UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

DIPARTIMENTO DI INFORMATICA 'GIOVANNI DEGLI ANTONI'
CORSO DI DOTTORATO IN INFORMATICA
XXXVI CICLO

TESI DI DOTTORATO DI RICERCA

ON THE PROBABILISTIC MODELLING OF PAIN

SSD 01/B1

Autrice
Sabrina Patania

Tutor
Prof. Giuseppe Boccignone

Coordinatore del Dottorato
Prof. Roberto Sassi

A.A. 2022-2023

“Ex malo bonum”

— Latin saying

Abstract

Pain is a complex subjective experience encompassing sensory, affective, and cognitive dimensions. Evidence challenges the linear relationship between nociceptive activation and pain perception, revealing scenarios where pain is felt without nociceptive input or vice versa. This research is based on the understanding that pain arises not merely from bottom-up stimuli but also from a blend of subjective components and contextual evaluations. The aim of this study is to develop a probabilistic and computational model of pain that aligns with both behavioural data and neurobiological constraints, aspiring to represent pain at various levels. This includes incorporating perceptual, affective, motivational, and social aspects into a comprehensive framework, and proposing implementation models tailored to specific contexts and varying levels of abstraction. The model uses Bayesian approaches to capture the probabilistic and dynamic nature of the phenomenon. The upcoming discussions will adopt a multidisciplinary lens, converging philosophical, psychological, and neurobiological insights to shape the envisaged model.

Contents

1 Prelude: The Conundrum of Pain **3**
1.1 The signs of Pain . . . 3
1.1.1 Self-reports: gold standard or fool's gold? . . . 4
1.1.2 Behavioural response . . . 6
1.1.3 Physiological responses . . . 10
1.2 Philosophical debate . . . 10
1.3 Dennett's criticism: exploring the computability of pain experience . . . 13

2 Most influential theories **15**
2.1 The first systematic study: René Descartes . . . 15
2.2 Specificity theory . . . 16
2.3 Pattern theory . . . 17
2.4 The two faces of pain . . . 17
2.5 Gate control theory . . . 18
2.6 Neuromatrix . . . 19
2.7 Motivation-decision theory . . . 20
2.8 Hierarchical processing of pain . . . 21
2.9 Cognitive-behavioural theory of pain . . . 22

3 Neurobiological Groundings: Shaping Model's Constraints **25**
3.1 Significant neural components . . . 26
3.2 Pain descending modulation . . . 29
3.3 A critical review of Pain Matrix . . . 30
3.4 The role of salience network . . . 32
3.5 From Reactive to Prospective Regulation: The Bayesian Brain Hypothesis 33
3.5.1 Three-Level Neural Hierarchy and Allostatic Control . . . 33
3.5.2 Allostatic Modulation in Pain Perception . . . 34

4 The Computational Modelling of Pain: An Overview and a Roadmap **35**
4.1 Multiple dimensions of pain, many models . . . 36
4.2 Computational modelling: A conceptual overview of the literature . . . 37

4.3 Epistemological interlude: levels of explanation . . . 41
4.4 The Bayesian roadmap . . . 42
4.5 Back to the literature: a principled taxonomy for statistical models of pain 48
4.5.1 The Class 0 model baseline . . . 49
4.5.2 Accounting for temporal change: Class 1 models . . . 52
4.5.3 Putting action into action: Class 2 models . . . 52
4.5.4 Class 3 models: the unexplored challenge of modelling social pain 54
4.6 Discussion . . . 56

5 The gentle start: A case for Class 1 models 57
5.1 Dataset . . . 57
5.2 A discriminative implementation model of a Class 1 model . . . 59
5.2.1 Graph architecture definition . . . 60
5.2.2 Node-level features representation . . . 61
5.2.3 Graph processing . . . 66
5.2.4 Simulation . . . 66
5.3 A generative implementation model of the Class 1 model . . . 68
5.3.1 Implementation model . . . 69
5.3.2 Clustering . . . 74
5.4 Implementation details . . . 77
5.5 Discussion . . . 81

6 Interlude: What is an emotion? 85
6.1 Emotion perspectives and theories . . . 85
6.2 Fear and the theory of constructed emotion: a brief tour . . . 88

7 Affective Alarms: Unpacking Fear Generalisation in Pain Contexts 95
7.1 How fear spreads: the mechanisms of generalisation . . . 96
7.2 Adaptive vs Maladaptive Fear Generalisation . . . 97
7.3 Description of the experimental setting . . . 99
7.3.1 Experiment description . . . 99
7.3.2 Data recording . . . 102
7.4 Bridging pain and fear . . . 103
7.5 A simple Class 2 model of fear generalisation . . . 105
7.5.1 Simulation and results . . . 107
7.6 The active inference model: a simulative approach . . . 109
7.7 Pain and fear in the eyes: theoretical and empirical consequences of the exploitation/exploration dilemma . . . 112
7.7.1 Foraging signatures in eye movements: a premise . . . 117
7.7.2 Preliminary analysis at the macroscopic level . . . 122
7.8 Discussion . . . 129

8 Beyond the Individual: The Social Resonance of Pain 131
8.1 Biopsychosocial model of pain . . . 132
8.2 Social communication model of pain . . . 133
8.3 Dyadic interaction: a transactional perspective . . . 135
8.4 Modelling patient-clinician dynamics . . . 137

8.4.1 Active inference POMDP model . . . 137
8.4.2 Simulations . . . 143
8.5 Pain communication and understanding as a pragmatic problem . . . 150
8.5.1 The RSA framework . . . 150
8.5.2 RSA-based patient-clinician interaction . . . 152
8.5.3 Implementation . . . 156
8.5.4 Simulations . . . 162
8.6 Discussion . . . 174

9 Conclusions **177**

A Probabilistic Graphical Models **181**
A.1 Forney-style Factor Graph . . . 186

B Active Inference: the bare essentials **191**
B.1 Generative models . . . 192
B.2 Variational Free-energy . . . 195
B.3 Perception and State estimation . . . 196
B.4 Expected Free-energy . . . 197
B.5 Planning and Decision Making: Updating plan distribution . . . 198

C Neurobiology of Fear Generalisation **201**
C.0.1 Neural substrates of positive generalisation . . . 204
C.0.2 Neural substrates of negative generalisation . . . 205
C.0.3 Animal evidence . . . 206

D A Brevia on Pragmatics: How Use Contributes to Meaning **207**
D.0.1 Austin: speech acts . . . 208
D.0.2 Grice: the inferential stance . . . 211
D.0.3 Sperber and Wilson: Relevance Theory . . . 212
D.0.4 At the origins of the communication act . . . 215
D.0.5 State of the art in computational pragmatics . . . 219

Bibliography **223**

Introduction

Pain, often viewed as a straightforward sensory response, is a multifaceted subjective experience intricately woven with sensory, affective, and cognitive threads. Delving into the intricacies of pain perception reveals an intriguing observation: the sensory component of pain does not consistently correlate with the experience itself. There are instances where nociceptive triggers fail to invoke pain, and conversely, pain sensations can emerge devoid of any clear nociceptive stimuli. An insightful reflection on this is evident in psychiatric contexts. For instance, individuals with schizophrenia frequently display a diminished sensitivity to pain (Rosenthal et al., 1990; Dworkin, 1994; Singh et al., 2006; Fishbain, 1982). On the flip side, chronic pain, a condition where pain lingers beyond the typical healing period, is often entwined with a host of emotional and cognitive challenges, including anxiety, depression, and broader mental health disorders (Fishbain et al., 1997; Asmundson and Katz, 2009; Bushnell et al., 2013). Such empirical observations underscore a deeper involvement of cognitive and emotional dimensions in the pain experience, complicating the previously assumed linear bond between nociceptive stimuli and pain perception.

Guided by these insights, this research is anchored in the belief that pain is not merely an outcome of stimulus-based, bottom-up mechanisms. Instead, it represents a compound phenomenon interlaced with an array of subjective elements ranging from memories and motivation to cognitive evaluations and emotional states. Contextual cues further augment this intricate tapestry.

Stepping into the realm of computational modelling, the overarching ambition of this research is to sketch a unified framework that can echo the nuances of both acute and chronic pain experiences. A pivotal consideration is an acknowledgement that pain can manifest without a direct nociceptive prompt, and the model should be receptive to the myriad physiological and psychological intricacies entwined with chronic pain manifestations. Venturing further, there is potential to explore realms of psychic pain or deeper emotional suffering.

A survey of the current landscape reveals a scarcity of computational models adept at capturing the nuances of pain, especially when it strays from the normative path of acute pain. Models often falter in encapsulating phenomena like chronic pain, phantom limb sensations, or neuropathic pain episodes, where pain perception is heightened

without a clear nociceptive trigger. Argüello et al. (2015) discern that this gap stems from the deterministic core of prevailing models, which clashes with the inherently probabilistic fabric of perception phenomena. To bridge this gap, the proposed model champions the inclusion of stochastic elements, echoing the unpredictability and dynamism of neural activities.

In the journey ahead, a comprehensive, multidisciplinary approach will be embraced. Delving deep into pain from varied lenses —philosophical, psychological, and neurobiological— will pave the way. These multifaceted insights will then be synergistically fused, informing and shaping the contours of the proposed computational model. As this thesis unfolds, readers should note the layered approach to our exploration. We aim to understand pain from multiple angles, integrating various levels of generalisation into our analysis. Starting with individual perceptual aspects, we will move to broader affective dimensions and then consider the social contexts of pain experiences. This progressive addition ensures a holistic understanding, allowing us to capture the richness and complexity of pain in its entirety.

To achieve these objectives, this thesis proposes a model pluralism approach, acknowledging the complexity of pain and the impracticality of a single model encompassing all facets. Starting with initial, straightforward proposals where the subject's inference from a physical cause is the sole focus of analysis, we progress to more sophisticated models. Utilising a Bayesian framework, particularly active inference (Friston et al., 2017), we aim to model components related to an agent acting within an environment. This approach begins with simple environmental contexts where action-perception loops are the main focus and extends to complex social contexts through dyadic modelling, incorporating communication dynamics to adhere to the biopsychosocial model of pain (Hadjistavropoulos et al., 2011). Moreover, the communication dynamics have also been examined through the lens of the pragmatics framework (see Appendix D). This approach allows for adaptable and context-specific models that can effectively represent the multifaceted nature of pain, from individual experiences to social interactions.

CHAPTER 1

Prelude: The Conundrum of Pain

1.1 The signs of Pain

> The action is initiated by an "incongruity" between the state of the organism and the state that is being tested for, and the action persist until the incongruity [...] is removed.
>
> — Miller et al. (1960)

From an evolutionary perspective, pain is, first of all, a call to action. It points to the need for an intervention to restore homeostatic balance. The generating aversive state requires a proper behavioural response to cope with pain by various means: communicating the condition to request assistance, finding an escape strategy, avoiding certain actions or fighting the threat. Besides, to build a computational model of pain, clarifying which measurable pain indicators are available and reliable to assess pain severity is a crucial issue to solve.

We can identify various measurable pain indicators that support pain assessment, some traditionally used in clinical settings and others specifically in computational models. In clinical practice, self-reports are the most commonly used tool, providing direct subjective accounts of pain from patients. However, computational models often rely on observational measures that are less subject to voluntary control and exhibit more automatic responses, such as facial expressions and physiological measures. Available measurements allow for the formulation of an external rating, which is crucial for clinicians or caregivers assisting a person in pain. From a computational perspective, these ratings are often used to label data for model training, ensuring that the system can accurately interpret and predict pain levels based on observable cues.

1.1.1 Self-reports: gold standard or fool's gold?

McCaffery (1968) pioneered the now widely accepted notion that "Pain is what the person experiencing it says it is, existing whenever they declare it does". This perspective shifted the focus to individuals, positioning them as the authoritative sources on their own pain experiences. This approach to pain assessment is not only equitable, aligning with the tenets of patient advocacy and ethical clinical practice, but it also established the groundwork for what has come to be regarded as the "gold-standard" in pain assessment: self-reporting. In more recent times, the classification of pain as the "fifth vital sign" (Society, 1999) has further entrenched the use of self-reporting in clinical settings.

In adults, the intensity of pain is commonly gauged using unidimensional scales. These may be numerical, ranging from 0, indicating "no pain", to 10, denoting "the worst imaginable pain", or employ a visual analogue scale—a 10 cm line with endpoints labelled "no pain" and "pain as bad as it could be". Alternatively, descriptive, categorical scales are used, which might range from "no pain" to "extremely intense pain" (Jensen and Karoly, 2011). For children, pain assessment often involves graphic facial scales that depict varying degrees of pain expression (Chambers et al., 2005). These methods, among others, aimed at quantifying pain through self-report, are often hailed as the gold standard in pain assessment. They prioritise the individual's subjective experience, underscoring the centrality of the patient's perspective, and are pivotal in clinical settings. When applied correctly, these tools can yield valid and reliable insights. Nonetheless, despite its esteemed status, the self-report approach has shortcomings.

The simplistic reliance on self-reported pain scales, whether numerical or descriptive, can obscure the nuanced and multidimensional nature of pain, which includes not only physical sensations but also emotional and psychological distress. This is especially problematic in populations such as children or cognitively impaired individuals, who may lack the necessary communication skills to accurately convey their pain. Moreover, factors like cultural background, personal pain tolerance, and the desire to avoid medication side effects can lead patients to under-report or even exaggerate their pain, further complicating the clinician's task of accurate pain assessment.

In pain assessment, a fundamental assumption often exists that patients aim to minimise their pain while clinicians strive to alleviate it. This assumption, termed the Assumption of Mutuality (AoM), predicates that patients truthfully communicate their pain and that clinicians respond with both attentiveness and skill (Schiavenato and Craig, 2010). For this dynamic to hold, it is presumed that patients can and will openly express their pain and that clinicians will accept these reports without skepticism. However, this ideal scenario often diverges from the complexities observed in clinical settings. Patients may hesitate to report their pain, perhaps assuming that clinicians can perceive their discomfort without explicit communication. Conversely, clinicians might expect patients to report pain whenever it is significant, regardless of the patient's reluctance or fear of being perceived as complainers (Watt-Watson et al., 2001).

These misalignments are compounded by various factors affecting both parties. Patients might modify their reports based on the perceived benefits or drawbacks in specific contexts—sometimes downplaying pain to avoid undesirable treatments or to

appear strong, or exaggerating it to receive more attention or medication. Similarly, clinicians bring their own biases, experiences, and professional judgements into the assessment process, which can skew their interpretation of patient reports. This complex interplay often obstructs clear communication and effective pain management, highlighting the need for a more nuanced approach that recognises and addresses these inherent challenges in pain assessment.

When assessing pain, clinicians often place more weight on behavioural cues than on verbal reports from their patients (Melzack, 1983). For instance, a patient displaying signs of distress through grimacing is more likely to receive pain medication than one who is smiling (McCaffery et al., 2000). Consequently, the notion that self-report is paramount in pain assessment is somewhat illusory. The complexity behind this issue stems from a variety of personal factors that may bias a clinician's response to a patient's self-report. Influences such as the patient's age, gender, ethnicity, socioeconomic background, lifestyle, legal issues, attractiveness, and demeanor have all been noted as potential factors that could sway clinical judgments in pain management (Hadjistavropoulos et al., 1990; Birdwell et al., 1993; Calvillo and Flaskerud, 1993; Tait and Chibnall, 1997).

All in all, self-report, unlike traditional vital signs that provide discrete physiological data crucial for sustaining life, represents a more complex and subjective measure. It is not an objective indicator of a physiological state but a symptomatic declaration that necessitates interpretation alongside other clinical data. This underscores the complexity of pain assessment, which cannot be simply quantified as an absolute measure but is better understood as a dynamic, transactional process.

Pain embodies a multifaceted experience that includes cognitive, emotional, and sensory components. It demands an assessment approach that integrates biophysiological and sociocultural factors, as argued by biopsychosocial models of pain (Asmundson and Wright, 2004). Such models reveal that pain comprises both interpersonal and intrapersonal dimensions, influencing each other in a systemic manner. This complexity often leads to the underestimation of pain in clinical settings, where psychological and social factors play significant roles.

Building on this understanding, we conceptualise pain assessment as a transaction between patient and clinician, involving several key elements (see Fig. 1.1). First, "Contributing factors" establish the personal and societal backdrop influencing both parties. Next, the "Assessment process" forms a cyclic feedback loop starting from the pain stimulus. This process extends from the patient's personal experience to the clinician's evaluation and actions, which in turn affect the patient's condition. Lastly, "Intervening steps" represent various potential actions within the assessment process, depicted as a continuum in our model, highlighting the interactive nature of pain assessment (Schiavenato and Craig, 2010).

This model posits that the patient-clinician relationship is inherently purposeful and goal-oriented, rooted in a mutual desire for pain relief and effective treatment. Trust is a fundamental component of this relationship, facilitating the meaningful exchange of information about pain. These interactions form the core of the transactional model of pain assessment, which will be further explored in the proposed social model in Chapter 8.

Much of the discussion thus far applies primarily within the context of clinician-

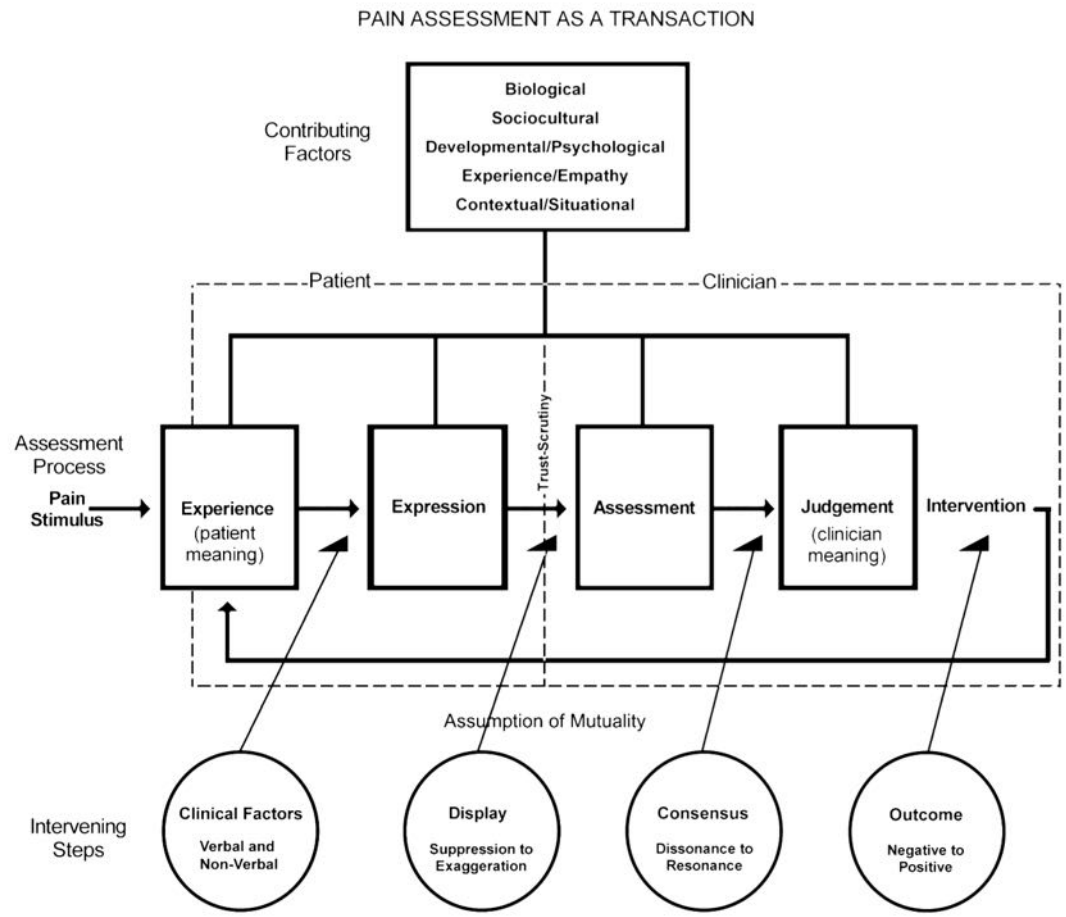


Figure 1.1: *Pain assessment as a transaction. From Schiavenato and Craig (2010).*

patient interactions. However, when considering other settings, such as experimental environments where healthy participants are subjected to painful stimuli and asked to provide a rating (essentially a form of self-report), the dynamics can differ. Despite these differences, any setting that involves interaction inherently engages social and communicative factors that play significant roles. Whether in clinical or experimental contexts, the impact of interpersonal dynamics and the complexities of communication influence how pain is expressed, perceived, and managed.

1.1.2 Behavioural response

If self-reports are not as immediately or fully reliable as one might initially believe, there are nevertheless responses that, although still subject to influence, are less frequently under deliberate control by higher cognitive functions. Indeed, behavioural responses such as facial expressions, changes in vocal tone, or even cries and groans, along with specific movements, can provide insights into an individual's pain state. These involuntary reactions are often automatic and driven by the immediate physiological impacts of pain, making them potentially more direct indicators of the sufferer's condition.

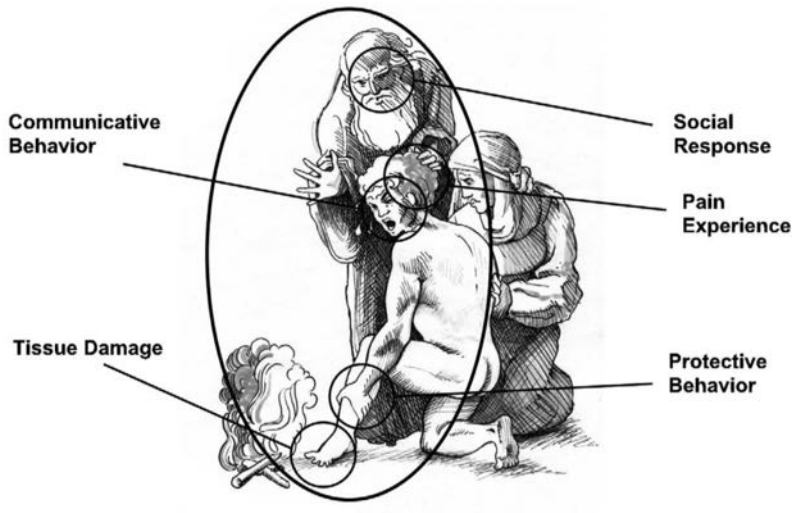


Figure 1.2: *Key aspects of the pain experience. From Sullivan (2008).*

What distinguishes persons with and without pain is not just how they feel but how they behave. The adaptive value of pain strongly depends on the behavioural response: the alert sensory system of pain only has a meaning if capable of triggering the actuator system for pain handling. Despite the robust relation between pain and behaviour, this is not considered as a defining dimension of the pain system according to IASP definition of pain. Sullivan (2008) has proposed to include this dimension in pain discussion and, with this aim, has drawn up a biopsychomotor model of pain. Starting from the naive sensory model of pain proposed by Descartes, Figure 1.2 shows prominent features of pain experience. These include communicative pain behaviours, protective pain behaviours, and social response behaviours (Sullivan, 2008).

Protective behaviours comprises functionally separable subgroups of behaviours, from reflexive withdrawal to complex avoidance mechanisms involving pain anticipation strategies (see Chapter 7), brought together by the function of protecting from a potential or actual damage and promoting recovery after an injury (Sullivan, 2008).

