences*, 17(2):187–202.

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193.

Jordan, M. I. (1998). *Learning in graphical models*, volume 89. Springer Science & Business Media.

Karnath, H.-O. (2001). New insights into the functions of the superior temporal cortex. *Nature reviews. Neuroscience*, 2(8):568.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current opinion in neurobiology*, 9(6):718–727.

Kilner, J. M., Friston, K. J., and Frith, C. D. (2007). The mirror-neuron system: a bayesian perspective. *Neuroreport*, 18(6):619–623.

Kim, Y., Lee, H., and Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3687–3691. IEEE.

Knill, D., Kersten, D., and Yuille, A. (1996). *Introduction: A Bayesian formulation of visual perception*, pages 1–21. Cambridge University Press.

Knill, D. and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719.

Koch, C. (1999). *Biophysics of Computation - Information Processing in Single Neurons*. Oxford University Press, New York.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press, Cambridge, MA.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. (2017). AFEW for valence and arousal estimation in-the-wild. *Image and Vision Computing*, pages –.

Kozuki, T., Toshinori, H., Shirai, T., Nakashima, S., Asano, Y., Kakiuchi, Y., Okada, K., and Inaba, M. (2016). Skeletal structure with artificial perspiration for cooling by latent heat for musculoskeletal humanoid kengoro. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 2135–2140. IEEE.

Kuppens, P., Oravecz, Z., and Tuerlinckx, F. (2010). Feelings change: accounting for individual differences in the temporal dynamics of affect. *Journal of personality and social psychology*, 99(6):1042.

Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R., and Nummenmaa, L. (2012). Naturalistic fmri mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Frontiers in human neuroscience*, 6.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Lang, P. J., Greenwald, M. K., Bradley, M. M., and Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Lawrence, I. and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of machine learning research*, 6(Nov):1783–1816.

Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16(329-336):3.

Lawrence, N. D. (2007). Learning for larger datasets with the gaussian process latent variable model. In *Artificial Intelligence and Statistics*, pages 243–250.

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press on Demand.

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

Levenson, R. W. (1988). Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. In *Social Psychophysiology and Emotion: Theory and Clinical Applications*. John Wiley & Sons.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Lim, A. and Okuno, H. G. (2014). The mei robot: towards using motherese to develop multimodal emotional intelligence. *IEEE Transactions on Autonomous Mental Development*, 6(2):126–138.

Lim, A. and Okuno, H. G. (2015). A recipe for empathy. *International Journal of Social Robotics*, 7(1):35–49.

Lopes, M. and Santos-Victor, J. (2005). Visual learning by imitation with motor representations. *IEEE Trans. on Sys., Man, and Cybernetics, Part B: Cybernetics*, 35(3):438–449.

MacKay, D. (2004). *Information Theory, inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK.

MacLean, P. D. (1949). Psychosomatic disease and the" visceral brain": Recent developments bearing on the papez theory of emotion. *Psychosomatic medicine*, 11(6):338–353.

Mahnke, R., Kaupuzs, J., and Lubashevsky, I. (2009). *Physics of stochastic processes: how randomness acts in time*. John Wiley & Sons.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, New York.

Marsella, S. C. and Gratch, J. (2009). Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90.

Masi, I., Tran, A. T., Hassner, T., Leksut, J. T., and Medioni, G. (2016). Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, pages 579–596. Springer.

Mattos, C. L. C., Dai, Z., Damianou, A., Barreto, G. A., and Lawrence, N. D. (2017). Deep recurrent gaussian processes for outlier-robust system identification. *Journal of Process Control*, 60:82–94.

McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., and Graesser, A. (2007). Facial features for affective state detection in learning environments. In *Proceedings of the Cognitive Science Society*.

Mehrabian, A. (1980). *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies*. Oelgeschlager, Gunn & Hain, Incorporated.

Mehrabian, A. and Russell, J. A. (1974). *An approach to environmental psychology.* the MIT Press.

Metallinou, A. and Narayanan, S. (2013). Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE.

Metta, G., Sandini, G., Natale, L., Craighero, L., and Fadiga, L. (2006). Understanding mirror neurons: a bio-robotic approach. *Interaction studies*, 7(2):197–232.

Meuleman, B. and Scherer, K. R. (2013). Nonlinear appraisal modeling: an application of machine learning to the study of emotion production. *IEEE Transactions on Affective Computing*, 4(4):398–411.

Milner, A. D. and Goodale, M. A. (1993). Visual pathways to perception and action. *Progress in brain research*, 95:317–337.

Minka, T. (1999). From hidden markov models to linear dynamical systems. Technical report, Technical report, MIT.

Molenberghs, P., Brander, C., Mattingley, J. B., and Cunnington, R. (2010). The role of the superior temporal sulcus and the mirror neuron system in imitation. *Human brain mapping*, 31(9):1316–1326.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press, Cambridge, MA.