Communication of pain is another way to protect our safety by resorting to assistance or care from others. The communicative dimension of pain has been widely established in research (Prkachin and Craig, 1995; Hadjistavropoulos and Craig, 2002; Craig, 2015). A part from linguistic report, the most common media used to communicate pain are facial expressions and vocalisations (Hadjistavropoulos and Craig, 2002). The configuration of the face rapidly communicates to perceivers information about the internal state of the person experiencing pain (Williams, 2002). As evidence suggests, we are fairly able to discern characteristic facial patterns and make inference on the presence of pain (Simon et al., 2008). Despite some evidence shows the innate nature of some specific facial patterns related to pain (Williams, 2002; Craig and Patrick, 1985), facial expressions can be overridden to some degree by conscious effort (Hill and Craig, 2002). Individual differences and communication-orientated vocation of facial expressions makes them not fully consistent with the actual pain perception, falling into the same issues affecting verbal reports. Even if a certain level of automaticity is always present (Hadjistavropoulos and Craig, 2002), personal propensities,

goals and plans may influence the overt behaviour depending on the context (Sullivan et al., 2006). For instance, exaggeration in a favourable environment could be preferred in order to receive maximum care, whereas withholding in a hostile environment could preserve reputation. In the vein of Duck and McMahan (2010), as with self-reports and any other form of overt communication used for pain assessment, the intentional component of facial display should be interpret as *communication as transaction* instead of as *communication as action* (Schiavenato and Craig, 2010).

For a more detailed investigation of social component we refer the reader to Chapter 8. Nevertheless, it is important to keep in mind that communicative and social dimensions are notably intertwined. In general, every behaviour, also protective, carries to some extent both communicative and social value, since insofar as it is perceived by others in a social environment.

Facial expressions

Among behavioural pain responses, facial expressions are the most investigated due to the well-established prominence of the face as a source of information compared to other channels of nonverbal communication such as paralinguistic vocalisation or involuntary and purposeful bodily activity. The natural ability of monitoring and interpreting facial expressions of other human beings, socially and phylogenetically determined, implies a certain measure of consistency across individuals. Psychological and neurobiological studies highlight the connection between pain and facial movements.

Experiments conducted by Prkachin and Solomon (2008), based on a sample of 129 subjects suffering from shoulder pain, identify several facial actions discriminating painful from non-painful movements with high validity and reliability. In this vein, Simon et al. (2008) proved the human capacity of discerning prototypical pain expression from other emotional and neutral facial reactions. These experiments argue for consistency of facial pain expression patterns by focusing on the spontaneous arising of specific facial reactions in a painful condition and the recognition of a painful experience of others by the expression indicator. Despite these findings, individuals have different pain thresholds (Prkachin and Craig, 1995) and expressiveness levels (Werner et al., 2017), making facial expression a weak stand-alone proxy of pain intensity level.

While reflexive components are certainly present in facial expressions, the communicative aspect also plays a crucial role, as also evidenced from an evolutionary standpoint. Drawing from evolutionary psychology, certain pain behaviours and the underlying propensities are likely to have been shaped by natural selection because they conferred survival advantages to individuals who effectively communicated their pain in the presence of potential helpers or known antagonists. Vigilance to observed pain cues co-evolved, as the expression of pain, when followed by helpful actions from observers, could significantly boost an individual's recovery and survival chances. This reciprocal relation is foundational to concepts like kin or reciprocal altruism, where the cost to helpers is low, yet the survival benefits to the pain sufferer are high (Williams, 2002).

Moreover, the development of these behaviours aligns with neo-Darwinism, which integrates Darwin's theory of selection pressures with genetic influences at the functional level rather than just the structural or behavioural levels. In this context, behaviours that enhance survival or reproductive success contribute to the likelihood of

genetic transmission to subsequent generations, shaping both physical and psychological adaptations to environmental challenges (Williams, 2002).

Specifically, behavioural routines that result from natural selection must confer advantages that outweigh their costs. This requires the behaviours to be both specific and detectable by others. The evolution of facial expressions of pain, like other non-verbal cues, likely evolved alongside language as a means of conveying crucial information within social groups. The theory of reciprocal altruism further suggests that such exchanges of aid are sustainable only if individuals remain vigilant to the possibility of social cheating - where one party may benefit without reciprocating or fulfilling their side of the social contract (Williams, 2002).

In sum, facial expressions of pain should not only be seen as involuntary or reflexive responses but as evolved, functional behaviours that enhance communication within social settings, potentially signalling the need for help. This ability to communicate pain effectively can be seen as a crucial adaptive mechanism, supporting not just individual survival but also facilitating social cohesion and mutual support within groups.

Other responses

Expanding on the variety of behavioural responses to pain, it is important to consider how these manifestations can offer critical clues about an individual's pain experience that go beyond verbal articulation. Body movements, for example, are often instinctual reactions to pain designed to protect an area of injury or to reduce further discomfort (Walsh et al., 2014). Such movements can range from reflexive withdrawal from a painful stimulus—a rapid jerk of a hand away from a hot surface—to more complex avoidance strategies such as limping to reduce weight-bearing on an injured leg. These protective responses are not just reactive but can also be anticipatory, as individuals learn to avoid movements or situations that previously resulted in pain.

Vocalisations represent another significant category of pain-related behavioural responses. These sounds are not merely expressions of discomfort; they can serve functional purposes by alerting others to the individual's need for help. Cries of pain can mobilise assistance from caregivers or signal to others the seriousness of an injury, potentially facilitating faster response times in emergencies. The nature and intensity of these vocalisations can vary dramatically depending on the severity of the pain and the individual's emotional state and personal pain tolerance (Thiam et al., 2016).

Moreover, changes in voice prosody, namely variations in the pitch, volume, and speed of speech, are also telling indicators. These alterations can subtly communicate pain intensity and emotional distress associated with pain, providing observers with cues that might not be as evident through facial expressions or self-report alone. For example, a person's speech may become slower and more monotone with increasing pain severity, reflecting their focus on coping with the discomfort (Tsai et al., 2017).

The context in which these responses occur also plays a pivotal role in how they are manifested and interpreted. Social norms and cultural background can influence how openly pain is expressed, with some cultures encouraging stoicism and others more open displays of suffering. Environmental factors, such as being in a public versus private setting, also modify how individuals express pain. In a clinical setting, the presence of healthcare professionals might lead individuals to amplify or downplay their pain expressions, influenced by the need to be taken seriously or fear of invasive

treatment options.

Collectively, these behavioural responses (body movements, vocalisations, and changes in voice prosody) form a complex repertoire through which pain is communicated. Understanding these responses in their contextual and cultural framework is crucial for accurate pain assessment and effective pain management, highlighting the need for a holistic approach that considers not just the physiological but also the psychosocial dimensions of pain.

1.1.3 Physiological responses

The autonomic nervous system (ANS) is a neural structure extensively involved in pain. In particular, a state of pain increases sympathetic activity, thus changes in physiological responses can be used as indicators of pain experience. Interestingly, physiological reactions cannot be reduced to proxy for intensity stimulation, but rather autonomic changes are dependent on the integrated activity of many centers in the central nervous system, including those responsible of affective-motivational modulation and cognitive reappraisal (see Figure 1.3) (Jänig, 2012).

More specifically, autonomic regulation leads to cardiovascular changes affecting heart rate (HR) and heart rate variability (HRV), increasing lower frequency power (Appelhans and Luecken, 2008). Moreover, pupil diameter changes in response to pain, since the sympathetic system controls the pupil dilation (Chapman et al., 1999). Nevertheless, skin conductance is shown to be the physiological signal more correlated to pain (Treister et al., 2012). Sweat glands, innervated by sympathetic excitatory efferent neurons, intensify their activity during pain, altering electrical properties of the skin (electrodermal activity, EDA) and hence increasing the electrical conductance (Storm, 2008). In this regard, a study by Treister et al. (2012) proves that a linear combination of different autonomic signals significantly differentiated between pain and no pain experiences as well as between three intensity levels of heat pain. Skin conductance level (SCL) and photoplethysmography amplitude (PPGA) were shown to be most sensitive, followed by the number of skin conductance fluctuations (NSCF) and HF-HRV.

However, physiological measures are not completely pain-specific, but they are mainly correlated with arousal, which is also present in other affective states and, as a consequence, it must be emphasised once again the role of the existing affective state in causing autonomic changes.

1.2 The ontology of pain: philosophical suggestions [2]

Pain is a slippery concept to pin down. The difficulty to define closely its nature is also a reflection of our way of referring to word "pain" to indicate apparently different experiences. We could say to feel pain as a consequence of injuries, disease, losses, stressful events in general or even if nothing seems to be affecting us. Either way, *it* hurts.

[2]The philosophical debate is far more complex than it may appear from these few lines. Positions can be schematically divided into sense-datum theories, direct and indirect perceptual theories, representational theories, and mixed theories, and different ways to tackle two classical problems: the problem of focus, which means pain reports are focused on the experience rather than on the tissue damage, and the problem of the affective dimension, namely the question about the representational content of the affective component of pain.

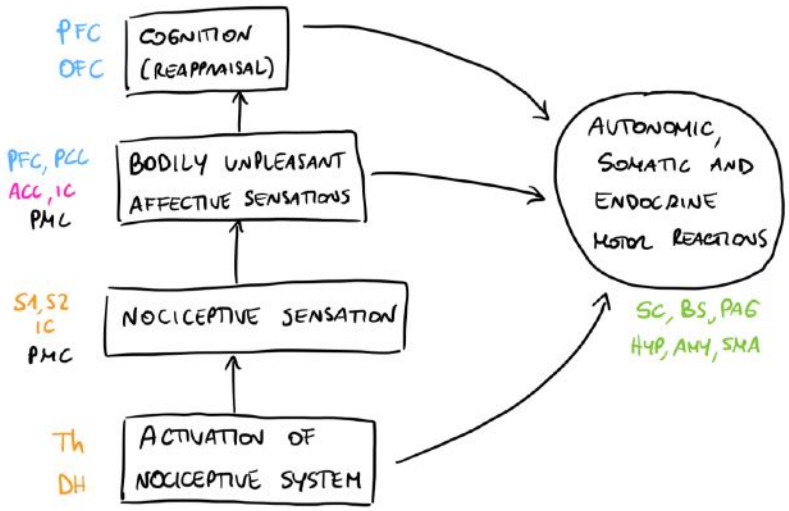


Figure 1.3: *Illustration of nervous system centers and their roles in autonomic nervous system reactions. Adapted from Jänig (2012).*

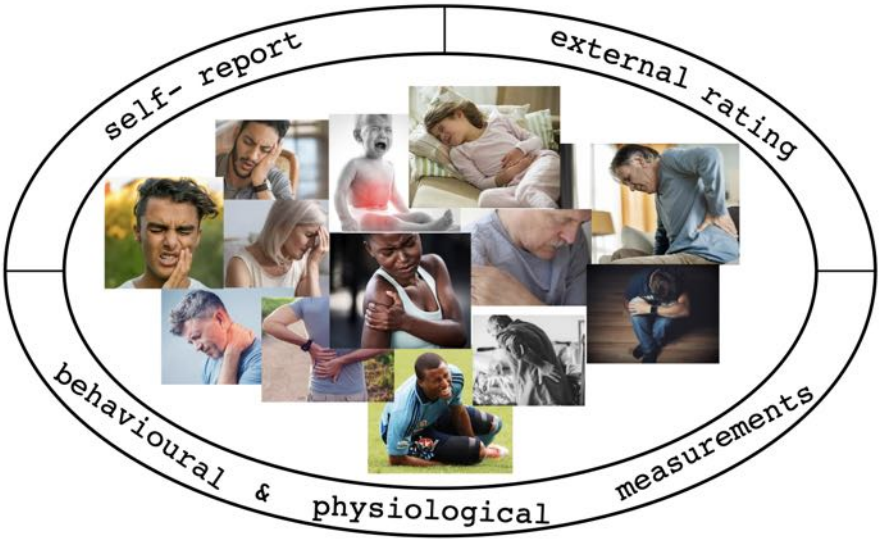


Figure 1.4: *Signs of pain.*

In the first place, we might be tempted to conceive pain as a particular instance of perception, then as a form of somatosensory perception (in particular "interoception"). Thereby, we may identify an object (cause?) of our experience in the environment (including our body), for example a tissue damage. Following this naive perceptual view, every pain experience not linked to perceivable source would be declassified as a delusion. In other words, we could be mistaken about our own pain. This contrasts with our intuitive conception of pain as an unquestionable and subjective experience, where did we go wrong?

Perception involves two planes: reality and appearance. We believe that objects

(colours, shapes, sounds, surfaces, etc.) exist in the external world regardless of our perceiving them and, as a consequence, we feel to reside in a public environment where perception can be easily shared. But pain is different from other perceptual experience in the fact that reality collapse into the appearance plane and the distinction between the perception and a certain state of affairs in the world seems to vanish. Pain seems to be the subjective experience itself instead of a representation of tissue damage (an object like another in the world) and, as such, is private and not perceptually shareable (see Figure 1.5) (Hardcastle, 2015).

Hence, the perceptual view of pain must endeavour to explain the peculiar nature of pain perception with its "sense-content" different from anything in the physical world. The most interesting attempt, proposed by Pitcher (1970), suggests to compare the concept of pain with the concept of glimpse. Glimpses are essentially caught, they have a *to-be-is-to-be-caught* feature, likewise pains have a *to-be-is-to-be-felt* quality[3]. Moreover, both are fundamentally private experience. Nevertheless, it is easy to agree on the fact that glimpses have an uncontroversial physical status in so far as catching a glimpse means just see something in the world. Similarly, Pitcher states that "to feel pain is to *perceive* the disordered state of a part of one's body". If this definition helps to bring back pain into the ground of perceptual phenomena, in particular in the somatosensory field, it still seems to leave partially uncovered the issue of pains without physical injury (e.g. chronic and phantom limb pain). He insists on the possibility to be mistaken in pain perception (e.g. localisation errors, overrated pain) and on the low incidence of those he calls "non-standard" pains. There are still missing pieces.

A different perspective may help us to mark the way of reconciliation. Intentionalism, that is the thesis that the propositional content of the experience exhausts its properties, is considered to have a promising chance to integrate philosophy and neuroscience of perception for the prominent role attributed to sense-content. In this paradigm, drawing a consistent theory of pain always seemed problematic because of the usual burden to identify the content of pain experience without forcing a dual-aspect theory of pain dividing sensory and motivational contents. As Klein (2007) suggests, this issue can be overcome considering non-representational contents. Such as not all the propositions are descriptive or declarative but also imperative, in the same way the content of pain could be considered as a command rather than a description. Then, pain is an imperative sensation demanding a certain kind of action to be satisfied, for example stopping to perform an action potentially damaging. This conception sees pain in a whole new light, stressing its action-related properties, usually overlooked, and making pain a call to action that only incidentally carry out an informative role[4]. Such a view seems to reconnect acute and chronic pain in a unified theory, pouring out specificity in the prescribed action.

Of course, none of these proposals is conclusive. A lot of work must still be done in order to grasp the complex nature of pain experience. While these philosophical reflections offer depth to our understanding, it is worth noting that the marriage between philosophy, consciousness, and computation has been a topic of long-standing debate. This is especially true when we encounter the quandary of simulating deeply subjective

[3]Philosophers usually agree on this pain quality, but there are still different positions.

[4]The debate on the purpose of perception is one of the most interesting in the history of philosophy of mind and cognitive science. It will suffice here to note that the seminal shift from perception as knowledge-oriented to perception as action-oriented is mirrored in these philosophical theories of pain.

experiences, such as pain, in computational models.

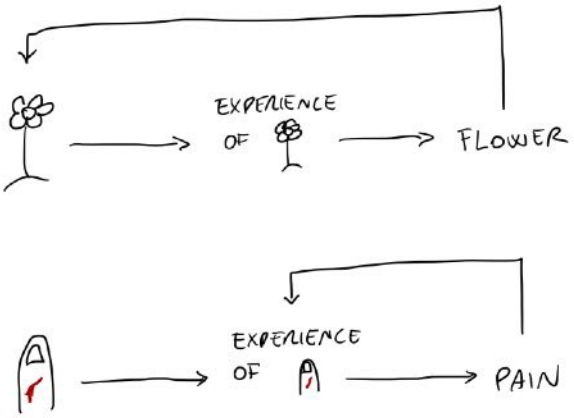


Figure 1.5: *Schematic representation of the difference between visual perception and pain perception to grasp the peculiarity of pain experience. The content of visual perception is the flower, whereas the content of pain perception is the experience itself.*

1.3 Dennett's criticism: exploring the computability of pain experience

The intricate conundrum of consciousness has often been at the forefront of philosophical, neuroscientific, and computational debates. The experiences that encapsulate our sentient existence, particularly those as fundamentally subjective as pain, serve as compelling arenas for these debates. Over the last half-century, philosophy of mind and neuroscience, the disciplines most involved in the dispute, have headed in the opposite direction. In the realm of scientific pain research, there has been a notable shift in the conceptualisation of pain. It has evolved from being perceived primarily as akin to the perception of objective reality to being likened more to emotional experiences, specifically affective states. This transformation began with the delineation of the sensory/affective distinction and a subsequent growing emphasis on the affective dimension of pain (Melzack et al., 1965; Melzack and Casey, 1968; Price et al., 1999; Chapman and Nakamura, 1999; Craig, 2003; Vogt, 2005). In contrast, the trajectory in philosophy, at least until recently, has followed a different course. With the rise of naturalism as a dominant paradigm in the latter half of the 20th century, philosophers increasingly sought ways to assimilate the experience of pain into the realm of ordinary perception, akin to vision and audition (see perceptual/representational views of pain (Bain, 2003; Tye, 2005; Cutter and Tye, 2011)).

Going forward in this line of argument, the philosophical trend of eliminative materialism fosters an intertheoretic reduction (Churchland and Churchland, 1992) of the folk psychological concept of pain to an explanation purely based on neural mechanisms (Baetu, 2020).

Advocating for this purpose, Daniel Dennett, in his influential paper "Why You Can't Make a Computer That Feels Pain" (Dennett, 1978) reflects on incongruous aspects of our common concept of pain. He stems from questioning the possibility of

simulating pain using computational tools. He elucidates on the nature of pain, the constraints of computation, and the gap that resides between them. Is it a chasm too vast to bridge, or are there untrodden pathways that lead towards a synthesis? Dennett's discourse serves as both a beacon and a cautionary tale for those attempting to computationally replicate the experience of pain through simulations.

A solid and clear definition of pain is a necessary condition to make a computationally operating model of pain and, in general, to tackle the issue from a rigorous scientific perspective. Dennett's criticism lies in this first unavoidable step. Pointing out the incongruity rooted in the folk notion of pain, whereby pain is characterised by both a negative valence (pain is deemed to be always unpleasant) and a total dependence on the subjective judgement (sole discretion of the sufferer)[5], he argues for a change of paradigm from the phenomenological personal level, that hinder a scientific analysis of pain, to a sub-personal level of neurophysiological mechanisms, where the concept of pain as we conceive it is doomed to disappear. Since folk psychological definitions and features of pain are logically inconsistent, they should be eliminated from a strictly scientific discourse at its finest level.

Dennett's stance is grounded in a potent blend of philosophical pragmatism and materialism. He challenges the notion that a computational device, under its design or complexity, could truly experience something like pain. Central to his argument is the differentiation between information processing and genuine conscious experience. For Dennett, pain is not merely a collection of data points to be processed, but a deeply subjective experience intertwined with consciousness. He emphasises that while computers can simulate responses and behaviours consistent with folk pain, the genuine phenomenological richness—the 'what-it-is-like' experience of pain—is absent in these digital entities. This raises compelling questions for computational modelling: If genuine pain involves a subjective quality beyond mere information processing, how do we approach its computational representation, especially when extending to affective and social domains?

The challenge, as inspired by Dennett's musings, is to discern where the line between simulation and genuine experience lies in computational models. Given the multilayered nature of pain, with its perceptual, affective, and social components, this endeavour becomes even more intricate. Traditional computational models have predominantly focused on the perceptual component, providing algorithms and networks that mimic the physiological and behavioural responses to noxious stimuli. Yet, as contemporary research indicates, pain's affective and social dimensions introduce layers of complexity that demand more sophisticated modelling approaches.

While Dennett's cautionary arguments present hurdles, they also provide a rich tapestry of considerations for researchers navigating the computational modelling of pain. Leveraging the psychological and neurobiological findings offers a roadmap that can guide the creation of comprehensive models that, if not capturing the full subjectivity of pain, can provide an enriched and nuanced understanding of its diverse components. Our proposed model, inspired by broader theoretical considerations, will be detailed in Chapter 4.

[5]Indeed, these two statements seem to be inconsistent in case of reactive dissociation, such as under lobotomy or morphine, where the affective component of pain (absent) seems to be dissociated from the sensory component (present).

CHAPTER 2

Most influential theories

2.1 The first systematic study: René Descartes

René Descartes (1596-1650), a French philosopher and mathematician, is often credited as the first to provide a mechanistic explanation for pain as part of his broader philosophical work (Descartes, 1969). His interpretation laid the foundations for many future theories by proposing that pain was a direct, mechanical process.

Descartes' concept of pain is most famously illustrated in his description of a boy touching a flame, leading to the disruption of skin and tissue, which in turn pulls on a tiny thread running through the nerve fibers to the brain, triggering the sensation of pain (Fig. 2.1). This model suggested a direct relationship between the source of pain and the perception in the brain, conceptualising pain as a simple message travelling along a fixed pathway.

His mechanistic model assumed a clear, linear relationship between the extent of an injury and the degree of pain felt, suggesting that pain intensity should be directly proportional to the tissue damage (Sullivan, 2008). This theory, however, oversimplified the complexities of pain perception, which we now understand involves not only the transmission of signals but also their modulation by various biological, psychological, and environmental factors.

Descartes' dualism also positioned the mind as a separate entity from the body, which interacted with the body only through the pineal gland, considered the "seat of the soul". This separation of mind and body has been influential, shaping ongoing debates in both philosophy and neuroscience about the nature of consciousness and its interaction with physical processes.

The advancements in neuroscience during the 20th century began to challenge Descartes' straightforward mechanistic model, leading to the development of more sophisticated

understandings of pain that incorporate neurobiological, emotional, and cognitive components. The Gate Control Theory of pain, proposed by Melzack et al. (1965), for example, introduced the idea that pain is not simply a direct result of tissue damage but can be modulated by psychological factors at the spinal level before the signals reach the brain.

Moreover, modern pain research emphasises the subjective nature of pain, recognising that individuals experience pain differently based on genetic makeup, past experiences, current expectations, and cultural contexts. This subjective experience of pain highlights the limitations of Descartes' model, which failed to account for the personal and subjective dimensions of pain, focusing instead on a purely physical and deterministic view of pain perception.

Overall, while Descartes' contributions laid essential groundwork in the study of pain, his theories now serve more as historical milestones that underscore the evolution of our understanding rather than definitive explanations of pain mechanisms. His work reminds us of the need for an integrative approach to studying pain, one that considers the intricate interplay of biological, psychological, and social factors.

Figure 2.1: *Descartes used this picture to describe his conception of pain perception. A body damage (frames in this case) would cause local pores to open so that tubes can convey the spirits to the brain.*

2.2 Specificity theory

The Specificity Theory, introduced by Frey (1895), marked a significant departure from ancient views of pain, which Plato and Aristotle saw as merely intense forms of regular sensory experiences. This theory posits pain as a fundamentally distinct sensory experience, mediated by specialised receptors called nociceptors. These receptors detect harmful stimuli and relay signals through designated nerve pathways to pain centers in the brain.

This delineation emphasises pain's unique physiological and psychological characteristics, distinguishing it from other sensory inputs. Unlike the classical understanding, which suggested that pain was an exaggerated sensation, Specificity Theory identifies

dedicated biological pathways and brain centers responsible for the perception of pain, highlighting its distinct nature.

This framework laid the groundwork for further advances in pain understanding, suggesting that pain is not merely a more intense form of touch but a separate and complex sensory modality. By establishing a clear physiological basis for pain through nociceptors and specific neural pathways, Specificity Theory played a crucial role in shifting pain research towards more nuanced theories that incorporate both biological and psychological aspects, thereby enriching our understanding of pain as a multi-faceted human experience.