Murray, R. M., Li, Z., Sastry, S. S., and Sastry, S. S. (1994). *A mathematical introduction to robotic manipulation*. CRC press.

Nakasone, A., Prendinger, H., and Ishizuka, M. (2005). Emotion Recognition from Electromyography and Skin Conductance. *The 5th International Workshop on Biosignal Interpretation*, pages 219–222.

Natale, L., Nori, F., Metta, G., Fumagalli, M., Ivaldi, S., Pattacini, U., Randazzo, M., Schmitz, A., and Sandini, G. (2013). The icub platform: a tool for studying intrinsically motivated learning. In *Intrinsically motivated learning in natural and artificial systems*, pages 433–458. Springer.

Nauta, W. J. (1971). The problem of the frontal lobe: A reinterpretation. *Journal of psychiatric research*, 8(3):167–187.

Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105.

Nishitani, N. and Hari, R. (2002). Viewing lip forms: cortical dynamics. *Neuron*, 36(6):1211–1220.

Oh, J.-H., Hanson, D., Kim, W.-S., Han, Y., Kim, J.-Y., and Park, I.-W. (2006). Design of android type humanoid robot albert hubo. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 1428–1433. IEEE.

Öngür, D., Ferry, A. T., and Price, J. L. (2003). Architectonic subdivision of the human orbital and medial prefrontal cortex. *Journal of Comparative Neurology*, 460(3):425–449.

Oravecz, Z., Tuerlinckx, F., and Vandekerckhove, J. (2011). A hierarchical latent stochastic differential equation model for affective dynamics. *Psychological methods*, 16(4):468.

Orozco, J., Rudovic, O., GonzÃlez, J., and Pantic, M. (2013). Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelids and irises. *Image and Vision Computing*, 31(4):322 – 340.

Ortony, A., Clore, G., and Collins, A. (1988). The cognitive structure of emotions. *CBO9780511571299*.

Ortony, A., Clore, G. L., and Collins, A. (1990). *The cognitive structure of emotions*. Cambridge university press.

Oztop, E., Kawato, M., and Arbib, M. (2006). Mirror neurons and imitation: A computationally guided review. *Neural Networks*, 19(3):254–271.

Oztop, E., Kawato, M., and Arbib, M. A. (2013). Mirror neurons: functions, mechanisms and models. *Neuroscience letters*, 540:43–55.

Palminteri, S., Wyart, V., and Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6):425–433.

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, 9(2):148–158.

Pessoa, L. and Adolphs, R. (2010). Emotion processing and the amygdala: from a 'low road' to'many roads' of evaluating biological significance. *Nature Reviews Neuroscience*, 11(11):773–783.

Picard, R. W. (1997). *Affective computing*, volume 252. MIT press Cambridge.

Picard, R. W., Vyzas, E., and Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191.

Plutchik, R. (1956). The psychophysiology of skin temperature: A critical review. *Journal of General Psychology*, 55(June):249–268.

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association.

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Pribram, K. H. (1970). Feelings as monitors. In *Feelings and emotions. The Loyola Symposium*, pages 441–453.

Pylyshyn, Z. (1984). *Computation and Cognition*. MIT Press, Cambridge, MA.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*. The MIT Press, Cambridge, MA, USA.

Reisenzein, R., Hudlicka, E., Dastani, M., Gratch, J., Hindriks, K., Lorini, E., and Meyer, J.-J. C. (2013). Computational modeling of emotion: Toward improving the inter-and intradisciplinary exchange. *IEEE Transactions on Affective Computing*, 4(3):246–266.

Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., Lalanne, D., and Schuller, B. (2015a). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30.

Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., and Pantic, M. (2015b). Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 3–8. ACM.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE.

Rizzolatti, G. and Sinigaglia, C. (2016). The mirror mechanism: a basic principle of brain function. *Nature Reviews Neuroscience*, 17(12):757–765.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. *Review of personality & social psychology*.

Roseman, I. J. (2001). A model of appraisal in the emotion system. *Appraisal processes in emotion: Theory, methods, research*, pages 68–91.

Rudrauf, D., David, O., Lachaux, J.-P., Kovach, C. K., Martinerie, J., Renault, B., and Damasio, A. (2008). Rapid interactions between the ventral visual stream and emotion-related structures rely on a two-pathway architecture. *Journal of Neuroscience*, 28(11):2793–2803.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Said, C. P., Moore, C. D., Engell, A. D., Todorov, A., and Haxby, J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of vision*, 10(5):11–11.

Salzman, C. D. and Fusi, S. (2010). Emotion, cognition, and mental state representation in amygdala and prefrontal cortex. *Annual review of neuroscience*, 33:173–202.

Sanborn, A. N. and Chater, N. (2016). Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893.

Sanborn, A. N. and Chater, N. (2017). The sampling brain. *Trends in Cognitive Sciences*, 21(7):492–493.