2.3 Pattern theory

Pattern theory presents an alternative to specificity theory by suggesting that pain does not have unique sensory receptors. Instead, it posits that the same receptors involved in other sensory processes are responsible for pain perception. According to this theory, pain is distinguished by specific patterns of neural activity that occur when neural firing exceeds normal thresholds in certain spatial and temporal configurations. This concept explains phenomena such as central summation, where prolonged or abnormal neural activity can initiate self-sustaining neural circuits, leading to persistent pain. This theory was instrumental in developing further understanding of how pain can arise from neural processes that are not exclusively dedicated to pain sensing, emphasising a more integrated view of sensory perception (Melzack et al., 1965).

2.4 The two faces of pain

Head and Holmes (1911) introduced a groundbreaking dual system model to elucidate the often puzzling discrepancy between actual nociceptive activation and the perceived intensity of pain. Their model posited two distinct pain-processing systems: the "epicritic" and the "protopathic" systems. The epicritic system is tasked with processing the sensory aspects of pain, such as its intensity and precise location, which is crucial for identifying the exact source and nature of the pain stimulus. Meanwhile, the protopathic system is primarily concerned with the affective and emotional aspects of pain, conveying the unpleasantness and the emotional response to pain.

This dualistic approach to understanding pain was revolutionary and has significantly influenced subsequent theories of pain perception. It laid the groundwork for later developments that conceptualise pain not just as a simple sensory experience, but as a complex, multidimensional phenomenon involving both physical sensation and emotional and motivational responses. The duality of pain is now commonly recognised in the form of sensory-discriminative and affective-motivational subsystems, as initially outlined by Melzack and Casey (1968). This framework has been elaborated upon by subsequent researchers, including Craig et al. (1994), who identified specific neural pathways associated with each subsystem. According to Craig's model, sensory-discriminative information about pain travels to the lateral nuclei of the thalamus and is then processed in the sensory cortex. In contrast, the affective-motivational aspects of pain are processed through pathways leading to the medial nuclei of the thalamus and are subsequently integrated in the anterior cingulate gyrus.

The continuing evolution of the dual systems theory underscores the complexity of pain as a sensory and emotional experience. It acknowledges that pain's protective purpose in signaling harm to the body is intricately linked with psychological factors that can modify pain perception. For instance, the context in which pain is experienced and the emotional state of the individual can profoundly influence the intensity and unpleasantness of pain. This recognition has important implications for clinical practice, as it suggests that effective pain management must address both the physical and emotional components of pain.

2.5 Gate control theory

The gate control theory, proposed by Melzack et al. (1965), represents a turning point in the story of pain comprehension. Taking the best from specificity and pattern theories, the gate control theory merges the concepts of physiological specialisation with the central summation and input control, in the light of physiological evidence on spinal mechanisms.

In the model of Melzack and Wall the substantia gelatinosa (SG) modulates the afferent patterns before they influence the T cells, acting as a gate control system. The afferent patterns in the dorsal horn activate selective brain processes which influence the modulation of the gate control system and the T cells activate neural mechanisms, in particular the action system in charge of response and perception (see Figure 2.2). Even without stimulation, in the spinal cord there is a continuous ongoing activity carried by small fibers that holds the gate in an almost open position. When the receptor-fibers are stimulated, the effect of the stimulus-evoked inhibition is determined by the total number of active fibers, the frequencies of their trasmitted impulses and the balance of activity in large and small fibers. As a consequence, the response of the T cells is not fully determined by the total input. In particular, pain experience arises when T cells reach a threshold.

The theory significantly highlights the role of central influences, such as attention, emotions, previous experiences, and cognitive evaluations, within the gate control system prior to the activation of the action system. The authors propose that either the dorsal column-medial lemniscus system, which transmits information from the body to the primary somatosensory cortex, or the faster dorso-lateral pathway, projecting to the brainstem and thalamus, could serve these modulatory functions. It is particularly noteworthy that central activities can influence pain perception at various stages throughout the temporal and spatial development of the sensory input.

For this study, it is pertinent that the action system, responsible for pain perception and response, is activated only after central modulation of the sensory input. According to this model, the "signs of pain" such as autonomic and behavioural responses, serve as indicators of the overall pain experience, rather than merely the sensory aspect.

Overall, this theory has the undoubted merit of having emphasised, for the first time, the role of the central nervous system (CNS) in the control of pain perception integrating upstream processes with downstream modulation, leading to a novel theory of pain going beyond the classical view of pain as an inevitable sensory response to tissue damage. Although the theory does not precisely account for long-term changes in the CNS, it is now clear that learning (or, more generally, plasticity) has a role in

pain and can contribute to explaining chronic pain, phantom limb pain, and synaptic potentiation occurring after repetitive noxious stimulation.

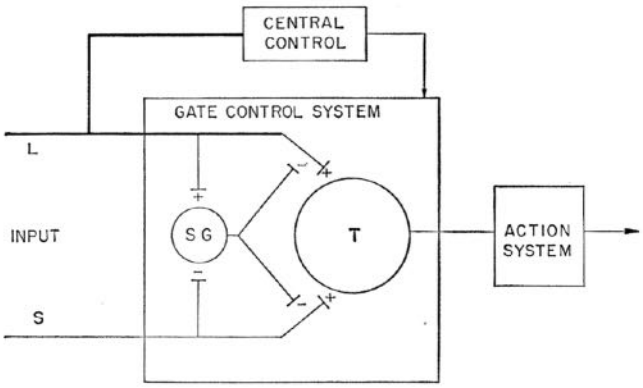


Figure 2.2: *Gate control mechanism with L, the large-diameter fibers, and S, the small-diameter fibers. The inhibitory effect of the substantia gelatinosa (SG) is increased by activity in L and decreased by activity in S. In turn, the central control system, receiving inputs from L, affects the gate control system. Image from Melzack et al. (1965).*

2.6 Neuromatrix

Building upon the recognition of the central nervous system's (CNS) pivotal role in pain processing, Melzack (1989) suggested the presence of a neural network, called the body-self neuromatrix, an effective brain system in charge of pain experience which integrates multiple inputs to produce the output pattern that evokes pain. The neuromatrix includes parallel somatosensory, limbic, and thalamocortical components that subserve the sensory-discriminative, affective-motivational, and evaluative-cognitive dimensions of pain experience (Melzack, 1999). This network acts as a neural program, genetically built, where the inputs are:

- cutaneous, visceral, and other somatic sensory inputs;
- visual and other sensory inputs that influence the cognitive interpretation of the situation;
- phasic and tonic cognitive and emotional inputs from other areas of the brain;
- intrinsic neural inhibitory modulation inherent in all brain function;
- the activity of the body's stress-regulation systems.

After the computation, the neuromatrix produces a "neurosignature" output consisting of patterns of nerve impulses of varying temporal and spatial dimensions which determines the particular qualities and other properties of the pain experience (see Figure 2.3).

This theory is another step toward perceiving pain as a multidimensional experience involving cognitive, emotional, and above all contextual processing of information beyond the pure physical injury.

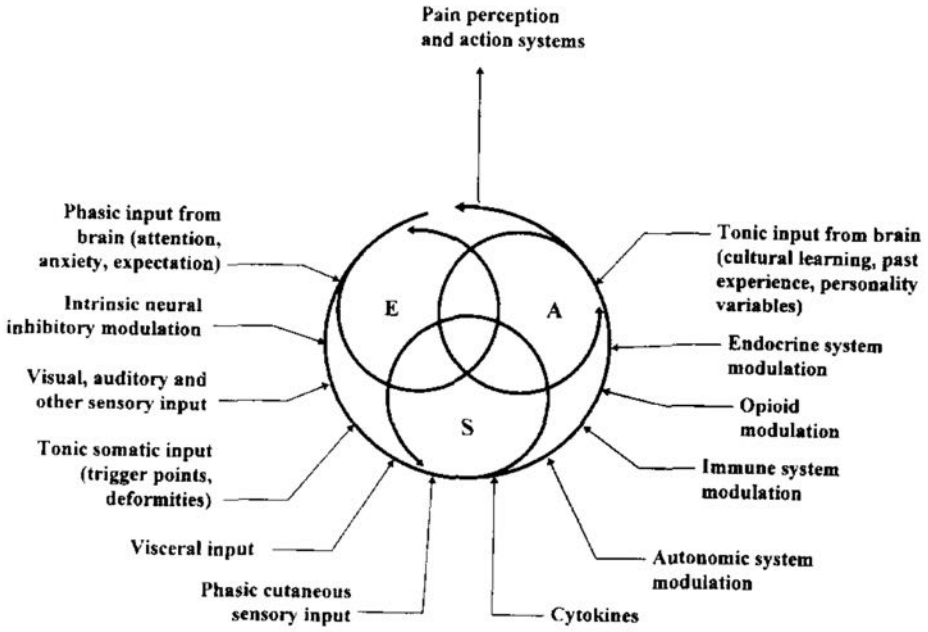


Figure 2.3: *The body-self neuromatrix from Melzack (1999).*

2.7 Motivation-decision theory

Expanding upon behavioral insights, Fields (2006) introduced a motivational approach to model pain to account for the different psychological factors influencing pain, such as attention, expectancy, and strong emotional states, intending to go beyond the limits of psychophysical findings which establish a relation between the subjective experience and the physical property of the stimuli, overlooking the processes in the middle. Rather, Fields posits that the variability observed between the intensity of a noxious stimulus and the resultant pain experience can be comprehended through the lens of a decision-making process.

This model recognises that individuals frequently face decisions between competing behaviours driven by conflicting motivations. In the context of pain, this may involve choosing between responding to the immediate motivations elicited by a noxious stimulus (such as avoidance or seeking relief) and pursuing actions driven by conflicting motivations like endurance or persistence for a delayed reward. This choice necessitates a level of nociceptive inhibition, facilitated by the pain descending modulation system, which has been previously noted. A pivotal feature of this modulatory circuit is its ability to exhibit ambivalence; both inhibition and facilitation are possible through the activation of OFF cells and ON cells, respectively. Consequently, pain perception can be moderated to enhance reward attainment or intensified if immediate pain relief is prioritised. Essentially, the direction of modulation hinges on a decision-making process, where the primary decision revolves around whether to engage with the noxious stimulus (illustrated in Figure 2.4).

The operational efficiency of this system is predicated on the immediacy of modulation, which ideally occurs before the conscious perception of pain. In this schema,

early sensory and contextual evaluations, along with the resultant decision-making process, do not culminate in an immediate conscious experience of pain. It is only after the decision-making phase that the psychophysical manifestations of pain become apparent. Fields proposes that projections in the Nucleus Accumbens (NAc) from putative nociceptive neurons in the dorsal horn contribute to this decision-making process without inducing a conscious perception of pain. Furthermore, he identifies the mesolimbic dopamine circuit as crucial for pain modulation, noting that the NAc projects to the hypothalamus and amygdala, which in turn communicate with the Periaqueductal Gray (PAG) and the Rostroventral Medulla (RVM).

What makes this theory particularly compelling is its portrayal of pain as a goal-directed experience, wherein the stimulus not only triggers a motivational state but also dictates the appropriate responsive action. This perspective resonates with Klein's conceptualisation of pain as an imperative sensation that demands specific actions (see Section 1.2), thus integrating the active participation of the individual in pain phenomenology.

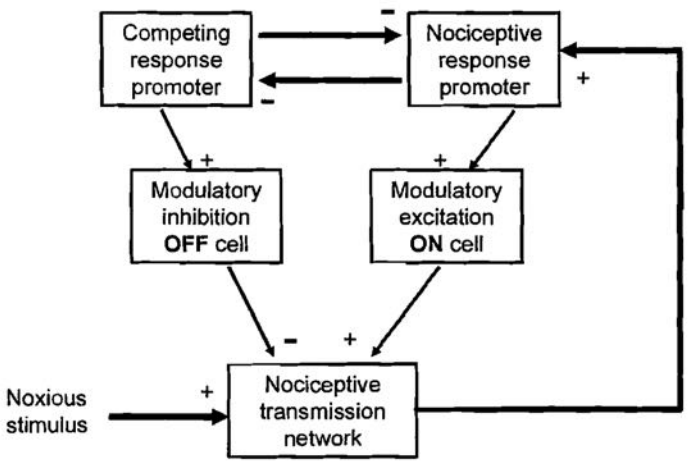


Figure 2.4: *Noxious stimulation activates promoters of nociceptive response which, in turn, trigger OFF or ON cells according to the selected strategy to either facilitate or inhibit pain perception. Image from Fields (2006).*

2.8 Hierarchical processing of pain

To deepen our understanding of the interplay between the sensory and affective components of pain, Price's Hierarchical Theory distinguishes two primary dimensions of pain: sensation and affection. The affection component is further subdivided into the immediate experience of unpleasantness, which encompasses feelings of distress and fear, and the secondary pain affect, which refers to the long-term emotional and cognitive implications of enduring pain (Price, 2000).

According to this theory, the sequential interaction among pain sensation, unpleasantness, and secondary pain affect is orchestrated along specific neural pathways and brain mechanisms, as illustrated in Figure 2.5. While the distinction between intensity and affective experiences is substantiated by various studies (Price et al., 1987;

Rainville et al., 1992), the direction or, possibly, the causal relation between these two components is less evident. The results reported in (Rainville et al., 1999) are given as evidence for the direction of causation between sensory and affective component of pain. In both experiments subjects underwent painful stimulation and hypnotic suggestions were used to sway pain unpleasantness in the first experiment and pain intensity in the second. Results show that both ratings were influenced by intensity suggestions, while unpleasantness indications modulate only unpleasantness ratings. These results support sequential model of intensity-unpleasantness-secondary affect and tally with other evidence (Price, 2000) (Price et al., 1987).

Interestingly, this theory introduces a new distinction between two different types of affective dimension, one more immediate and related to the classical affective sphere and the other one featured with more cognitive elements. The secondary pain affect involves elaborate reflections grounded in memory and imagination about meaning and implications for the future, resting on a deeper level of abstraction.

Moreover, this theory foreshadows recent findings about hierarchical brain elaboration of signals and the way their features are transduced by the sensory surfaces of the body. Progressively, during the elaboration, modality-specific high-dimensional features are summarised in lower-dimensional features in a process of abstraction (Katsumi et al., 2021), along what is called *representation-modulation* or *transmodal-unimodal* gradient (Margulies et al., 2016). Along these lines, in the context of pain, the sensory component could represent just the first step of elaboration and the affective features, requiring higher levels of abstraction, may show up after an integration process between exteroceptive and interoceptive instances.

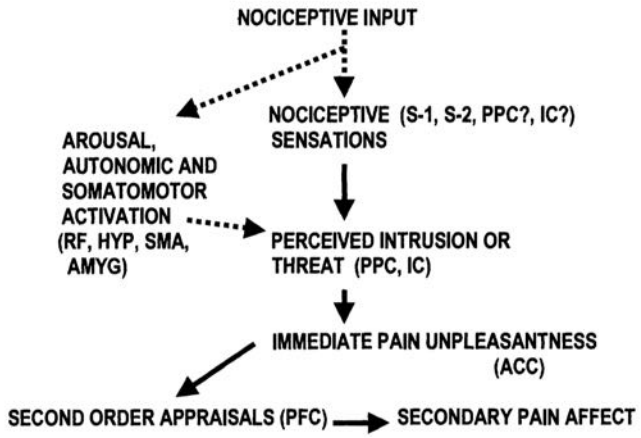


Figure 2.5: *The scheme illustrates the interaction between the three components of pain according to Price's Hierarchical Theory: pain sensation, pain unpleasantness and secondary pain affect. Image from Price (2000).*

2.9 Cognitive-behavioural theory of pain

Diving into psychological learning theories and the resulting strategies to treat pain can give us another perspective on pain-related mechanisms. Essentially, the difference

between behavioural and cognitive learning theories lies in the object of learning. The former view considers specific actions as the main result of every learning process. Thus, learning means finding the right association between the stimulus and the set of possible actions to better cope with the situation. Whereas the latter goes further considering also and especially the mental representation, namely the influence of the cognitive categorisation of the situation.

From these theoretical foundations, pain is understood through various lenses. The behavioural perspective, which has been prevalent in clinical practice and psychological research for decades, posits that pain behaviours are governed by conditioning influences similar to other behaviours (Fordyce, 1996). This framework supports behavioural interventions for pain, which have proven effective over recent years (Sharp, 2001). However, the success of these interventions does not unequivocally validate the behavioural model. Closer examination suggests alternative interpretations, as pointed out by Turk (1996), particularly highlighting the model's failure to account for patients' interpretations of environmental changes which can be pivotal in many scenarios.

Cognition's role in the emergence and modulation of pain is increasingly recognised, influencing reports of pain intensity, coping strategies (Jensen et al., 1991), mood, and pain-related disability (Turk and Rudy, 1992; Williams and Keefe, 1991; Wilson et al., 1993). Over the years, research has acknowledged the significance of cognitive factors such as appraisals, interpretations, expectations, and perceptions of control, particularly in the context of chronic pain. There is strong evidence linking pain-related cognitions and beliefs with the use of coping strategies, which in turn affects the extent of pain's impact on an individual's life. In chronic conditions, these cognitions and beliefs are further reinforced by avoidance behaviours and stress-induced autonomic arousal, creating a feedback loop that exacerbates pain perception and disability (Sharp, 2001).

Sharp (2001) has proposed a comprehensive cognitive-behavioural model that integrates these elements with existing psychological evidence (see Figure 2.6). Notably, the model introduces two critical elements to our discussion: learning history and cultural background. These factors shape individuals' pain-related cognition and beliefs, based on their personal and cultural experiences with pain, influencing how they categorise and respond to pain.

This approach aligns with empirical evidence and is particularly insightful regarding the appraisal of autonomic responses, where physiological reactions are interpreted and assigned meaning within an interoceptive context. The role of anxiety, catastrophisation, and fear in modulating pain is evident, emphasising how these factors influence the interpretation of physical sensations.

Lastly, the model highlights the potential iatrogenic effects of inappropriate or excessive medical interventions, which may reinforce pain symptoms and disability. This phenomenon might be explained by patients' attempts to legitimise their pain, as suggested by Kouyanou et al. (1998). The clinician's response can significantly influence patients' beliefs about their pain, thereby affecting the perceived intensity of pain and the level of disability experienced. This interaction underscores the importance of considering placebo and nocebo effects, which, though rooted in prior beliefs, can be seen as a form of iatrogenic influence.

In conclusion, given the limitations of operant models, an updated cognitive-behavioural

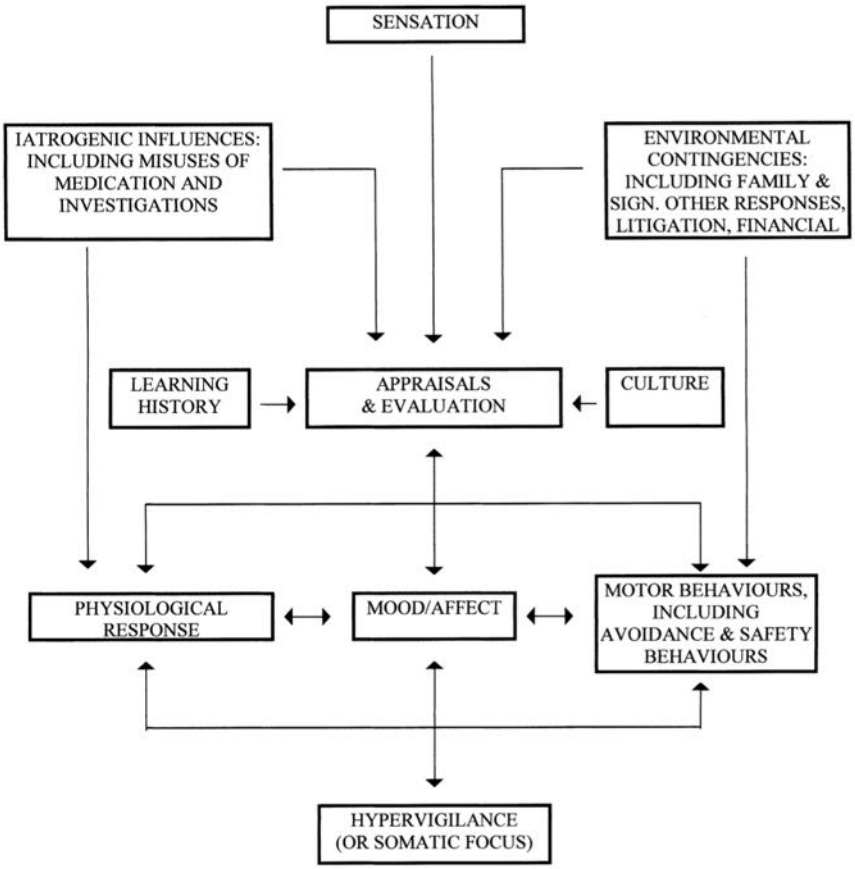


Figure 2.6: *The cognitive-behavioural model of chronic pain from Sharp (2001).*

framework appears well-suited for integrating the latest scientific findings with the rich tradition of cognitive and behavioural theories in psychology.

CHAPTER 3

Neurobiological Groundings: Shaping Model's Constraints

> The concept of a "pain center" in the brain is totally inadequate to account for the sequences of behavior and experience. Indeed, the concept is pure fiction, unless virtually the whole brain is considered to be the "pain center".
>
> — Melzack et al. (1965)

There is no single center of pain in the brain specifically accountable for pain information processing, nevertheless, neurobiological researchers strive to identify a minimal network of structures receiving parallel inputs from multiple nociceptive pathways, the so-called "Pain Matrix". Pain imaging studies are not always consistent in identifying specific activation sites due to differences in technical procedures.[1] Despite this divergence, Apkarian et al. (2005) propose a meta-analysis that identifies six most commonly reported areas (ACC, S1, S2, IC, Th, PFC). This network of somatosensory (S1, S2, IC), limbic (IC, ACC) and associative (PFC) structures is consistent with most of the analysed studies.

Somatosensory cortex (S1, S2) is related to the processing of pain's sensory features. In contrast, insular cortex (IC) and anterior cingulate cortex (ACC) are associated with the affective component of pain and, lastly, prefrontal cortical areas (PFC) seem to be implicated in cognitive processing of pain (attention, memories, expecta-

[1] EEG and MEG analyses struggle with the increase of distance from the scalp but benefit from a better temporal resolution. On the contrary, hemodynamic imaging studies (typically PET and fMRI) are more sensitive to detect activation in deeper regions but are lacking in temporal resolution. As a consequence, pain-related areas are best identified by hemodynamic imaging methods, while the temporal sequence and time delays to activating different cortical regions are best studied with EEG and MEG methods (Apkarian et al., 2005).

tion and evaluation). Of course, it is necessary to point out that these areas are strictly interconnected and constitute a nondecomposable system (Pessoa, 2008), then, this conventional separation is just a useful approximation.

In physical pain (acute and chronic as well), these structures are activated by nerve impulses transmitted from the dorsal horn (DH) by projection neurons, once the peripheral noxious stimuli are processed at spinal level. From DH the nociceptive signal is propagated in the rostral ventromedial medulla (RVM) and periaquieductal gray (PAG) through the spinobrachial pathway (Ossipov et al., 2010).

Examining the neural areas involved in pain processing and what we know about their role and their interactions, it is easier to understand the complexity of pain processing.

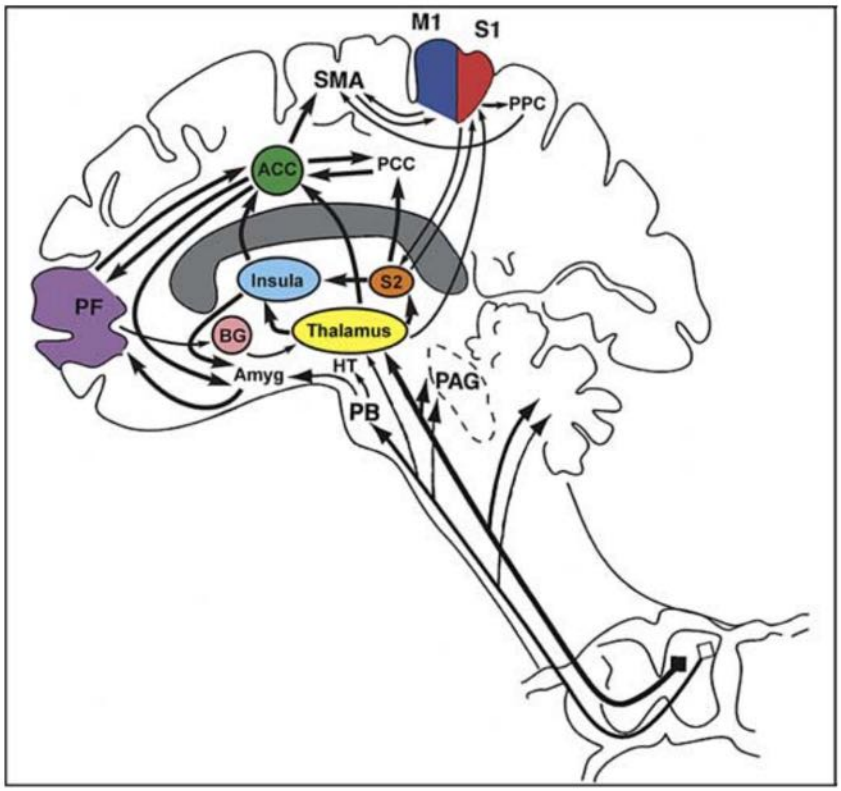


Figure 3.1: *Cortical and sub-cortical regions involved in pain perception, their interconnectivity and ascending pathways (image from (Apkarian et al., 2005)).*

3.1 Significant neural components

The principal brain regions involved in pain processing, with their commonly assigned function, are presented below.

OFC. The orbitofrontal cortex, along with the amygdala and the anterior insula, is a key structure for central emotional control (Adolphs, 2002). Its activation during pain experience is well documented and its connections support an involvement in the

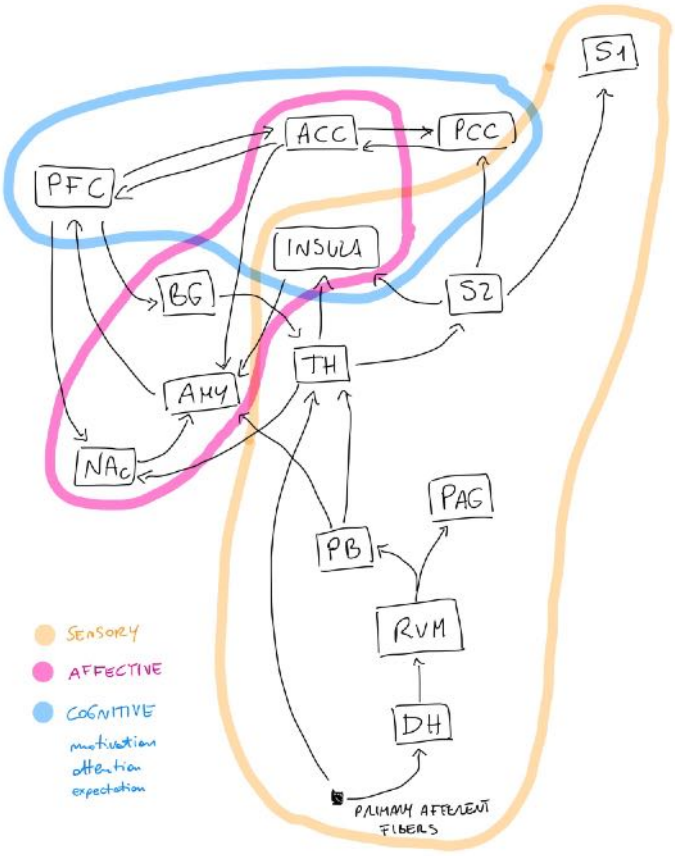
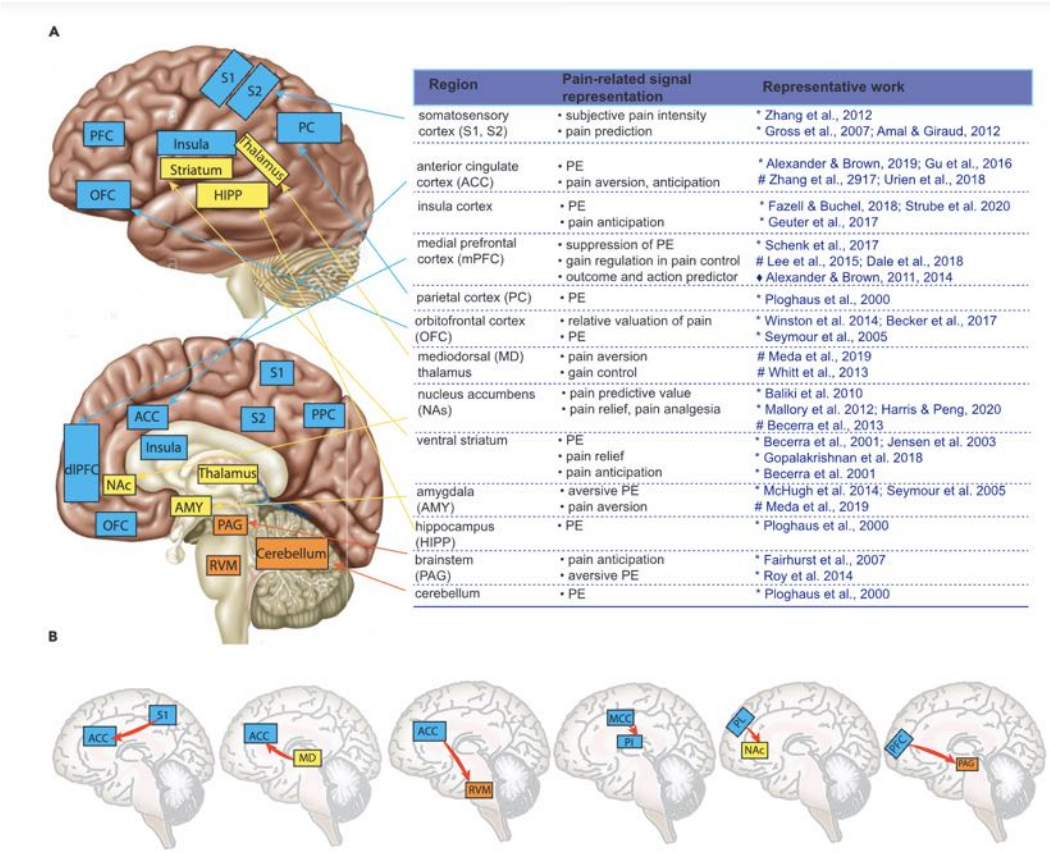


Figure 3.2: *Neural mechanistic diagram of regions involved in pain perception, their interconnectivity, ascending pathways and their role in sensory, affective and cognitive processing. Adapted from (Apkarian et al., 2005). Note that the tripartition is a suggestive oversimplification.*

integration of sensory and autonomic information (Kringelbach, 2005). Indeed, OFC receives input from all sensory modalities (including visceral) and has direct reciprocal connections with brain regions involved in pain perception, such as ACC and PAG (Cavada et al., 2000).

PFC. The lateral prefrontal cortex (lPFC) is indicated as the center for cognitive process and for integration of this processed information with affective and motivational information (Buhle et al., 2014) (Ochsner and Gross, 2005). Recent findings indicates the existence of a negative association between catastrophising and pain-anticipatory brain activity, this reduced activity contributes to the hyperalgesic effect of catastrophising (Loggia et al., 2015).

The ventrolateral prefrontal cortex (vlPFC) and anterolateral prefrontal cortex (alPFC) are involved in emotion regulation and implementation of cognitive strategies that reduce negative emotional experience (reappraisal). In particular, an activation in this area increases during self-controlled pain stimulation and it is negatively correlated



Region	Pain-related signal representation	Representative work
somatosensory cortex (S1, S2)	• subjective pain intensity • pain prediction	* Zhang et al., 2012 * Gross et al., 2007; Amal & Giraud, 2012
anterior cingulate cortex (ACC)	• PE • pain aversion, anticipation	* Alexander & Brown, 2019; Gu et al., 2016 # Zhang et al., 2917; Urien et al., 2018
insula cortex	• PE • pain anticipation	* Fazell & Buchel, 2018; Strube et al. 2020 * Geuter et al., 2017
medial prefrontal cortex (mPFC)	• suppression of PE • gain regulation in pain control • outcome and action predictor	* Schenk et al., 2017 # Lee et al., 2015; Dale et al., 2018 ♦ Alexander & Brown, 2011, 2014
parietal cortex (PC)	• PE	* Ploghaus et al., 2000
orbitofrontal cortex (OFC)	• relative valuation of pain • PE	* Winston et al. 2014; Becker et al., 2017 * Seymour et al., 2005
mediodorsal (MD) thalamus	• pain aversion • gain control	# Meda et al., 2019 # Whitt et al., 2013
nucleus accumbens (NAs)	• pain predictive value • pain relief, pain analgesia	* Baliki et al. 2010 * Mallory et al. 2012; Harris & Peng, 2020 # Becerra et al., 2013
ventral striatum	• PE • pain relief • pain anticipation	* Becerra et al., 2001; Jensen et al. 2003 * Gopalakrishnan et al. 2018 * Becerra et al. 2001
amygdala (AMY)	• aversive PE • pain aversion	* McHugh et al. 2014; Seymour et al. 2005 # Meda et al., 2019
hippocampus (HIPP)	• PE	* Ploghaus et al., 2000
brainstem (PAG)	• pain anticipation • aversive PE	* Fairhurst et al., 2007 * Roy et al. 2014
cerebellum	• PE	* Ploghaus et al., 2000

Figure 3.3: *Brain regions involved in the representation, prediction, and pain regulation. The lateral and medial views of the brain map highlight key nodes of the pain matrix in the brainstem, cerebellum, and cerebral cortex. The accompanying table summarises representative studies, including human functional imaging studies (*) and electrophysiology studies in animals (#), as well as computer simulations (A). Additionally, the illustration depicts several causally identified direct pathways in pain regulation (B). Figure from Chen and Wang (2023).*

with subjective pain intensity (Wiech et al., 2006).

Different studies suggest that these regions are implicated in high-level pain-modulatory mechanisms involving expectations, beliefs, and judgements about pain (Wiech et al., 2006) (Salomons et al., 2007) (Wiech et al., 2008).

ACC. Anterior cingulate cortex seems to be implicated in high level pain processing, integrating cognitive and emotional features with sensory component. The dorsal part (dACC) is connected with PFC which, as already depicted, process cognitive information, whereas the ventral part (vACC), connected with the amygdala, is involved in assessing emotional and motivational salience (Ossipov et al., 2010). ACC seems to encode (at least partially) pain unpleasantness as the result of an evaluation of the aversiveness of pain (Bushnell et al., 2013).

The long latency of its activation corroborates the role of ACC in cognitive-evaluative stages of pain processing (Apkarian et al., 2005).

Amygdala. The amygdala is a key structure for the emotional-affective dimension of pain and for pain modulation but its connections with cortical sites suggest a role in cognitive aspects and likely mediates expectation and placebo analgesia by PAG activation (Ossipov et al., 2010).

Insular Cortex. The anterior part (AIC) is anatomically contiguous to PFC, while the posterior part (PIC) is closer to sensory component of pain. Due to the heterogeneity of its anatomy it may represent the integration hub of emotional, cognitive and perceptual processing (Craig, 2003) (Brooks and Tracey, 2007).

NAc. Activity in NAc has a key role in motivated behaviour (Salamone and Correa, 2012) (Schwartz et al., 2014) and its activation correlates with both the subjective experience of pain as well as the transition to chronic pain (Scott et al., 2006) (Baliki et al., 2010). The NAc seems to encode mainly affective value and salience of the stimulus rather than sensory information (Brischoux et al., 2009)

PAG. Periaqueductal gray has a pivotal role in descending pain modulation, channelling information from higher structures into the medulla. It is the target structure for chronic pain intervention since its direct stimulation causes a significant analgesia (Richardson and Akil, 1977).

Imaging studies show its involvement in pain modulation through attention; distraction increases activity within the PAG and consequent pain rating, suggesting a part in attentional analgesia (Tracey et al., 2002).

RVM (ON cells and OFF cells). The rostral ventromedial medulla receives inputs from the thalamus, the parabrachial region and the noradrenergic locus coeruleus, and is considered to be the final common relay in descending modulation of pain, projecting to the spinal dorsal horn (Ossipov et al., 2010). In this region, neurons defined as ON and OFF cells (according to their role in pain enhancement or inhibition) underpin pain modulation (Urban and Gebhart, 1999).

3.2 Pain descending modulation

The descending pain modulatory system is a well-characterised network regulating nociceptive processing to produce pronociception (pain facilitation) and antinociception (pain inhibition) acting mainly on the dorsal horn. Spinal cord excitability is influenced by descending input deriving from higher centers of the brain. The brain regions most involved in this descending modulation are frontal lobe, ACC, IC, amygdala, thalamus, PAG and RVM (Ossipov et al., 2010). Along descending path, the afferent noxious stimuli are modulated by inhibitory neurons recruited by stimulation of non-nociceptive Aβ fibres and excitatory cells recruited through nociceptive Aδ and C fibres (see Figure 3.4). The inhibitory role of Aβ fibres underpins the conditioned pain modulation (CPM), phenomenon in which a conditioning stimulus influences the test stimulus, but it should be noted that a distinct endogenous pain modulation system is implicated in placebo/nocebo effect not involving Aβ fibers, denoting the presence of a higher level pain control system (Damien et al., 2018).

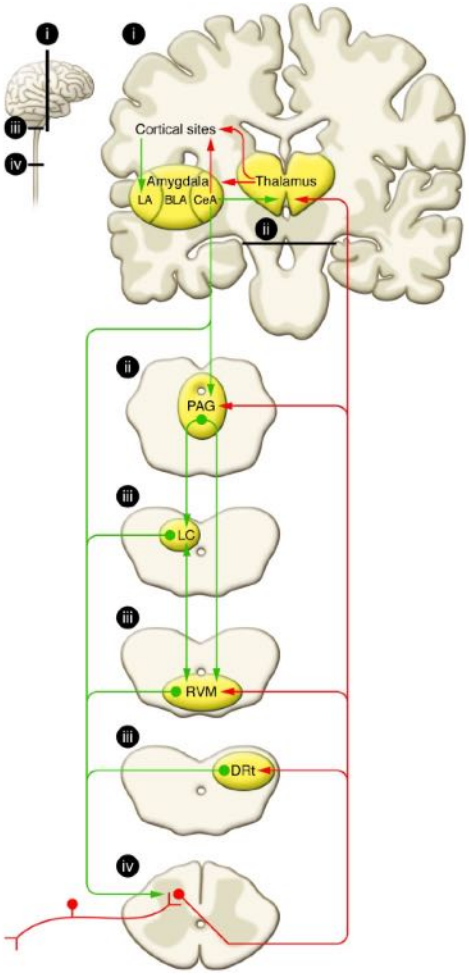


Figure 3.4: *Ascending nociceptive processing (red) and descending modulatory path (green). Image from (Ossipov et al., 2010).*

3.3 A critical review of Pain Matrix

The idea of a "Pain Matrix" as an ensemble of circuits and anatomically determined regions in the brain responsible for pain perception has originated from a distortion of the "neuromatrix" concept proposed by Melzack in 1989 (Melzack, 1989) (see Section 2.6). In brief, the original idea stemmed from the awareness of a lack of evidence for specific cortical regions responsible for pain perception. Hence, the neuromatrix was conceived as a widespread and distributed ensemble of nonnociceptive and nociceptive neurons for which pain was just one of many possible outputs, clearly not specific to pain (Iannetti and Mouraux, 2010). However, in the last few decades, a stronger stance has spread among pain researchers, that is the idea of a Pain Matrix as an enumeration of pain-specific structures in the brain, with some areas having specialised subfunctions (for example, see (Brooks and Tracey, 2005), (Ingvar, 1999) and (Ossipov et al., 2010)).

However, as highlighted by Iannetti and Mouraux (2010), the state of affairs is more tangled and subtle. Even if there are evidences for specific cortical structures activated by nociceptive stimuli and sensitive to pain intensity, there are compelling evidences

against this view that seems to pave the way for a different hypothesis. Starting from anatomic findings, in the primary somatosensory cortex neurons responding preferentially to nociceptive stimuli have never been identified and, the so-called "nociceptive-specific neurons", which respond to high-intensity sensory stimuli, are sparsely distributed in different areas. Second fact, several studies have shown that the magnitude of brain responses can be dissociated from both the intensity of the stimuli and the intensity of perceived pain (Iannetti et al., 2008). In addition, the magnitude of the events-related potentials elicited (ERP) in these areas could not be related to perceived pain even when it is consistent with nociceptive stimuli, for example when two equal stimuli are presented with a short interval (Lee et al., 2009). But interestingly, the stimulus repetition affects the magnitude of the response only for constant inter-stimulus intervals, that is to say when the stimulation is predictable. A certain "novelty effect" can explain this modulation, suggesting a relation between the pain matrix and unpredictable stimuli. Further, other studies suggest that the Pain Matrix, as it was characterised, mainly reflects multimodal neural activity; nociceptive, somatosensory, auditory, and visual stimuli elicited indistinguishable responses, with no nociceptive-specific neural activity (Mouraux et al., 2011) (Mouraux and Iannetti, 2009).

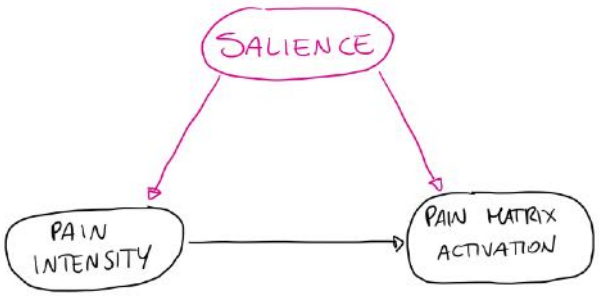


Figure 3.5: *Graphic representation of salience as a confounding variable for pain intensity and pain matrix activation. This hypothesis suggests that the observable correlation between pain intensity and Pain Matrix activation is a spurious association caused by the intrinsic salience of pain experience.*

These findings, taken together, point to a suggestive relationship between the Pain Matrix and the processing of salience, where salience can be seen as a confounder of pain intensity and neural activation (see Figure 3.5). To support this connection, the pain matrix response appears to be largely correlated with the salience of sensory inputs (Downar et al., 2000) and, furthermore, the identified "salience network" includes several regions of the Pain Matrix (Seeley et al., 2007) resulting in a consistent overlap.

However, it is important to acknowledge that the discussion presented here simplifies some aspects of the attentional processes involved in pain perception. The concept of salience, while crucial, may be more accurately interpreted when complemented by the notion of top-down relevance. Although salience can be considered a function of the physical characteristics of stimuli (bottom-up), the organism, following new experiences, attaches a homeostatic value to the sensory event, modulating its behavioural relevance (Valentini et al., 2023) (Torta et al., 2017). This top-down relevance reflects

how prior experiences, expectations, and cognitive evaluations influence the salience attributed to stimuli, as discussed by Valentini et al. (2023). Thus, the interaction between bottom-up salience and top-down relevance offers a more nuanced understanding of how the brain processes and prioritises pain.

Of course, following this theory, the persuasion of the Pain Matrix as an ultimate pain processing center and the related experimental evidences can be explained with the intrinsic high saliency content of a noxious stimulation. Likewise, this view can easily account for the Pain Matrix activation watching someone else's pain (Singer et al., 2004) (Jackson et al., 2005) or cues indicating the delivery of a noxious stimulus (Cheng et al., 2007) and also experiencing social rejection (Eisenberger et al., 2003).[2]

Ultimately, the aim to isolate a pure neural representation of pain in the brain seems doomed to fail. Nevertheless, a different placement of pain as a multimodal high-saliency phenomenon can provide novel insights into its nature and mechanisms.

3.4 The role of salience network

As we have already seen, salience, that is the importance we assign to a given signal, has been hypothesised to have a critical role in pain processing. However, we are left with the problem of finding an interpretation of phenomena such as chronic pain, placebo effect, and subliminal priming pain modulation in light of the pain-salience relationship.

Starting from its function, salience mechanism allows us to establish an importance order among multiple sensory inputs through an amplification of neural processing of salient inputs to enhance adaptive responses and promote an effective interaction with the environment. The brain circuit responsible for this task appears as a well-defined unique network encompassing different regions including the cingulate cortex, IC, basal ganglia, temporal lobe, and medial prefrontal cortices (Borsook et al., 2013), with ACC and IC likely the focal points of the network. Interestingly, aberrant salience processing is documented for several diseases such as Parkinson, obesity, schizophrenia, addiction, and, most importantly in this context, chronic pain. In this regard, several studies suggest that chronic pain syndromes may share a common predisposition to morbid activation of salience network due to a different baseline or "salient state" leading to a top-down control which prioritises pain-related perceptual contents also when they are non-painful *per se* (see Shimo et al. (2011) for visual and Eck et al. (2011) for textual stimulation).

Analysing the functional connectivity of the salience network in subjects with chronic pain, Cauda et al. (2010) report an alteration of the dynamics of the salience network of the resting state comparable to the one detectable in clinical conditions characterised by impaired attention. Other fMRI evidence agrees on the dysfunctional salience network for subjects with different types of chronic pain (migraine, irritable bowel syndrome, temporomandibular disorders) (Borsook et al., 2013). As suggested by behavioural evidences, the salience network aberration makes non-pain related stimuli less salient

[2]Note that, within this theoretical horizon, the occurrence of chronic pain may appear unexplained if not conflicting. Subjects suffering from chronic pain usually show an increased expectation of pain that seems inconsistent with this salience-based perspective. A way out is to detach the concepts of salience and surprise, in the sense that something not surprising may still be salient.

resulting in dysregulated cognitive processing, which means slower reaction times and poor performance in various cognitive tasks (Eccleston and Crombez, 1999). Moreover, subliminal priming with pain-related cues can alter brain and behavioural response in healthy subjects, reducing pain tolerance also when they are unaware of the displayed cues (Meerman et al., 2011).

Following these suggestions, chronic pain could be explained as a salience impairment, where the salience can be manipulated by suggestion, distraction and placebo. Also, it is worth noting the similarity of the salience condition in chronic pain disorder and in psychosis or schizophrenia. Not surprisingly, chronic pain is often coupled with other disorders including anxiety, depression, and catastrophising.

3.5 From Reactive to Prospective Regulation: The Bayesian Brain Hypothesis

> If we want everything to stay as it is, everything has to change.
>
> — *The Leopard*, Tomasi di Lampedusa

The transition from a reactive to a prospective regulation in neural systems represents a significant evolution in our understanding of brain function. Traditional models have primarily viewed physiological responses as mechanisms that defend against deviations from a homeostatic setpoint. However, recent theories suggest a more dynamic framework underpinned by the principles of allostasis.

The concept of allostasis introduces a proactive system that adjusts internal parameters in anticipation of changing demands, rather than merely reacting to disturbances. This paradigm shift is encapsulated by the idea of the brain as a predictive organ, continually forecasting and adjusting based on expected future states. This approach incorporates the Bayesian brain hypothesis, which posits that the brain constructs and continuously updates a probabilistic model of the world through sensory input (Corcoran et al., 2020).