Saper, C. B. (2002). The central autonomic nervous system: conscious visceral perception and autonomic pattern generation. *Annual review of neuroscience*, 25(1):433–469.

Sariyanidi, E., Gunes, H., and Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24.

Schachter, S. and Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5):379.

Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion*, 7(3-4):325–355.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, 92(120):57.

Selvaraj, N., Jaryal, a., Santhosh, J., Deepak, K. K., and Anand, S. (2008). Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography. *Journal of medical engineering & technology*, 32(6):479–484.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, 17(11):565–573.

Sloman, A. (1992). Prolegomena to a theory of communication and affect. *Communication from an Artificial Intelligence Perspective*, pages 229–260.

Sporns, O., Chialvo, D. R., Kaiser, M., and Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in cognitive sciences*, 8(9):418–425.

Sprevak, M. (2016). Philosophy of the psychological and cognitive sciences. In *The Oxford Handbook of Philosophy of Science*, page 92. Oxford University Press.

Stuart, A. M. (2010). Inverse problems: a bayesian perspective. *Acta Numerica*, 19:451–559.

Tani, J., Ito, M., and Sugita, Y. (2004). Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using rnnpb. *Neural Networks*, 17(8):1273–1289.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Titsias, M. and Lawrence, N. (2010). Bayesian Gaussian Process Latent Variable Model. *Artif. Intell.*, 9:844–851.

Tomkins, S. (1962). *Affect, imagery, consciousness*, volume 1. Springer, New York.

Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528.

Tramacere, A. and Ferrari, P. F. (2016). Faces in the mirror, from the neuroscience of mimicry to the emergence of mentalizing. *Journal of Anthropological Sciences*, 94:113–126.

Trovato, G., Zecca, M., Kishi, T., Endo, N., Hashimoto, K., and Takanishi, A. (2013). Generation of humanoid robot's facial expressions for context-aware communication. *International Journal of Humanoid Robotics*, 10(01):1350013.

Turkle, S. (1984). *The Second Self: Computers and the Human Spirit*. Simon & Schuster, Inc., New York, NY, USA.

Uhlenbeck, G. E. and Ornstein, L. S. (1930a). On the theory of the brownian motion. *Physical review*, 36(5):823.

Uhlenbeck, G. E. and Ornstein, L. S. (1930b). On the theory of the brownian motion. *Phys. Rev.*, 36:823–841.

Ungerleider, L. G. and Desimone, R. (1986a). Cortical connections of visual area mt in the macaque. *Journal of Comparative Neurology*, 248(2):190–222.

Ungerleider, L. G. and Desimone, R. (1986b). Projections to the superior temporal sulcus from the central and peripheral field representations of v1 and v2. *Journal of Comparative Neurology*, 248(2):147–163.

Van Kampen, N. (2011). *Stochastic Processes in Physics and Chemistry*. Elsevier.

Van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry*, volume 1. Elsevier.

Venkatraman, A., Edlow, B. L., and Immordino-Yang, M. H. (2017). The brainstem in emotion: A review. *Frontiers in neuroanatomy*, 11.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., and Schroeder, M. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE J AC*, 3(1):69–87.

Vitale, J., Williams, M.-A., Johnston, B., and Boccignone, G. (2014). Affective facial expression processing via simulation: A probabilistic model. *Biologically Inspired Cognitive Architectures Journal*, 10:30–41.

von Helmholtz, H. (1858). Über integrale der hydrodynamischen gleichungen welche den wirbelbewegungen entsprechen. *Crelles, J.*, 55:25–55.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*, volume 9. Voss.

Wang, B. and Titterington, D. (2004). Variational bayesian inference for partially observed diffusions. Technical report, Technical Report 04-4, University of Glasgow. http://www. stats. gla. ac. uk/Research/-TechRep2003/04-4. pdf.

Wang, J., Hertzmann, A., and Fleet, D. J. (2006). Gaussian process dynamical models. In *Advances in neural information processing systems*, pages 1441–1448.

Wang, J. M., Fleet, D. J., and Hertzmann, A. (2008). Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298.

Wiese, E., Metta, G., and Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, 8:1663.

Winsberg, E. (1999). Sanctioning models: The epistemology of simulation. *Science in context*, 12(2):275–292.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 358(1431):593–602.

Wood, A., Rychlowska, M., Korb, S., and Niedenthal, P. (2016). Fashioning the face: sensorimotor simulation contributes to facial expression recognition. *Trends in cognitive sciences*, 20(3):227–240.

Yang, Y., Saleemi, I., and Shah, M. (2013). Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1635–1648.

Zafeiriou, S., Papaioannou, A., Kotsia, I., Nicolaou, M., and Zhao, G. (2016). Facial affect"in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–47.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Proc. of IEEE CVPR*, pages 2879–2886.

Ziemke, T. and Lowe, R. (2009). On the role of emotion in embodied cognitive architectures: From organisms to robots. *Cognitive computation*, 1(1):104–117.