3.5.1 Three-Level Neural Hierarchy and Allostatic Control

In the context of the Bayesian brain, control systems are envisaged as operating across a three-level neural hierarchy, which enables an integration of various physiological signals and cognitive assessments to optimise behavioral responses (Stephan et al., 2016). At the lowest level, reflexive neural circuits handle immediate physiological responses. The intermediate level integrates these reflexes into larger behavioural strategies that consider past experiences and predicted future states, thereby maintaining physiological coherence across varied environmental contexts.

At the highest level, cognitive processes evaluate the potential long-term consequences of different behavioural strategies, integrating complex variables including social interactions and distant future needs. This hierarchical structure allows for a nuanced modulation of responses, where lower levels handle habitual or reflexive actions and higher levels manage complex decision-making scenarios involving uncertainty and prediction.

3.5.2 Allostatic Modulation in Pain Perception

Allostatic modulation allows the organism to preemptively adjust to expected challenges or stressors, thereby minimising potential disruptions. The effectiveness of such a system is based on its capacity to reduce the surprise or prediction error associated with unexpected events. In the realm of pain management, this implies a sophisticated interplay between sensory inputs, expected pain levels, and cognitive evaluation of potential threats or injuries (Peters et al., 2017).

Pain, often seen merely as a sensory experience, can also be understood through the lens of the Bayesian brain as a complex predictive model in which the brain anticipates potential damage and modulates the perceptual and emotional experience of pain accordingly. This predictive model integrates past pain experiences, current physiological states, and contextual cues that might predict future pain, allowing a more adaptive and proactive pain response.

In summary, the Bayesian brain hypothesis offers a compelling framework for understanding the brain's predictive capabilities, particularly in how it modulates physiological responses like pain through allostatic mechanisms. This perspective not only improves our understanding of brain function, but also opens new avenues for therapeutic interventions that can preemptively modulate pain perception, targeting predictive rather than reactive components of pain management.

CHAPTER 4

The Computational Modelling of Pain: An Overview and a Roadmap

This chapter introduces a draft of a general model framework that serves as a pivotal foundation for the analysis and discussion of various pain models presented in the literature. This chapter aims to outline a systematic approach to understanding and categorising these models, facilitating a deeper insight into their mechanisms and applications. By establishing a blueprint for defining models across different levels of complexity, this framework seeks to bridge the gaps between theoretical understanding and practical application.

The chapter starts by critically examining existing models, assessing their strengths and weaknesses in capturing the multifaceted nature of pain. Through this analysis, it becomes clear that while a theoretical model should encompass the sensory, cognitive, and affective dimensions of pain to ensure completeness, the practical implementation of such models may vary. Depending on specific circumstances and objectives, it is essential to discern which class of models is most appropriate for the situation at hand. This realisation leads to the proposal of a flexible approach to model development, capable of adapting to the dynamic requirements of different application contexts.

Furthermore, this chapter sets the stage for deploying this general model framework as a tool for both refining existing models and guiding the development of new ones. It offers a strategic blueprint for constructing models that are not only scientifically robust but also clinically relevant, providing insights that can lead to tailored pain management strategies.

4.1 Multiple dimensions of pain, many models

Previous chapters have addressed the multiple dimensions of pain phenomenon. Constraints have been discussed at the philosophical, psychological and neural level that provides a complex picture of the pain concept. As proposed by Williams and Craig (2016) pain is "a distressing experience associated with actual or potential tissue damage with sensory, emotional, cognitive and social components". Such complexity is indeed and eventually reflected in the Revised IASP Definition of Pain (2020):

An unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage.

Summing up the different aspects previously touched (Raja et al., 2020):

- Pain and nociception are different phenomena. Pain cannot be inferred solely from activity in sensory neurons.
- Pain is always a personal experience that is influenced to varying degrees by biological, psychological, and social factors.
- Through their life experiences, individuals learn the concept of pain.
- A person's report of an experience as pain should be respected.
- Verbal description is only one of several behaviours to express pain; inability to communicate does not negate the possibility that a human or a non-human animal experiences pain.
- Although pain usually serves an adaptive role, it may have adverse effects on function and social and psychological well-being.

The main result here is that pain is a multidimensional phenomenon, thus

- It should be assessed (not measured) using multiple dimensions:
  - **Measurement** refers to the process of quantifying physical properties in terms of a standard unit or fixed amount, typically using an instrument or container marked off in units. This approach is useful for obtaining precise data on observable phenomena.
  - **Assessment**, on the other hand, involves a comprehensive evaluation that goes beyond mere measurement to critically analyse and definitively judge the nature, significance, status, or merit of a condition. When applied to pain, assessment entails a holistic appraisal of the pain experience, encompassing not just the intensity but also the personal and subjective aspects of the experience. Rather than rigorous evaluation in a bias-free manner, it connotes a more generalised judgment that is made for clinical purposes (McGuire, 1992).
- It should be modelled as a multiple component system: Fig. 4.1 sketches the different components, at different analysis level (social, conceptual and perceptual).

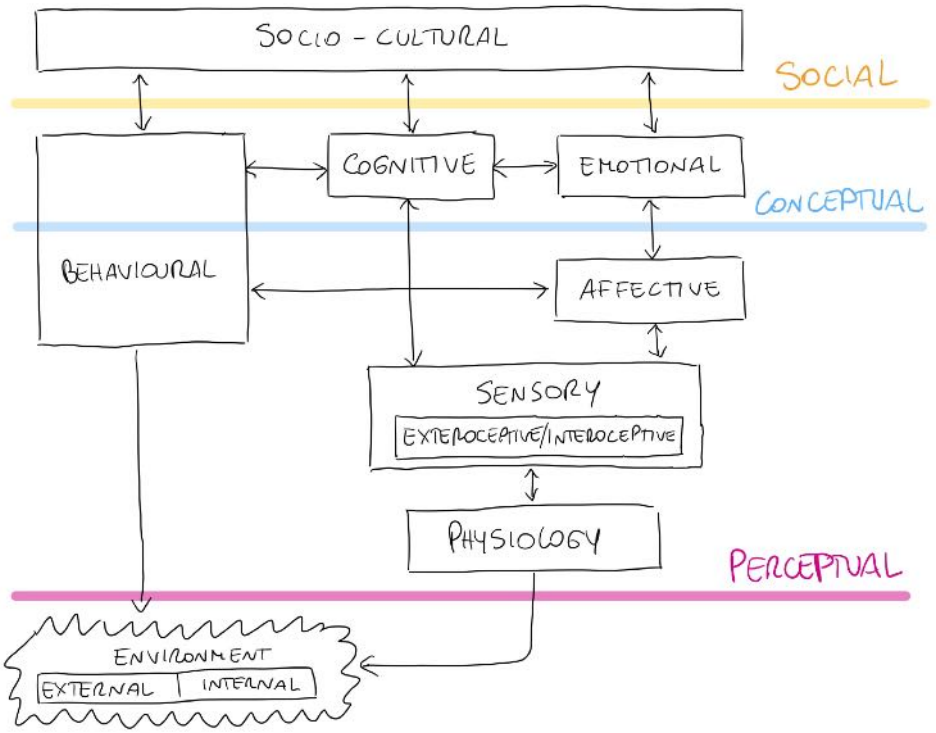


Figure 4.1: *This functional representation is best conceived as that of a state space defined by an ensemble of components dynamically changing in time, at different time scales, the "rate of change" decreasing from bottom to top representation levels.*

Note that the two above are strictly intertwined, since any model, beyond its explicatory and predictive capabilities, is a measuring device (Gerlee and Lundh, 2016). While theories are grand statements about the constitution of the world, models, in order to act as "measuring devices" have to be carefully aligned with their purpose in order to that tell us something about a specific part of reality (Gerlee and Lundh, 2016).

4.2 Computational modelling: A conceptual overview of the literature

Due to the complexity of pain it is not easy to provide a comprehensive and straightforward review of the existing literature that concerns computational models. There are many reasons for such circumstances.

First, there is no single prevailing theory of pain; meanwhile, pain can be studied at different scales, such as the biochemical/cellular, neural, functional brain areas and psychological scales (Lang et al., 2021; Britton and Skevington, 1996).

Hence, currently, there is an overwhelming literature that has been produced in the foundational neurobiological and psychological/behavioural realms, which feeds on such complexity. Such circumstances are depicted at a glance in Figure 4.2.

The variety of neurobiological and psychological models offers the conceptual basement (Gerlee and Lundh, 2016) upon which computational models can be conceived. Here, the idea of conceptual basement is borrowed (and freely adapted to our purposes) from (Gerlee and Lundh, 2016) "house of models" idea, illustrated in Figure 4.3.

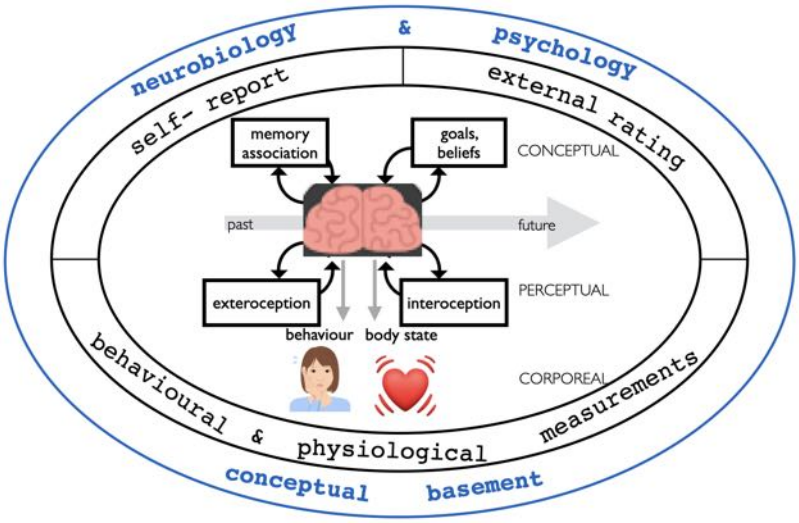


Figure 4.2: *The "agent-in-pain" involving the different analysis levels outlined in Fig. 4.1. Neurobiological and psychological models, as gauging devices must rely on subjective reports, third-person ratings and behavioural and physiological measurements, the experience of pain being a subjective experience.*

In principle, computational models are invaluable tools also from a clinical standpoint. They allow for examining the full spectrum of conditions and combinations of parameters can be studied faster than with classic clinical studies, which cannot test exhaustively (Gerlee and Lundh, 2016). In general, computational models of pain can be designed such that they may reveal undiscovered pain characteristics, explain clinical observations, or challenge current "truths" about pain mechanisms.

In spite of these possible advantages, recent literature surveys indicate that modelling has not been a tool hitherto widely employed (nor accepted) in the study of pain. This is notwithstanding its proven utility in helping to understand complex biological processes, such as in exploring the neural dynamics of vision and in predicting the spread of infectious diseases (Gerlee and Lundh, 2016).

On the other hand statistical and Bayesian approaches, in particular, have gained currency in formalising, at the theoretical level, the conceptual basement of pain (Tabor et al., 2017). We argue that a possible taxonomy of computational models could be clustered as in Fig. 4.4 relying on the prediction/explanation dimensions.

Roughly, we might consider the following.

- *Mathematical models*. These are mostly symbolic and analogue in Gerlee and Lundh's terms, and mostly aim at explanation. By and large they might exploit differential equations to address lower description scales of pain (Britton and Skevington, 1996). One paradigmatic example is Britton and Skevington's work to model Melzack's gate control theory of pain (Britton and Skevington, 1989), by simulating acute pain for a single transmission unit via partial differential equations (PDEs) based on the Wilson-Cowan model for synaptically coupled neuronal networks (Wilson and Cowan, 1972).

- *Predictive, black-box models*. In Gerlee and Lundh's terms these are functional,

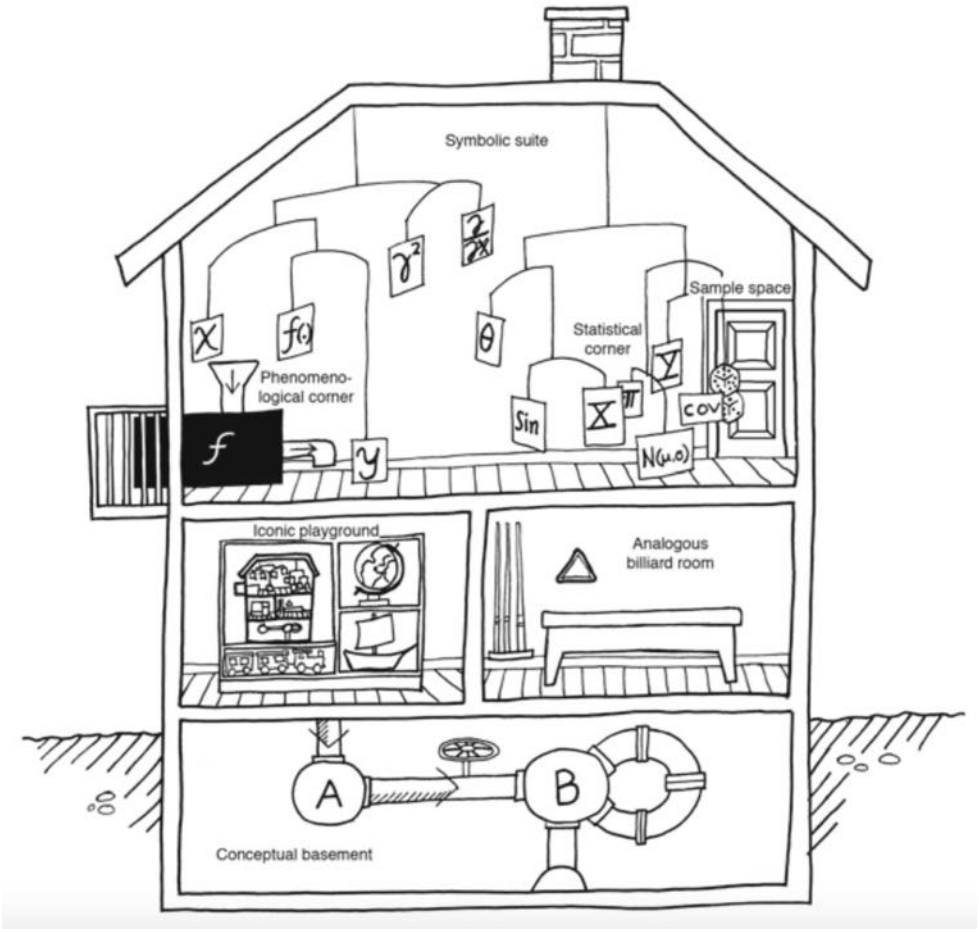


Figure 4.3: *Gerlee and Lundh "House of models" idea.* ***Conceptual models*** *are the most basic type and serve as the foundation for more concrete and mathematical models; they manifest themselves as ideas and notions about mechanisms and entities within a system.* ***Iconic models*** *are direct representations of a system; atoms being modelled as plastic spheres that can illustrate how molecules are structured and interact are one example.* ***Analogous models*** *are not classified by the process of construction devised in analogy with a known system; one striking example is the model of atom describing the orbits of electrons in analogy with the planets orbiting the sun.* ***Symbolic models*** *make use of symbols and formal systems in order to describe phenomena; in physics partial differential equations, which describe how continuous quantities such as a magnetic field or wind speed changes in space and time are but one example.* ***Phenomenological models*** *are often symbolic in nature but are used when the end result is prioritised and capturing the actual internal mechanisms within the system is viewed as less important; Such models, where only the outcome is of importance, are often viewed as "black boxes"; most deep-learning based model might be ascribed to this family.* ***Statistical models*** *are a subset of symbolic models that make use of tools from probability theory (random variables and distributions); with the aid of statistical models it is possible, by observing and analysing data from a phenomenon, to determine which interactions are important and what variables are relevant. Adapted from Gerlee and Lundh (2016).*

phenomenological and data-driven models that aim at identify pain parameter-

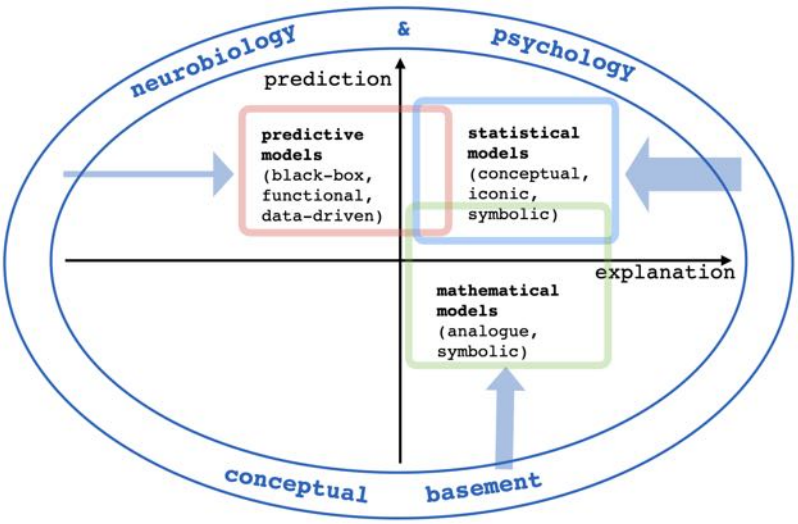


Figure 4.4: *A qualitative taxonomy of computational models of pain. Size of shaded arrows is proportional to the capability of each family of models to account for and accomodate different levels of neurobiological and psychological conceptual explanation.*

s/features that are crucial to the presence or absence of pain (classifying or regressing levels of pain) (Lang et al., 2021). This is an area which is best known as automatic "Pain Recognition", strictly related to the affective computing research realm, and which has recently benefitted from the machine learning and deep learning advancements. An in-depth overview can be found in Werner et al. (2022).

- *Statistical models*. These are conceptual/symbolic approaches that address the computational problem of pain at different explanation levels (in the sense of Marr and Vaina (1982)). Typically such models have addressed both the agent's inferential problem via Bayesian approaches and agent's learning via Reinforcement Learning approaches. In general terms, pain is addressed in the framework of the perception-action cycle common to all goal-directed behaviours. Such models have the potential to deepen the understanding of pain processes at different levels but also generate new predictions to validate current and future experiments. Further, these models when implemented can take advantage of computational techniques developed in current advanced machine learning. A principled discussion of this area of research is given by Chen and Wang (2023); Seymour and Mancini (2020).

Clearly, the above is a crude taxonomy, and there might obviously be overlaps between the three areas. Yet, it has the merit to shed some clear light on a really messy landscape of studies.

In the remainder of this dissertation, we shall be mostly concerned with statistical theories and models, the latter having the potential to not only improve on both prediction and explanation dimensions, but also hierarchically account for a plurality of pain analysis levels from the bottom/physiological to the upper/social as we will show.

Eventually, statistical and markedly Bayesian model offer a natural framework for

addressing the epistemological issues famously raised by David Marr and that have been deemed cogent for pain too (Chen and Wang, 2023; Seymour and Mancini, 2020).

This latter aspect deserves a specific interlude before delving into modelling details.

4.3 Epistemological interlude: levels of explanation

Within the realms of cognitive and behavioural sciences, computational models serve dual purposes. They can act as instruments to interpret empirical data or as embodiments of cognitive theories, as discussed by Palminteri et al. (2017). The emphasis of this research aligns more with the latter, delving into the realm of pain's computational representation.

It's imperative to appreciate the granularity with which these models can interpret pain. Palminteri et al. (2017) highlight two primary model types: aggregate and mechanistic. While aggregate models offer overarching behavioural patterns using mathematical frameworks, mechanistic models delve into the underlying processes that generate these behaviours.

Building on this foundation, Marr and Vaina (1982) introduced a triad of descriptive layers:

- the *what/why* facet, aligning with computational theories and identifying models for specific behavioural phenomena;
- the *how* facet, synonymous with algorithmic processes;
- the realisation aspect, focusing on the tangible implementation.

Marr's tripartite distinction has become emblematic in cognitive science research, as supported by Dennett (1987), facilitating rigorous philosophical discourse. For the purpose of this thesis, Marr's model suggests that complex systems, like living organisms experiencing pain, demand multifaceted explanations encompassing psychological, neurological, cellular, and biochemical facets.

Anderson (1991) further emphasised the dichotomy between computational theories, which he termed "rational explanations", and algorithmic perspectives, which he labelled "mechanistic explanations". A primary challenge lies in determining the constraints of computational theory to minimise the arbitrariness of cognitive models at the algorithmic level. This problem of model underdetermination, although a broad challenge in scientific explanation, has a heightened impact on cognitive exploration.

The transition from classical to embodied cognitive science brought in additional dimensions of environment and body-based constraints. The challenge of grounding different levels into a cohesive relationship remains an ongoing research conundrum, as discussed by Boccignone and Cordeschi (2015). Even Marr grappled with the ambiguity between algorithmic and implementation aspects.

Recent adaptations, considering the rise of Bayesian methods in cognitive sciences, propose a distilled two-tier hierarchy: the computational theory, grounded in Bayesian logic, and an implementation theory that consolidates Marr's latter two levels.

This thesis gravitates towards this dual-level perspective. The terms "model" and "theory" are interchangeably used to denote these levels: the theoretical model and

the implementation model. At each stratum, the cognitive scientist is crafting models that incorporate constraints rooted in both biological laws and theoretical hypotheses, aiming to explain specific phenomena—like the multifaceted experience of pain.

4.4 The Bayesian roadmap

Multilevel analysis provides a structured approach to navigating the multifaceted nature of pain, which is intricately woven with neurobiological, psychological, and social dimensions, as highlighted in the biopsychosocial model (Hadjistavropoulos et al., 2011). Pain, akin to cognitive phenomena such as emotions, spans various temporal and spatial dimensions.

Through a Bayesian lens applied to Marr's multilevel framework, researchers are equipped with a refined methodology to dissect the complexities of pain. Distilling the key points:

- The concept of architecture is pivotal. It represents the set parameters or constraints that researchers operate within, tailored to specific levels of examination. In the context of pain, this can relate to neurobiological pathways, psychological processes, and the social dynamics surrounding pain.
- At the base lies the neurobiological component, offering a foundational blueprint. It's poised to articulate the broader computational theory, branching into specific realms from individual cellular activities to expansive neural networks.
- Marr's algorithmic approach, when adapted to pain, isn't an isolated level of understanding. It bridges simulations across spectrums: from overarching simulations encapsulating pain's psychological and social interactions, right down to the nuanced neural underpinnings.

The top-down perspective ensures cohesive integration across these tiers, where broader dimensions like psychological understandings and social contexts shape and are influenced by the foundational neurobiological mechanisms. On the flip side, the bottom-up approach shines a light on the emergent intricacies of pain, deriving from foundational neurobiological patterns. These base patterns are the springboards from which the broader psychological and social aspects of pain emerge.

To make this discussion concrete, the functional architecture presented in Figure 4.1, which in turn has been shaped from the underlying structures devised at the neurobiological level, can be iconically abstracted in the concept-driven perception-action cycle shaped in the form of a dynamic Probabilistic Graphical Model presented Figure 4.5.

More precisely, assume that the world's state at time t, $\mathcal{W}_t$ comprises both the external environment and the agent's body *internal milieu* and denote the following.

- $\mathcal{C}_t$: the ensemble of conceptual state variables and goals. Here, we use the abstract term "concept" to refer to any entity, idea or abstract noun that we can think, reason or talk about. A concept is something that has specific properties, truths and beliefs relating to that concept. We can refer to the name of the concept as the concept label. More specifically, in the context of pain and emotion theory the

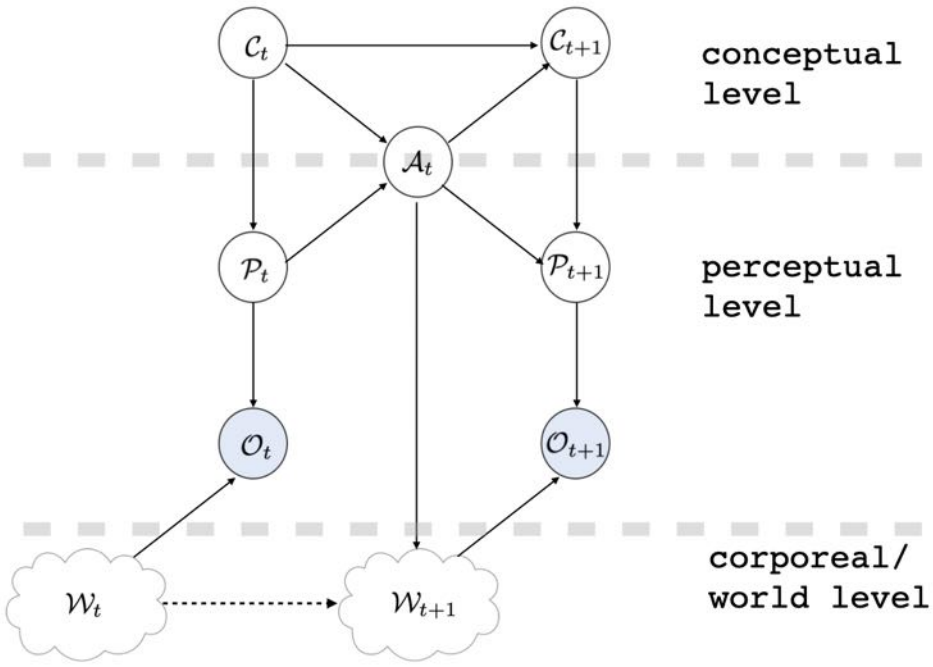


Figure 4.5: *A first essential model in the shape of a dynamic Probabilistic Graphical Model $\mathcal{GM}$. Arrows denote conditional probability dependencies between the ensemble of random variables. The world ensemble $\mathcal{W}_t$ and its process dynamics is a kind of special ensemble since it can only be accessed through observations $\mathcal{O}_t$. Dependencies between ensembles indexed by $t \rightarrow t+1$ transition represent model dynamics. This can be also considered a time-slice representation of the inferential dynamics allowed by the model.*

term concept has been defined by Barrett (2006). Thus, for instance, at a higher cognitive level the term "pain" labels a concept much like the term "fear" does.

- $\mathcal{P}_t$: the ensemble of perceptual state variables.
- $\mathcal{O}_t$: the ensemble of sensation/observation state variables. These variables provide the only access to the "true" state of the world at time t, $\mathcal{W}_t$
- $\mathcal{A}_t$: the ensemble of controls, actions, and sequences of actions, namely policies $\pi_t = a_1, a_2, \cdots$. At the lowest level, actions impinge either on the external or internal milieu components of the world $\mathcal{W}_t$ and contribute to the world dynamics $\mathcal{W}_t \rightarrow \mathcal{W}_{t+1}$

The PGM shown in Figure 4.5. addresses a very high-level abstract representation of our problem. As we will see, the abstract structure needs to be further detailed in terms of the definition of the random ensembles introduced in order to operationalise it. Indeed, PGM-based systems are suitable to provide model components that hide underlying complexity. In general they are able to express many components of a theoretical model, yet they lack expressivity as to the core of such models, concerning the stochastic processes behind the structure and anything dependent on those processes and their dynamics. Further, parts/components of the overall structure might evolve along a possible simulation so that their structure might not constrain to a fixed topology. One

example, in our case, could be the unfolding of the action planning component/ensemble. Thus, it is sometimes necessary, for actual problems, to describe the model as an unbounded stochastic loop or recursion over potential PGMs. Also, notwithstanding the time-slice representation the inferential dynamics is not readily apparent.

These expressivity problems can be solved using universal probabilistic programming languages (PPLs), a kind of modelling approach that has a long history in Computer Science, but which has gained currency in recent years (Goodman, 2013; Ghahramani, 2015), see van de Meent et al. (2018) for an in-depth introduction.

The basic idea in probabilistic programming is to use computer programs to represent probabilistic models. One way to do this is for the computer program to define a generator for data from the probabilistic model, that is, a simulator. This simulator makes calls to a random number generator in such a way that repeated runs from the simulator would sample different possible data sets from the model. This simulation framework is more general than the PGM framework since computer programs can allow constructs such as recursion (functions calling themselves) and control flow statements (for example, "if" statements that result in multiple paths a program can follow), which are difficult or impossible to represent in a finite graph (Ghahramani, 2015).

Thus, on the one hand, a "universal PPL" (UPPL) which is generally defined as an extension of a Turing-complete general-purpose language, can express models with an unbounded number of random variables. This means that random variables are not fixed statically in the model (as they are in a finite PGM) but can be created dynamically during execution. On the other hand, due to recent and exciting advancements in this research field, concrete PPLs have been specified that can rely on sophisticated algorithms and tools developed in the machine learning community based on more recent advancements in Markov Chain Monte Carlo (MCMC) and variational inference techniques. These efforts have produced powerful PPL platforms that can "compile" a theoretical probabilistic model into an implementation model suitable to work in the real world (Bingham et al., 2019; Tran et al., 2016; Salvatier et al., 2016).

In brief, like probabilistic graphical modelling, PP allows one to capture abstract, conceptual knowledge as generative models. Instead of a graphical representation, PP represents conceptual knowledge as stochastic programs—chunks of code that embed randomness into their execution. The core idea is representing a probabilistic model as specified through a $\mathcal{GM}$ in terms of probabilistic programs. Thus, unlike deterministic programs that always produce the same output when given the same input, probabilistic programs instead produce samples from a distribution of possible outputs. This allows explicit modelling of uncertainty, whether such uncertainty arises from (i) incomplete knowledge about the world and agents' unobservable mental states, (ii) incomplete theory, or (iii) inherent randomness in the generative process.

UPPLs (UPPLs) solve the expressivity problem by providing additional expressive power over PGMs. A PPL model description is essentially a simulation program (or generative model). Each time the program runs, it generates a different outcome. Theoretically, if it is executed an infinite number of times, we obtain a probability distribution over outcomes. Probabilistic programs are usual functional or imperative programs with two added constructs: (1) the ability to draw values at random from distributions, and (2) the ability to condition values of variables via observation.

Conceptually, conditioning needs to compute input states of the program that generate data matching the observed data. Canonical programs are conceived to run from inputs to outputs, conditioning involves solving the inverse problem of inferring the inputs (in particular the random number calls) that match a certain program output. Such conditioning is performed by a "universal inference engine", usually implemented by Monte Carlo sampling over possible executions of the simulator program that are consistent with the observed data.

Thus, a UPPL provides two special constructs, one for drawing a random variable from a probability distribution, e.g., "$\sim$" and one for conditioning a random variable on observed data, say "OBSERVE". The former is a way to define $P(Z, Y)$ and the latter is the same as standard Bayesian conditioning $P(Z|Y)$. These special constructs are used by the PPL inference algorithms to manipulate executions of the program during inference. Many PPLs are embedded in existing programming languages, with these two special constructs added.

Below, for simplicity, we use a simple, abstract PPL-like specification of the model. This will suffice for the current purposes.

The generative/predictive dynamics of the agent based on the $\mathcal{GM}$ unfolds as follows (see Algorithm 1).

Algorithm 1 Simulation-based one-step dynamics

Input: Agent's state $\mathcal{S}_t = (\mathcal{C}_t, \mathcal{P}_t)$ and related state distribution (prior); current observed outcome $\mathcal{O}_{t+1}$; the previous action $\mathcal{A}_t$ and its distribution (evidence)

Output: Agent's state $\mathcal{S}_{t+1}$ and next action $\mathcal{A}_{t+1}$ with updated state distribution (posterior)

Conceptual sampling:
$\mathcal{C}_{t+1} \sim P(\mathcal{C}_{t+1} \mid \mathcal{C}_t, \mathcal{A}_t)$ ▷ conceptual belief update

Perceptual sampling:
$\mathcal{P}_{t+1} \sim P(\mathcal{P}_{t+1} \mid \mathcal{A}_t, \mathcal{C}_{t+1})$ ▷ exteroceptive/interoceptive sampling

Action/plan sampling:
$\mathcal{A}_{t+1} \sim P(\mathcal{A}_{t+1} \mid \mathcal{C}_{t+1}, \mathcal{P}_t)$ ▷ external / internal action sampling

Observation: $\texttt{OBSERVE}(P(\mathcal{O}_{t+1} \mid \mathcal{W}_{t+1}, \mathcal{P}_{t+1}) : \mathcal{O}_{t+1})$ ▷ sensing the state world

The above model is a general one and can further specified in its components. To this aim the basic principles to be considered are the following.

1. **Principle of Hierarchy**. Each ensemble can be defined/refined at multiple levels ℓ (Mesulam, 2008). For example, in a certain modelling context, proprioception might be defined as a two-level perceptual component:

   - $\mathcal{P}^{(\ell)}$: proprioception of body movement;
   - $\mathcal{P}^{(\ell-1)}$: proprioception of head or eye movements;

   This obviously entails that in the most general case one could consequently deal with proprioceptive "observations" at different levels, e.g. $\mathcal{O}^{(\ell)}$ and $\mathcal{O}^{(\ell-1)}$.

   A corollary of this principle is that Different levels might operate at different time scales, resulting in a level-dependent time or clock $t^\ell \in [1 \cdots T^\ell]$. For example, in robotics one might differentiate robot movement occurring at:

   - $\ell = 3$, t^3 unfolding at 0.5 Hz for decision making;

- $\ell = 2$, t^2 unfolding at 25 Hz for limb stability and control;
- $\ell = 1$, t^1 unfolding at 500 Hz for joint level control;

A representation of such circumstances is sketched in Figure 4.6

2. **Principle of Heterarchy**. At any level ℓ, one might define/differentiate any ensemble via different components or factors. For example, the general perceptual ensemble can be differentiated in its exteroceptive and interoceptive components: $\mathcal{P}^{(\ell)} = (\mathcal{P}^{(\ell,ext)}, \mathcal{P}^{(\ell,int)})$ A heterarchy permits different factors at the same level to cooperate whilst individually optimising different success criteria (McCulloch, 1945).

Indeed, a heterarchy may be orthogonal to a hierarchy, subsumed to a hierarchy, or it may contain hierarchies; the two kinds of structure are not mutually exclusive. In fact, each level in a hierarchical system is composed of a potentially heterarchical group which contains its constituent elements Mesulam (2008).

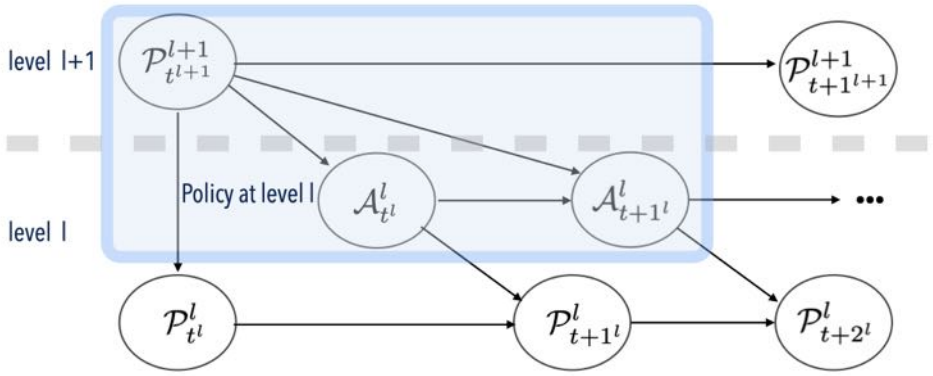


Figure 4.6: *Considering different time scales, for instance, at the perceptual level The shaded box collects the RVs from which a policy* $\pi^l_t = a^l_{t^l} a^l_{t+1^l} \cdots a^l_{T^l} \mid P^{l+1}_{t^{l+1}}$ *can be instantiated at level* l *depending on a perceptual state* $P^{l+1}_{t^{l+1}}$ *at level* $l+1$*. The mechanism can be replicated* mutatis mutandis *at the conceptual level.*

Under such circumstances, a general model can be formalised as in Figure 4.7, where the generic ensemble $\mathcal{S}^{(\ell)}_{t^{(\ell)}}$ stands for the state of any collection of ensembles of random variables (conceptual, perceptual, etc) at level ℓ, whose dynamics is represented by the random process $\mathcal{S}^{(\ell)} = \{\mathcal{S}^{(\ell)}_{t^{(\ell)}} : t^{(\ell)} \in \left[t^{(\ell)}_1, t^{(\ell)}_2\right]\}$ indexed over the subset of reals $[t^{(\ell)}_1, t^{(\ell)}_2]$.

The generative model corresponding to the PGM in Figure 4.7 reads:

$$P(\mathcal{O}, \mathcal{S}, \mathcal{A}) = \prod_{\ell} P(\mathcal{O}^{(\ell)} \mid \mathcal{S}^{(\ell)}) P(\mathcal{S}^{(\ell)} \mid \mathcal{S}^{(\ell+1)}, \mathcal{A}^{(\ell)}) P(\mathcal{A}^{(\ell)} \mid \mathcal{S}^{(\ell+1)}) \tag{4.1}$$

Here without loss of generality we have introduced two simplifying assumptions.

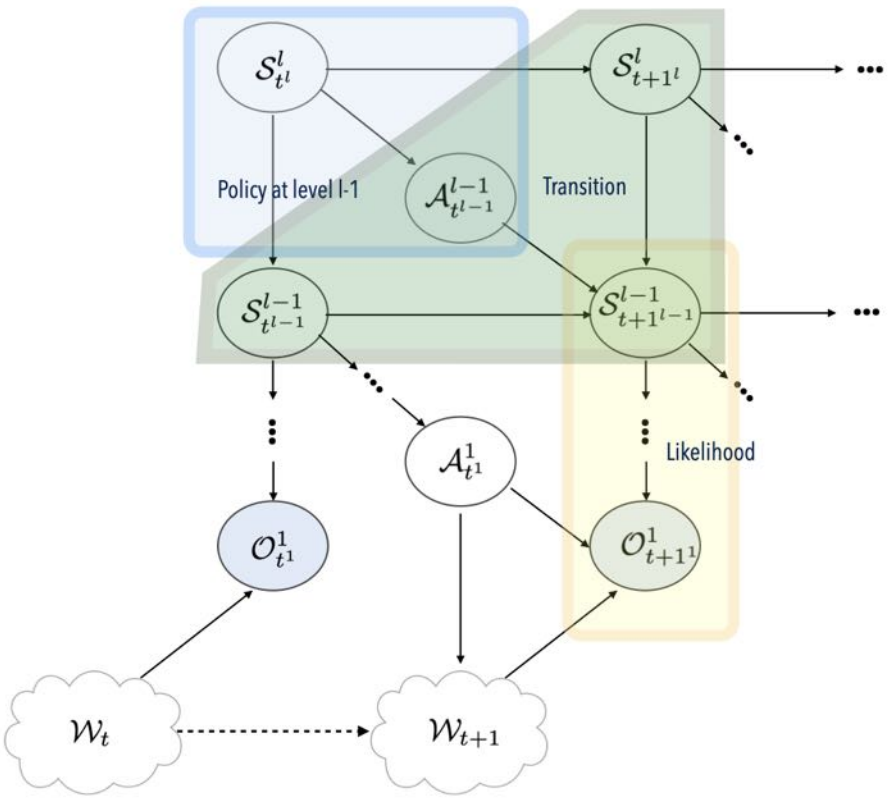


Figure 4.7: *Generalising the model in Fig. 4.5. For simplicity, the observation level has been only represented at the bottom of the hierarchy to gauge world dynamics, thus omitting possible observational states* $\mathcal{O}^l_{t^l}, l > 1$

Our first simplification is to discretise the timeline into a set of time slices, that are measurements/assessment of the system state taken at intervals that are regularly spaced with a predetermined time granularity Δ. Thus, we can restrict our set of RVs for example to $\mathcal{X}_0, \mathcal{X}_1, \cdots$, 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 RVs to representing distributions over countably many RVs, sampled at discrete intervals.

Under such assumption, consider a distribution $P(\mathcal{X}_{0:T})$ 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:

$$P(\mathcal{X}_{0:T}) = P(\mathcal{X}_0) \prod_{t=0}^{T-1} P(\mathcal{X}_{t+1} \mid \mathcal{X}_{0:t}) \tag{4.2}$$

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. A dynamic system over the variable $\mathcal{X}$ satisfies the Markov assumption if, $\forall t \geq 0$,

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

The Markov assumption allows us to define a more compact representation of a state space distribution as:

$$P(\mathcal{X}_{0:T}) = P(\mathcal{X}_0) \prod_{t=0}^{T-1} P(\mathcal{X}_{t+1} \mid \mathcal{X}_t). \tag{4.3}$$

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

For instance, under such assumption for an $L = 3$ level PGM, Eq. 4.7 would write:

$$P(\mathcal{O}^{(1)}_{0:T^{(1)}}, \mathcal{S}^{(1:L)}_{0:T^{(l)}}, \mathcal{A}^{(1:L)}_{0:T^{(l)}}) = \prod_{\ell=1}^{L} P(\mathcal{S}^{(\ell)}_0) P(\mathcal{A}^{(\ell)}_0) \tag{4.4}$$

$$\prod_{t^{\ell=1}}^{T^{\ell=1}} \prod_{t^{\ell=2}}^{T^{\ell=2}} \prod_{t^{\ell=3}}^{T^{\ell=3}} P(\mathcal{A}^{(\ell=2)}_{t^{(\ell=2)}} \mid \mathcal{S}^{(\ell=3)}_{t^{(\ell=3)}}) P(\mathcal{S}^{(\ell=3)}_{t^{(\ell=3)}} \mid \mathcal{S}^{(\ell=3)}_{t^{(\ell=3)}-1}) \tag{4.5}$$

$$P(\mathcal{A}^{(\ell=1)}_{t^{(\ell=1)}} \mid \mathcal{S}^{(\ell=2)}_{t^{(\ell=2)}}) P(\mathcal{S}^{(\ell=2)}_{t^{(\ell=2)}} \mid \mathcal{S}^{(\ell=2)}_{t^{(\ell=2)}-1}, \mathcal{S}^{(\ell=3)}_{t^{(\ell=3)}}, \mathcal{A}^{(\ell=2)}_{t^{(\ell=2)}-1}) \tag{4.6}$$

$$P(\mathcal{O}^{(\ell=1)}_{t^{(\ell=1)}} \mid \mathcal{S}^{(\ell=1)}_{t^{(\ell=1)}}) P(\mathcal{S}^{(\ell=1)}_{t^{(\ell=1)}} \mid \mathcal{S}^{(\ell=1)}_{t^{(\ell=1)}-1}, \mathcal{S}^{(\ell=2)}_{t^{(\ell=2)}}, \mathcal{A}^{(\ell=1)}_{t^{(\ell=1)}-1}) \tag{4.7}$$

Each "possible world" in the probability space defining the agent is then a *trajectory*, namely, an assignment of values to each RV of interest, collected in $\mathcal{S}$, for each relevant time t.

4.5 Back to the literature: a principled taxonomy for statistical models of pain

In the previous section we have devised a general hierachical concept-driven perception/action model that, by constraining levels (hierarchy) and factors (heterarchy) at the different levels, can frame and instantiate different pain models.

The major advantage of such step lies in that now we can afford to distinguish and characterise in a principled way the complexity and expressiveness of the wide number of statistical/probabilistic models of pain that have been proposed so far in the literature. On this basis, we can establish a "gradient" of models based on their probabilistic structure and the information they encode.

In other terms, the hierarchical/heterarchical nature of generative models allows for the description/grouping of proposed pain models of increasing model complexity.

For instance, we might minimally aggregate agent-in-pain models according to the following classes.

- *Class 0 models*. The lowest complexity is a simple, single-time-point model of pain perception.
- *Class 1 models*. More complex pain perceptual models can include anticipation of future observations, that is they account for dynamics.

- *Class 2 models*. Complexity increases when a model incorporates action selection and must therefore anticipate the observed consequences of different possible actions (plans or policies).

  At this level, we can include all statistical methods that involve pain in learning issues, such as in the well known fear-generalisation task.

- *Class 3 models*. At the highest complexity level, we can consider pain models that allow for agent communication, e.g., the sufferer and the caregiver, thus accounting for the social dimension of pain.

It goes without saying that each class, in turn, can provide lodging to pain models of different complexity in terms of hierarchy depth L and heterarchy dimension (cardinality of factors $|F|$) at level (ℓ).

In the following, we provide some examples to illustrate the above taxonomy.

4.5.1 The Class 0 model baseline

The baseline model provides a crude operationalisation of the idea that sensory processing of pain signals in the brain might comprise some sort of statistical inference about the cause of incoming nociceptive information, cfr. Seymour and Mancini (2020) for a discussion. This is a longstanding idea, following parallels with other sensory domains such as vision and audition.

Precisely, consider a simple 1 level / 1 factor perceptual model where $\mathcal{S}, \mathcal{O}$ boil down to the following:

- $\mathcal{S} = \mathcal{P} = P$: the RV denoting the generic pain percept
- $\mathcal{O} = O$: a RV denoting the bottom-up nociceptive input from the periphery

The model proposes a mechanism for determining the hidden (latent) causes $P = p$ of encountered sensory information $O = 0$, summarised in the generative model

$$P(O = o, P) = P(P)P(O = o \mid P) \tag{4.8}$$

In Bayesian terms, this is achieved through the weighted integration of prior experience $P(P)$ and current (potentially multisensory) likelihood information $P(O = o \mid P)$, represented using probability distributions that reflect the agent's subjective uncertainty. In Eq. 4.8, $P(P)$ stands for the expectation of pain or physical threat.

Perceptual experience of pain follows from the optimal integration of these probability distributions via Bayes' rule:

$$P(P \mid O = o) = \frac{P(P)P(O = o \mid P)}{P(O = o)} \tag{4.9}$$

Prior $P(P)$ denoting top-down pain expectation is usually assumed to encode prior experience. Thus, in general, how we ultimately perceive pain ("pain belief", $P(P \mid O = o)$, including placebo or nocebo effects as an example) does not only depend on how the nociceptive input, sensory input (such as vision), and proprioceptive input signaling alarms ("likelihood of injury or harm") but also on our attention, expectation, context, and environment ("prior"). Through evolution and experiences, our brains

have built empirical internal predictive mapping models for the likelihood of an event to occur (Chen and Wang, 2023; Seymour and Mancini, 2020; Tabor et al., 2017).

This basic scheme has produced a wide number of works in the field even limiting to the simple instantiation of uncertainties in terms of Gaussian approximations, which is worth recalling here.

Assume a Gaussian or Normal prior $P(P) = \mathcal{N}(\mu_P, \sigma_P)$ and likelihood $P(O \mid P) = \mathcal{N}(\mu_O, \sigma_O)$. Then the posterior expectation of pain is also Gaussian $P(P \mid O) = \mathcal{N}(\mu_{P|O}, \sigma_{P|O})$ with mean and variance written as

$$\mu_{P|O} = \frac{\sigma_O^2}{\sigma_O^2 + \sigma_P^2}\mu_P + \frac{\sigma_P^2}{\sigma_O^2 + \sigma_P^2}\mu_O \tag{4.10}$$

$$\sigma_{P|O}^2 = \frac{\sigma_P^2 \sigma_O^2}{\sigma_O^2 + \sigma_P^2} \tag{4.11}$$

Sophisticated analyses can be conducted even by resorting to this minimal scheme with few variations, see for instance the recent work by Strube et al. (2023).

It is often useful to express variances in terms of precisions $\gamma_O = \frac{1}{\sigma_O^2}, \gamma_P = \frac{1}{\sigma_P^2}$ so that

$$\mu_{P|O} = \frac{\gamma_O}{\gamma_{P|O}}\mu_O + \frac{\gamma_P}{\gamma_{P|O}}\mu_P \tag{4.12}$$

$$\gamma_{P|O} = \gamma_O + \gamma_P \tag{4.13}$$

Alternatively, the posterior mean can be rearranged as an incremental update form from the prior

$$\mu_{P|O} = \mu_P + \frac{\gamma_O}{\gamma_{P|O}}(\mu_O - \mu_P), \tag{4.14}$$

where the relative precision $\frac{\gamma_O}{\gamma_{P|O}}$ can be interpreted as a learning rate and $(\mu_O - \mu_P)$ as a prediction error (PE).

This paves the way for the formal introduction of the predictive coding framework in pain modelling (Wiech, 2016).

With Marr in mind, whereas perception may be approximately Bayes optimal in some cases, it doesn't necessarily mean that the brain is implementing Bayesian inference directly. One reason for questioning the plausibility of Bayesian models at an algorithmic level is that representing full probability distributions across a neural population is unrealistic unless the distribution is simple, which is rarely likely to be the case. This has led to algorithmic approaches to perceptual inference that approximate inference. One such approach is Predictive Coding (PC) (Rao and Ballard, 1999; Friston and Kiebel, 2009; Aitchison and Lengyel, 2017).

In PC models, inference occurs on a hierarchy in which higher level features are inferred on the basis of lower level features. A particular level encodes a prior expectation of what its input will be and sends this information to the level below. This lower level compares this prediction with the observed features and computes a prediction error (PE); the prediction error is sent back to the higher level and can be used to update the prior expectations to minimise surprise. The amount of belief updating is determined

by the relative precision of prior and incoming information, allowing information to be weighted by its uncertainty.

In the case of pain (Wiech, 2016), as shown in Figure 4.8, sensory input or pain-related cues trigger a pain-related expectation. Subsequently, nociceptive input is compared to the expectation that reflects prior information. If incoming information is in line with prior assumptions, the expectation is confirmed. If they diverge, a prediction error signal is generated and the expectation is updated through a learning rule. Note that the generation of a prediction error might not necessarily lead to a revision of the expectation; following up on prediction errors might selectively be impaired in a pathological state and contribute to aberrant learning in the context of pain.

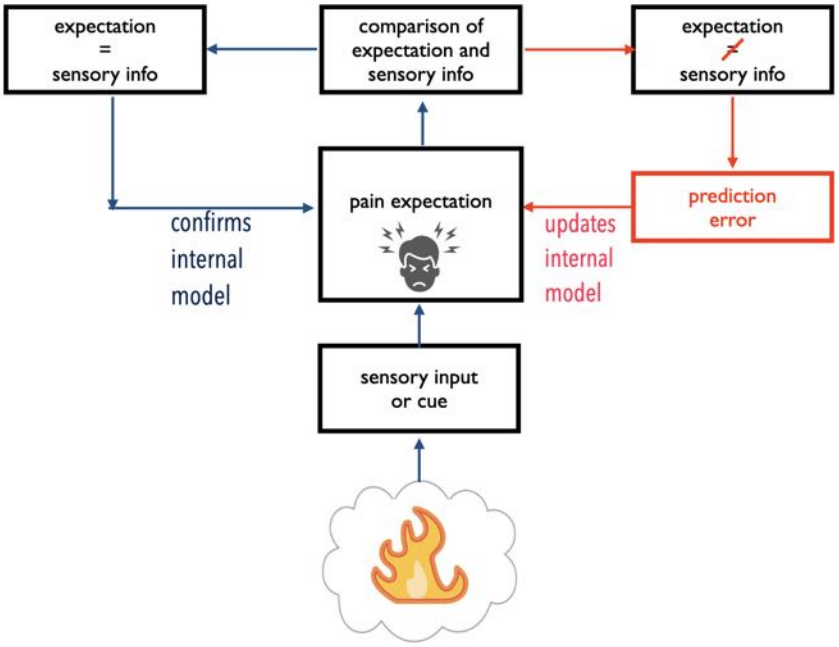


Figure 4.8: *Prediction error processing and learning in the context of pain: a schematic overview. Adapted from Wiech (2016).*

Note that PE computation in terms of the difference written as in Eq. 4.14, is a specific case of a more general approach known as Variational Bayes.

In fact when the Normal assumption is not met, then an approximating distribution $Q(P \mid O) \approx P(P \mid O)$ can be designed and learnt, by lower bounding the log-evidence or *negative surprise* $\log P(O)$ via the Variational Free Energy (VFE) $\mathcal{F}[Q]$

$$\log P(O) = \log \int_P P(P, O) \frac{Q(P \mid O)}{Q(P \mid O)} dP \geq$$
$$\int Q(P \mid O) \log \frac{P(P, O)}{Q(P \mid O)} dP = \mathcal{F}[Q] \tag{4.15}$$

This is the key concept to formal Predictive Processing and Active Inference approaches that have currently gained traction in theoretical neurobiology and neuropsychology (Friston and Kiebel, 2009).

Further applications of the Class 0 model concern, for instance, extending the observation ensemble $\mathcal{O}$, to multiple observational cues such as nociceptive and visual, i.e., $\mathcal{O} = \{O^{vis}, O^{noc}\}$, see e.g. Tabor et al. (2017).

4.5.2 Accounting for temporal change: Class 1 models

Up to this point Class 0 models have considered a static analysis of the experience of pain and how this maps directly to the mathematical principles of Bayesian inference.

Yet, pain experience does not simply reflect prior expectations and current evidence but also incorporates changing experience over time. The agent-in-pain, when the world is uncertain, needs to continually update their estimations in a changing environment.

As pointed out by Tabor et al. (2017) a large body of research has addressed the impact that prior experiences and motivations have on pain, but there are few studies that study the temporal integration of this information. This state of affairs is likely due to the main difficulty with experimentally testing the effects of temporal integration over relatively short timescales, necessary for the majority of lab experiments, where adaptation effects occur. A stimulus, repeated over long time scales, may be perceived as progressively less painful, an effect that can be related to prediction. Hence, to properly explore the effects of temporal integration in a lab setting, we need to minimise this adaptation effect. Switching stimulus locations would be one option, which would assume that pain locations are not static. Using pulsed noxious stimuli is another possibility. In accordance with the principles of temporal integration, previous experiences inform current experiences. If pain is temporally integrated across pulses, then the stimulus to the right arm should be perceived as more painful than the identical stimulus to the left arm. Beyond experimental subtleties, a simple model would be

$$P(\mathcal{O}_{1:T}, \mathcal{S}_{0:T}) = P(\mathcal{S}_0) \prod_{t=1}^{T} P(\mathcal{O}_t \mid \mathcal{S}_t) P(\mathcal{S}_t \mid \mathcal{S}_{t-1}) \tag{4.16}$$

Depending on whether distributions are continuous or discrete (e.g., an HMM model) we have here a sequential Bayesian model that opens here avenues that appear to be relevant for pain research in many ways (Tabor et al., 2017).

One example has been recently set by Eckert et al. (2022).

4.5.3 Putting action into action: Class 2 models

Inferential theories of pain processing (Class 0 and 1) leave open an account of how the motivational function of pain is directed.

Central to Class 2 models is the idea that pain leads to some action, ranging from simple motor reflexes and physiological responses to conscious, deliberative decisions. Such actions should be appropriate to the pain experienced, either preparing for, reducing, or completely avoiding it if possible (Seymour and Mancini, 2020).

Thus, unlike purely inferential or perceptual models, this approach assumes that continuous updating of the agent's estimations should be based on the consequences of its actions within a changing environment. The consequences of these actions, along with potential costs and benefits, vary over time and contribute to the pain experience.

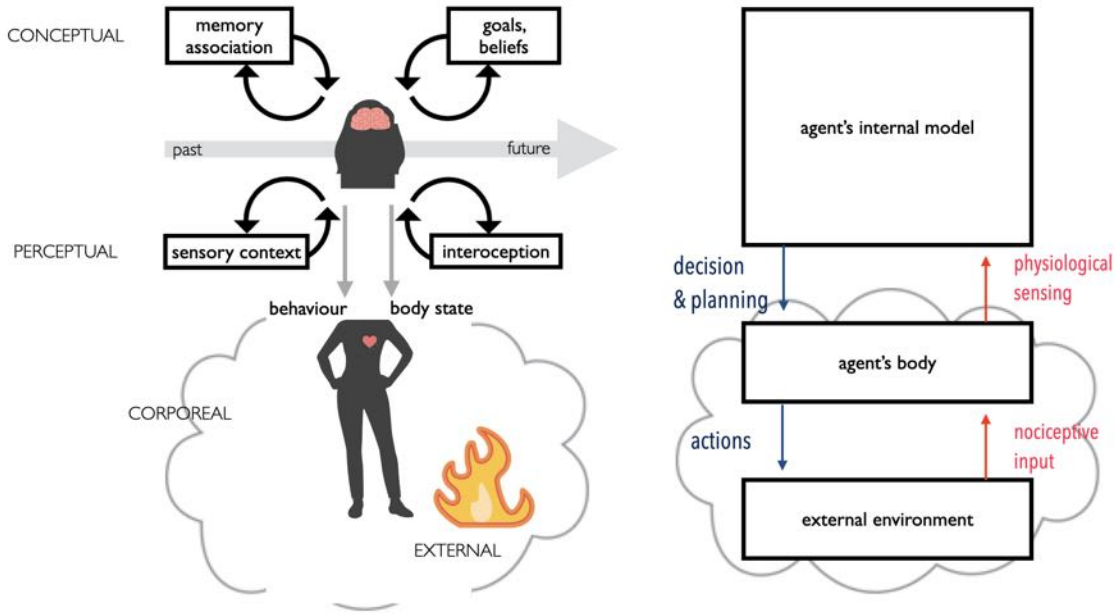


Figure 4.9: *The agent-environment interface. The agent learns values, updates beliefs over states concerning the external environment that contains the sensed objects and selects actions to interact with the environment. Nociceptive input experienced as pain can stand as a negative reward signal to condition the learning process.*

The very fact that pain elicits immediate defensive responses (withdrawal, orientation, etc.) naturally leads to the assumption that as such pain must also guide learning to optimise future responses. In other terms, pain is a kind of (negative) reward from the external environment (see Figure 4.10). More precisely, it can be conceived as an optimal control signal to minimise current and future nociceptive stimuli.

In these terms, models here fall within the domain which is known as Reinforcement Learning (RL, Sutton and Barto (2018)).

However, the idea that harm minimisation is a clearly definable, objectively measurable function and is based on the ability to learn from trial-and-error interaction with the world, finds its roots in early studies of humans and animal learning using Pavlovian (classical) and instrumental (operant) conditioning. In this perspective an early and foundational computational learning model was the Rescorla-Wagner model (R-W) (Rescorla, 1972). The R-W model is usually applied to one-step learning and uses a prediction error term—the difference between what was expected and observed—to update future predictions. In this perspective, RL can be considered an extension of early psychological learning models, such as the R-W model.

Indeed, threat learning has been and still is a central field of research, which we will also address in this dissertation (Chapter 7). However, it is out of scope of this thesis to provide a complete review of RL in pain research; we urge the reader to refer to in-depth overviews by Chen and Wang (2023) and Seymour and Mancini (2020).

Clearly, at this point, the modelling question arises on how Bayesian approaches to perceptual inference of pain can be connected to RL-based methods. Perceptual

inference and action control seem to represent distinct computational processes, both utilising computational architectures that are at least partially hierarchically organised. This leads to the question of how they interact. For instance Seymour (2019) propose a hybrid architecture where the RL control hierarchy has a corresponding sensory processing hierarchy, with crude spinal and brainstem nociceptive input feeding into lower controllers, and conscious pain feeding into higher controllers at the top. The function of sensory processing hierarchies is to allow the optimal estimate of the properties—such as the intensity—of the external stimulus. Based on the assumption that the incoming nociceptive input is inherently noisy, inference will improve the estimate of the true intensity. Computationally, sensory inference is typically proposed to approximate some sort of Bayesian inference (Seymour, 2019).

In a different vein, a way to reconcile the Bayesian perception-as-inference approach and the RL framework is likely to be found by turning back to recent developments in theoretical neuroscience where the predictive brain hypothesis has been formalised into the active inference framework.

Active inference is a mathematical framework that applies the free-energy principle (cfr. Eq. 4.15) to the behavioural norms of biological agents. Since future observations are unobservable until an action is executed, in active inference, decision-making agents are assumed to act to minimise a measure called the expected free energy (EFE). The EFE is composed of value and information gain terms. Thus, by minimising the EFE, it is possible to naturally balance the tradeoff between exploitation and exploration in a principled manner. Additionally, active inference can easily incorporate prior information about the probability distribution of outcomes that agents desire to observe, which is called a prior preference (a.k.a "evolutionary prior"). By adjusting this distribution, agent behaviour can be varied. Adjustment of the prior preference values corresponds to a type of reward tuning required in typical sequential decision-making problems. Yet, specifying the probability of observed outcomes desired by the agent arguably provides a theoretically intuitive and elegant alternative to classic RL to specify numerical reward values over intermediate actions and states.

4.5.4 Class 3 models: the unexplored challenge of modelling social pain

In the minimal case of a sufferer (S) / caregiver (C) dyadic interaction, S and C interact to reliably infer each other's mental states, the beliefs of S ('my concept/experience of pain') generate S's observable actions ('my behaviour'). The actions of S, in turn, cause the (attended) sensory states of C. Attention directed towards S by C in turn enables the observations generated by S to entrain the hidden states of C ('your version of my concept/experience of pain'). This is just some hypothesis entertained by C about the causes of C's observations (i.e., about the mental states generating S's actions). To increase or maintain the reliability of C's hypothesis, C must then act on the niche ('your behaviour') to test C's hypothesis about hidden causes, as it were (that is, to check for mutual understanding, for instance). C thereby causes S's attended observations and, hence, S's mental states ('my version of your concept/experience of pain'). This recursive dynamics continues until both agents might infer "alignment" (Vasil et al., 2020). Central here is that S is attending to the sensory states generated by C (and vice versa) because the only way to gather evidence for the adaptive prior that mental states are aligned is to attend to the sensory effects of one's actions; and evidence for hypotheses

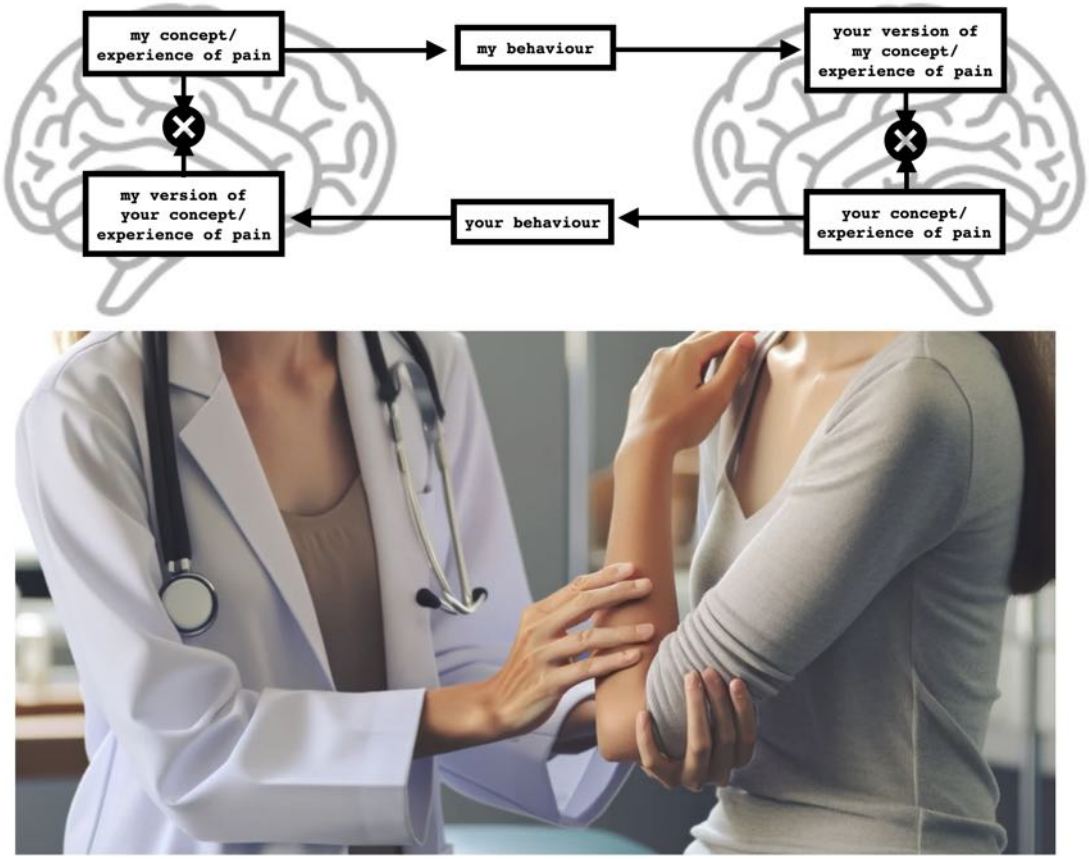


Figure 4.10: *The agent-environment-agent communication loop. This can be seen as the coupled action-perception cycle of two (dyadic) or more (collective) agents over time. This may vary as a function of context, common ground and experience of communicating agents. Adapted from Vasil et al. (2020).*

about the sensory effects of one's actions can only be given (in the present context) by the actions of the other agent.

It is important to note that there is a distinct lack of models specifically tailored for social pain within this domain, apart from what can be found in the broader literature on Theory of Mind and RSA/linguistic communication. These theoretical frameworks often encompass a wide array of interpersonal dynamics and cognitive processes, which include but are not limited to, the negotiation of shared understanding and the recursive prediction of each other's mental states. Despite this, the direct application to social pain remains underexplored, presenting a significant opportunity for future research to bridge these gaps and develop more nuanced models that can effectively capture the complex interplay of social interactions and pain perception.

Proposals for modelling these aspects will be presented and discussed in Chapter 8.

4.6 Discussion

The investigation into the computational models of pain has highlighted the intricate and multifaceted nature of pain perception and response. As discussed, pain is not merely a straightforward sensory experience but a complex interplay of sensory, cognitive, and emotional dimensions. The integration of these dimensions within a computational framework provides a deeper understanding of how pain is processed and experienced.

One of the central challenges in modelling pain, as highlighted by Seymour and Mancini (2020), is determining exactly what is being inferred during pain perception. Classical sensory inference tasks are typically exteroceptive, focusing on identifying observable properties of external objects. However, pain perception involves both exteroceptive and interoceptive inference, complicating the modelling process. To encompass these various components and their interactions, integrating same-level factors, we have specifically proposed a heterarchical computational model.

The hierarchical nature of pain processing further complicates the development of comprehensive models. Pain involves multiple levels of processing, from basic nociceptive signals in the spinal cord to complex cognitive evaluations in the prefrontal cortex. This hierarchical organisation requires models that can account for the dynamic interactions between different levels of the nervous system. For instance, the integration of sensory and motivational processes is not fully understood but is crucial for accurately simulating pain behaviours and responses.

The class division introduced in this chapter further refines our understanding by categorising models based on their complexity and focus.

The theoretical framework presented emphasises that while a comprehensive model integrating sensory, cognitive, affective, and social dimensions of pain—potentially with dynamic and active capabilities—is ideal, it is not always necessary or practical for all applications. Depending on the specific circumstances and objectives, it is essential to recognise which class of models is most suitable. This pragmatic approach allows for the development of tailored pain management strategies that are both effective and contextually appropriate.

CHAPTER 5

The gentle start: A case for Class 1 models

This section explores two distinct models applied to the Class 1 theoretical framework for video-based pain classification, representing the first complexity level in our pain modelling framework. They are characterised by their focus on temporal changes and dynamic patterns in pain expression, which are crucial for understanding how pain manifests and evolves over time. The choice of DCGNN and IO-HMM aligns with the objectives of Class 1 models in the following ways because they account, in different and peculiar ways, for temporal dynamics: both models explicitly account for the temporal aspect of pain expression. The DCGNN does this through its ability to process sequences of frames, while the IO-HMM models the temporal evolution of pain states. The first model employs a discriminative approach, utilising a deep convolutional Graph Neural Network (DCGNN) to infer pain from observable sufferer behaviour. The second model opts for a generative strategy through an Input-Output Hidden Markov Model (IO-HMM). Both models utilise the same dataset for their simulations, which is detailed to clarify the experimental context.

5.1 Dataset

Considering the subjective and multifaceted nature of pain, which varies across individuals and involves multiple dimensions, it is crucial to utilise a multimodal dataset to accurately reflect the complexity of pain perception. Therefore, the BioVid Heat Pain Dataset was chosen for its comprehensive collection of videos and biological signals documenting various pain expressions under controlled conditions. This dataset, derived from an experiment that involved 90 participants in three age groups, provides a diverse and balanced mix of gender and age, ensuring a representative sample for studying pain. It includes biopotential measurements capturing key physiological indicators,

such as:

- The Electrocardiogram (ECG), captured via two electrodes, monitors the heart's average action potential on the skin. Essential ECG-derived metrics include heart rate (BPM), interbeat intervals, and heart rate variability (HRV). Notably, HRV offers valuable insights into the autonomic nervous system's state, making it a critical factor in the psychophysiological evaluation of individuals.
- Electrodermal Activity (EDA), also known as galvanic skin response (GSR), is utilised to track changes in skin conductance levels. It serves as a dependable measure of an individual's psychological condition, solely reflecting the activity of the sympathetic nervous system without influence from the parasympathetic activity. This makes EDA a widely used tool for assessing the presence of pain, evidenced by a swift surge in EDA signals within one to three seconds following a stress stimulus.
- The Electromyogram (EMG) records electrical activity across muscles such as the corrugator, zygomaticus, and trapezius. This measurement is indicative of psychophysiological arousal. Essentially, an increase in muscle activity signals a rise in sympathetic nervous system engagement, indicative of heightened arousal. Conversely, reduced muscle activity suggests a dominance of the parasympathetic nervous system, associated with a state of relaxation.

Regrettably, the EMG data for the corrugator and zygomaticus muscles were not recorded accurately, which limited EMG analysis to the trapezius muscle alone.

The second segment of the database comprises video recordings of the experiment, captured using three AVT Pike F145C cameras at a frame rate of 25 Hz and a resolution of 1388 x 1038 coloured pixels. These cameras were positioned, one directly in front of the participants and two at the sides, to ensure complete capture of all facial movements. Two examples from the BioVid Heat Dataset are illustrated in Fig. 5.1.

Figure 5.1: *Two examples of video frame from BioVid Heat Dataset.*

Before starting the experiment, a crucial preparatory step involved calibrating pain thresholds for each participant. Subjects were required to establish their pain threshold T_P — the point at which pain first becomes perceptible — and their tolerance threshold T_T, the highest level of pain they could withstand. Based on these benchmarks, two additional intermediate pain levels were defined, resulting in four distinct pain levels

overall. Table 5.1 displays the average temperatures used for each level of heat stimulation across all participants, with BL1 denoting the "no stimulation" condition.

Stimulation label	MEAN+SD
BL1	31.9 ± 0.01
PA1	46.7 ± 2.61
PA2	47.8 ± 2.17
PA3	48.9 ± 1.8
PA4	50.1 ± 1.7

Table 5.1: *Average temperatures for each heat stimulation.*

During the 25-minute experiment, each participant underwent 100 consecutive tests (trials) with five different levels of heat stimulation to observe their reactions. The highest temperature for each pain level was maintained for 4 seconds, with each level being applied 20 times, totalling 80 stimuli. A random pause of 8-12 seconds was introduced between stimuli, as depicted in Fig. 5.2.

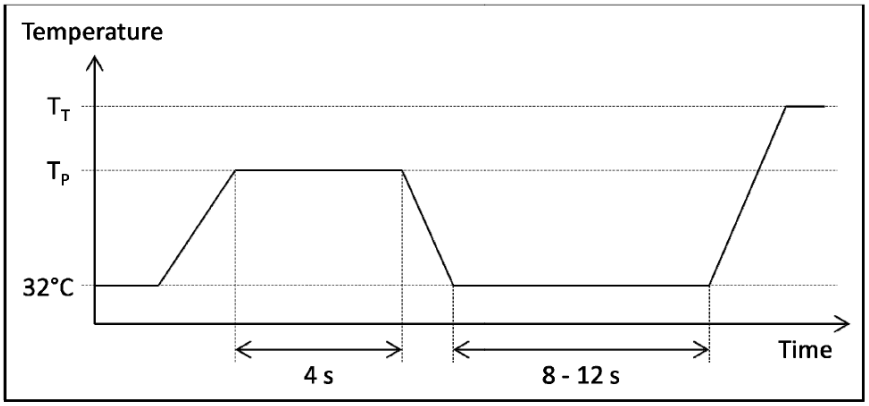


Figure 5.2: *BioVid experiment settings. Between two heat stimuli, a randomised pause is placed.*

5.2 A discriminative implementation model of a Class 1 model

We propose a discriminative model for pain detection that offers a pragmatic approach to identifying the presence of pain for immediate applications. While effective in operational settings, it's important to acknowledge that such models may not fully encapsulate the multifaceted nature of pain or contribute to expanding our understanding of the phenomenon. Discriminative models focus on surface manifestations rather than delving into the underlying complexities or generating new insights into the intricacies of pain, nevertheless, they can be crucial in contexts where the operational aspects are overriding. Indeed, when self-reports (most reliable and valuable methods for pain assessment) are not an option, automatic pain recognition systems based on facial reactions may provide a valuable alternative. In clinical contexts, for example, patients

may not be able to communicate verbally and medical personnel cannot continuously monitor them (Boccignone et al., 2020). In such a scenario, typical in intensive care units (ICU), automatic pain recognition systems based on behaviour and physiological responses could support the clinical routine of pain management.

Among behavioural pain responses, facial expressions are the most investigated (Calvo and D'Mello, 2010; Werner et al., 2022), primarily due to the significant role of the face as an information source, surpassing other nonverbal communication channels such as paralinguistic vocalisations or involuntary and deliberate body movements.

For instance, Prkachin and Solomon (2008), examining 129 individuals with shoulder pain, pinpointed specific facial movements that reliably distinguish between pain-related and non-painful activities. Similarly, Simon et al. (2008) demonstrated the human ability to differentiate between typical expressions of pain and other facial expressions, whether emotional or neutral. These studies highlight the uniformity of facial expressions associated with pain by emphasising the natural occurrence of certain facial responses to pain and the ability to identify pain in others through facial cues.

Leveraging this feature, our classification system profits from the natural graph representation of face landmarks (Shi et al., 2006), and relies on features describing local dynamics evolution. Hence, we take into account the progression of pain expression over time, thus overcoming the inherent limitations of frame-level approaches. To this end, we employ a DCGNN architecture able to capture the expression semantics connecting local motion information from face landmarks to the holistic view coming from the relationships between fiducial points.

Figure 5.3 illustrates the components of the implementation model in the context of the reference hierarchical model. From the observations, in this instance the videos, specific local perceptual elements, or features, are extracted. The more conceptual perspective comes into play when all elements are considered together, including their relationships, hence the graph, leading up to the final categorisation.

The video-based facial pain expression recognition we propose consists of three main steps: graph architecture definition, node-level feature representation, and graph processing.

5.2.1 Graph architecture definition

In our approach, we generate graphs from facial video sequences where each node represents a critical facial landmark. Nodes are interconnected if they fall outside a specific proximal range, determined by a set threshold of Euclidean distance. To effectively model transient expressions in lengthier footage, we segment videos into shorter clips. For each segment, every node is equipped with a feature vector designed to capture the landmark's dynamic behavior within that specific timeframe.

More specifically, given a video v, if it is longer than f frames, it is split into short clips $v^i, i \in 1...k$, $|v^i| = f$[1]. On each frame in v^i the method extracts a set of fiducial points. In current implementation, we use the method presented in Kartynnik et al. (2019) (see Fig. 5.4), deriving a dense map of fiducial point that we lighten applying a uniform subsampling (see Fig. 5.5).

Each clip v^i is modelled by a graph G_v^i with nodes corresponding to the n selected

[1] In case of videos with length not multiple of f, the last shortest video clip will be discarded

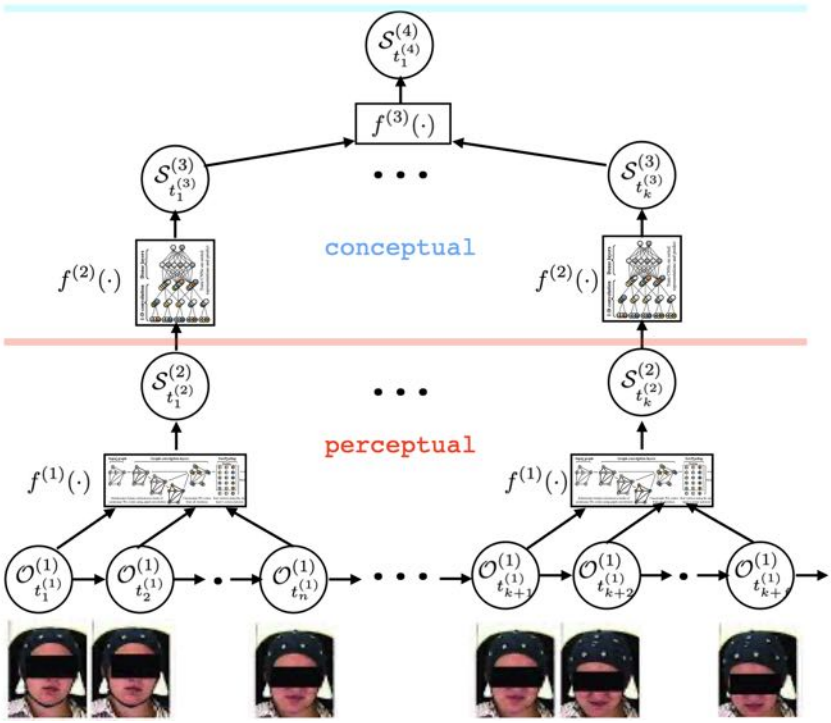


Figure 5.3: *A Forney-like factor graph that provides a discriminative implementation model of a Class 1 model. Circles represent observational/perceptual and conceptual/categorical states. Squares denote factors/functions $f^{(3)}(\cdot)$ operating at the different levels. Note that the reversed arrows (from bottom to top direction) represent the model at the inference/discriminative stage*

landmarks, and edges created connecting nodes outside the close neighbourhood according to the Euclidean distance between each pair of landmarks, using an experimentally fixed threshold (see Fig. 5.6). This way, the local information can be shared between distant areas, fostering the message passing all over the graph. As detailed in Sec. 5.2.2, the node characterisation is conceived to produce a feature vector capturing the trajectory followed by the corresponding landmark in the clip at hand.

5.2.2 Node-level features representation

Each fiducial point f is characterised considering its trajectory as a 2-dimensional stochastic process, which (x_f, y_f) coordinates are assumed to be independent from one another. As a consequence, we end up examining $2n$ time-series in total.

Each trajectory is characterised using a set of complexity-related measures, delivering insights concerning the dynamics and predictability of the time series, and spectral attributes summarising the properties of the signals in the frequency domain. In particular, we consider the following features:

Approximate Entropy (ApEn)

ApEn (Pincus, 1991) is a statistical measure used to quantify the amount of regularity of fluctuations in time-series data. Larger values indicate higher complexity or irregularity

Figure 5.4: *Face mesh based on Kartynnik et al. (2019).*

Figure 5.5: *Face mesh following the subsampling.*

in the data. ApEn has been extensively used for the analysis of physiological time-series (Richman and Moorman, 2000; Sabeti et al., 2009).

Sample Entropy (SampEn)

As ApEn, SampEn is another measure of the complexity of a signal. Large values indicate high complexity, whereas smaller values characterise more self-similar and regular signals. The *SampEn* of a signal x is defined as:

$$SampEn(x, m, r) = -\log \frac{C(m+1, r)}{C(m, r)} \tag{5.1}$$

where m is the embedding dimension (in our experiments we set $m = 2$) and r is the radius of the neighbourhood (in our case $r = 0.2 * std(x)$). $C(m+1, r)$ and $C(m, r)$ are the number of embedded vectors of length $m + 1$ and m respectively, having a Chebyshev distance inferior to r.

Figure 5.6: *Example of edges for a single node. The radius of the red circle represents the minimum distance for connection.*

Permutation Entropy (PermEn)

The *PermEn* is a complexity measure for time-series first introduced by Bandt and Pompe (2002).

Given a signal x, it is defined as:

$$PermEn = -\sum p(\pi)\log_2(\pi) \tag{5.2}$$

where π is the set of $p!$ permutations of x of order p. In our experiments we set $p = 3$. As with $ApEn$ and $SampEn$, the smaller $PermEn$ is, the more regular and more deterministic the time series is. In contrast, higher values of $PermEn$, suggest more noisy and random time series.

SVD Entropy (svdEn)

SVD Entropy (Roberts et al., 1999) indicates the number of eigenvectors that are needed for explaining the data. In other words, it measures the dimensionality of the data.

Define an embedding matrix Y of a signal x as:

$$y(i) = [x_i, x_{i+\text{delay}}, \ldots, x_{i+\text{ (order-1)*delay}}]$$
$$Y = [y(1), y(2), \ldots, y(N - (\text{ order } - 1)) * \text{ delay })]^T$$

where $delay = 1$ and $order = 3$ represent the considered time delay and the length of the embedding dimension, respectively.

The SVD entropy is then obtained as:

$$svdEn = -\sum_{i=1}^{M} \bar{\sigma}_i \log_2(\bar{\sigma}_i) \tag{5.3}$$

where M is the number of singular values of the embedding matrix Y and σ_i are the normalised singular values of Y. As for the previous measures of Entropy Rate, *svdEn* is lower for simpler time series and higher for more complex ones.

Detrended Fluctuation Analysis (DFA)

DFA (Peng et al., 1995) is a method for determining the statistical self-affinity of a signal. Similarly to the Hurst Exponent, it is useful for analysing the the signal correlation behaviour and it allows the detection of the long-range dependencies. However, differently from Hurst exponent, DFA may also be applied to signals whose underlying statistics (such as mean and variance) or dynamics are non-stationary (changing with time). The computation of DFA, goes as follows. The original signal x on length N is first integrated and its average is subtracted:

$$X = \sum_i (x_i - \langle x \rangle) \tag{5.4}$$

The resulting cumulative sum X is divided into chunks of length c, within which the linear trend Y is computed. Let Y_i indicate the resulting piece-wise sequence of straight-line fits representing the linear trends estimated via least square fitting in each window. Then, the root-mean-square deviation from the trend (the fluctuation) is calculated as:

$$F(c) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X_i - Y_i)^2} \tag{5.5}$$

Detrending followed by fluctuation measurement is repeated over a range of different window sizes c and a log-log plot of $F(c)$ against c is constructed, upon which a straight line is fitted. The slope of this line represents the scaling exponent α delivering information about the self-affinity of the process.

Higuchi Fractal Dimension (HFD)

HFD is a method for approximating the fractal dimension of a time series. HFD measures the rate of increase in the difference of signal amplitude while the signal samples are picked in an increasingly sparse way. HFD is computed as follows. Given a time series x, For each $k \in \{1, \ldots, K\}$ and $m \in \{1, \ldots, k\}$ define the length $L_m(k)$ by:

$$L_m(k) = \frac{N-1}{\left[\frac{N-m}{k}\right] k^2} \sum_{i=1}^{\left[\frac{N-m}{k}\right]} |X_N(m+ik) - X_N(m+(i-1)k)|$$

Total average length $L(k)$ is computed as:

$$L(k) = \frac{1}{k} \sum_{m=1}^{k} L_m(k)$$

The HFD is represented by the slope of the best fitting straight line on the log-log plot of $\frac{1}{k}$ against $L(k)$. In our experiments we set $K = 10$.

Petrosian Fractal Dimension (PFD)

PFD represents another method for estimating the fractal dimension of a signal. In particular, the Petrosian fractal dimension of a time-series x is defined as:

$$PFD = \frac{\log_{10}(N)}{\log_{10}(N) + \log_{10}\left(\frac{N}{N+0.4N_s}\right)} \tag{5.6}$$

where, N is the length of the time series, and N_s is the number of sign changes in the signal derivative.

Katz Fractal Dimension (KFD)

Katz Katz (1988) proposed yet another method for estimating the fractal dimension of a time-series. Specifically, KFD can be computed as:

$$KFD = \frac{\log_{10}(L/a)}{\log_{10}(d/a)},$$

where L is the sum of distances between successive points, a is their average, and d is the maximum distance between the first point and any other point of the considered signal.

Zero-Crossing Rate (ZCR)

The zero-crossing rate (ZCR) is the rate at which a time-series changes from positive to zero to negative, or from negative to zero to positive. Formally, given a signal x of length N, ZCR can be defined as follows:

$$ZCR = \frac{1}{N-1}\sum_{i=1}^{N-1} 1_{\mathbb{R}_{<0}}\left(x_i x_{i-1}\right)$$

where $1_{\mathbb{R}_{<0}}$ is the indicator function.

Mel Frequency Cepstral Coefficients (MFCCs)

Besides considering complexity-related measures of the time domain signals represented by (x, y) coordinates of the facial landmark points, we augment the feature set associated to each node with some spectral features. In particular, we compute the first 13 Mel Frequency Cepstral Coefficients (MFCCs) on each trajectory.

MFCCs are coefficients derived from a representation of the short-term power spectrum of a signal, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. They have been extensively used in speech and sound processing as delivering compact and informative summary of the spectral content of a signal. Specifically, MFCC computation is carried out as follows:

- each signal describing the trajectory of a landmark w.r.t. its x or y coordinate, is first transformed to the frequency domain;

- a Mel filter-bank is applied to the spectrum and the energy in each filter su summed
- the logarithm of all filter-bank energies is taken
- the DCT of the log filter-bank energies is computed
- the first 13 coefficients are eventually kept.

Once all the complexity measures and MFCC features are extracted, they are concatenated to form a 44-dimensional feature vector for each node.

5.2.3 Graph processing

Given a dataset of videos $D = \{v_j, j \in 1..d\}$, each one labelled by $l_v \in \{Pain, notPain\}$, we split D into train and test sets, and for each video v, the corresponding clip graphs G_v^i (possibly 1) are computed as described in Sec. 5.2.1 and Sec. 5.2.2.

To solve the binary classification over graphs, we resort to the DGCNN (Zhang et al., 2018). This is trained on the pairs $\{G_v^i, l_v\}$ in the training set, l_v being the label of the video the clip belongs to.

In testing phase, for each video v we create its graphs G_v^i, and evaluate them by collecting the classifications $C_v = \{\hat{l}_v^i\}$. The video classification $\hat{l}_v$ is finally computed as the median over C_v.

5.2.4 Simulation

Acted pain classification

In the first experiment, we use the Part D of the database, selecting pain and neutral videos for a total of 178 videos. Each video is divided into 200-frame sequences, reaching a total of 1245 samples. This implementation choice is motivated by the very nature of pain expression dynamics, typically discontinuous. Hence, in order to discriminate the presence of pain in a video, we analyse short windows and thence make a global prediction at the video level. Moreover, this approach eases the comparison between the two experimental settings, characterised by videos with a significant difference in duration.

Then, for each sequence we collect the trajectories (in x and y dimensions) of 94 face landmarks, obtained by applying a uniform subsampling to the 468 ones delivered by the MediaPipe Python library (Lugaresi et al., 2019), and finally we derive the set of 44 features per landmark. This information is then associated to each node of the graph, afterwards completed by edges between pairs of landmarks (i.e. nodes) whose distance exceeds an experimental fixed threshold equal to double the distance between the eyes. The number of edges obtained is 4032 on average, given that the amount of edges hinges on the facial configuration in the specific sequence.

Following the feature extraction step, the DGCNN classifier is trained using the Adam optimisation algorithm to minimise the binary cross-entropy loss and evaluated via 5-fold cross-validation where the train/test split is performed to avoid the simultaneous presence of sequences taken from the same video in train and test set. In addition to cross-validation, we implemented dropout and early stopping to further mitigate the

risk of overfitting. Dropout, applied at a rate of 50%, helps reduce model complexity by randomly deactivating half of the neurons during training, while early stopping halts training when no improvement in validation loss is observed over a certain number of epochs. We modified the network structure, making some variations to the model proposed in (Zhang et al., 2018). First, we add a graph convolutional layer to the original network and double the number of kernels, reaching a total of five graph convolution layers with 64, 64, 64, 64, 1 output channels, respectively. Also, the SortPooling layer is revised to keep the first 40 sorted nodes. Moreover, the hyperbolic tangent activation function is replaced with rectified linear units (ReLU).

The results obtained are presented in the Tab. 5.2 in comparison to a baseline Support Vector Machine (SVM) model created in order to have a benchmark, since, as far as we know, there are no works in literature adopting the acted part of the database for pain classification. In order to have an appropriate data structure for the SVM classifier training, the graph structure was discharged, flattened to a 4324-dimensional feature vector, and then reduced to a 200-dimensional vector using PCA. It is worth noting that the information carried by edges is unavoidably lost in the baseline model.

Further, we compare our DGCNN classifier with an SVM adopting Action Units (AUs) intensities as features. This more standard approach obtains almost the same performance as the SVM baseline, pointing out the importance of the graph structure for learning effectiveness over the concatenation of the feature sets.

In Tab. 5.2 we report the evaluations of the proposed method (CM+DGCNN), and the baselines (AUs+ SVM and CM+SVM), proving the effectiveness of the adopted features (CM) in combination with the graph structure and learning.

Model	Video-level accuracy
AUs+SVM	0.669 ± 0.146
CM+SVM	0.714 ± 0.102
CM+DGCNN	$\mathbf{0.834} \pm 0.116$

Table 5.2: *Results on acted pain videos (Part D) using proposed complexity measures (CM) in combination with DGCNN model compared to standard AUs intensity features and SVM.*

Spontaneous pain classification

The spontaneous pain discrimination task is, in general, more worthwhile but also challenging. For this experiment we refer to the short video sequences (5.5 seconds) included in the Part A of the database, taking into account only the sequences labelled as pain-free (0/4) and with maximum pain intensity (4/4). In doing so, we obtain 40 videos per participant (87 subjects altogether), 20 for each label, totalling 3480 videos.

There are no differences in the feature extraction step and the network structure compared to the acted experiment. Although, in this session there is a one-to-one correspondence between videos and graphs motivated by the shortness of video sequences and by the presence of a single painful stimulation per video. For this reason, the video-level accuracy is equivalent to standard accuracy of the DGCNN.

For this experiment we evaluate our approach, CM+DGCNN, and compare it to both the baseline method AUs+SVM, and the results reported in Werner et al. (2016) and Werner et al. (2017). As shown in Tab. 5.3, our results are slightly above the state of the art on this Database.

Model	Accuracy
AUs+SVM	0.648 ± 0.068
Werner at al. (2016)	0.700
Normalised Werner at al. (2016)	0.724
Werner et al. (2017)	0.718
CM+DGCNN	$\mathbf{0.732} \pm 0.139$

Table 5.3: *Results on spontaneous pain videos (Part A) in comparison with the state of the art and a plain AUs-based classifier. Werner et al. (2016) reports two results. The second adding a feature standardisation per subject.*

5.3 A generative implementation model of the Class 1 model

As with the model just presented, data-driven models are proficient at detecting the presence of pain through facial expressions or physiological signals, yet they often fail to grasp the multifaceted nature of pain itself. By relying heavily on identifying patterns within datasets, these models overlook the deeper, underlying mechanisms that contribute to the experience of pain. They lack the capacity for explanatory depth, cannot engage in counterfactual reasoning, and neglect the importance of contextual factors that might influence the perception and expression of pain. This limitation becomes particularly evident when applying these models to new datasets, where their generalisation capabilities fall short.

In contrast, theory-driven models, anchored in psychological and neuroscientific theories of pain, offer a more nuanced understanding. These models attempt to encapsulate the complex processes underlying pain experiences, incorporating insights from research on how the brain processes pain, the psychological factors at play, and the body's physiological response to pain stimuli. For instance, some models might integrate the concept of pain appraisal, recognising that cognitive evaluations of stimuli significantly impact pain perception and response. Others may explore the interconnections between pain, cognition, and behaviour to provide a more holistic view of pain experiences. However, the custom nature of these models, designed to reflect specific theoretical frameworks, limits their flexibility and applicability across diverse situations. Despite their potential to more accurately reflect the complexity of pain, their specificity and lack of adaptability pose significant challenges for broader application, underscoring the need for models that can navigate the complexities of pain experiences across various contexts and conditions.

Merging the strengths of both data-driven and theory-driven strategies, a probabilistic approach offers a pathway to encapsulate the multifaceted nature of pain experiences characterised by their inherent uncertainty and the often incomplete understanding of

an individual's state. Generative models are adept at capturing the multifaceted nature of pain, marked by uncertainty, randomness, or incomplete knowledge of an individual's physiological and psychological states. Its modular design enables the crafting of complex programs that can mirror various abstraction levels within a hierarchical framework, facilitating context-specific learning and generalisation across different scenarios. Hence, this methodology facilitates the creation of models grounded in psychological theory, enabling hypothesis testing and scientific experimentation.

The selection of a psychological framework for implementation is critical. While traditional models based on basic emotion theories or appraisal theories have struggled to encapsulate the full breadth of pain experiences, a newer perspective has gained traction (Ong et al., 2018), inspired by human intuitive understanding of their surroundings and interactions. Humans possess an inherent "intuitive physics" that enables them to predict physical phenomena, alongside an "intuitive psychology" for making sense of others' mental states. This "folk theory", a comprehensive ontology encompassing concepts such as goals, behaviours, and their interrelations, extends into the domain of pain. It embodies an "Intuitive Theory of Pain", enriching our understanding of how people reason about pain, its causes and effects.

In this framework, an observer uses intuitive theories to infer the pain states of an agent experiencing pain, considering motivationally significant events and the agent's beliefs and desires. This process, termed "third-person appraisal", reflects the observer's beliefs about the agent's pain experience, influenced by their own past encounters and cultural background. Thus, this approach provides a scaffold for exploring pain in a way that accommodates individual variability and cultural diversity, offering a comprehensive and adaptable model for pain perception and response.

In probabilistic terms this model can be represented via the PGM presented in Fig. 5.7.

Applying the concepts to the data available and the experimental setting described in this chapter, we can differentiate between the generative process and the generative model. The former occurs when subjects experience pain through a heat stimulus and subsequently exhibit a reaction, engaging in an action that could be facial expression (exteroceptive) or interoceptive regulation, such as thermal regulation. Participants are not required to provide a live evaluation of their pain level; hence, they do not perform an explicit conceptualisation. If we aim to create a model for assessing the pain experienced by participants, essentially modelling the process, we can envisage the presence of an external rater or observer tasked with inferring the perceived pain of the participants based on observations and measurements, thereby simulating their approximation of the process (thus, activating their model). Fig. 5.8 illustrates the distinction and interplay between process and model.

5.3.1 Implementation model

The simulation relies on the *naive* Observer's model illustrated in Figure 5.9, which reduces and instantiates the general model outlined in Figure 5.7.

The Observer conjectures that a painful stimulus u_t will be provided to the agent, which will affect the Agent's physiological and behavioural (facial AUs) responses $\mathbf{O}_t = \{O_t^{EDA}, O_t^{HR}, O_t^{EMG}, O_t^{AU}\}$ that can be directly observed or at least measured.

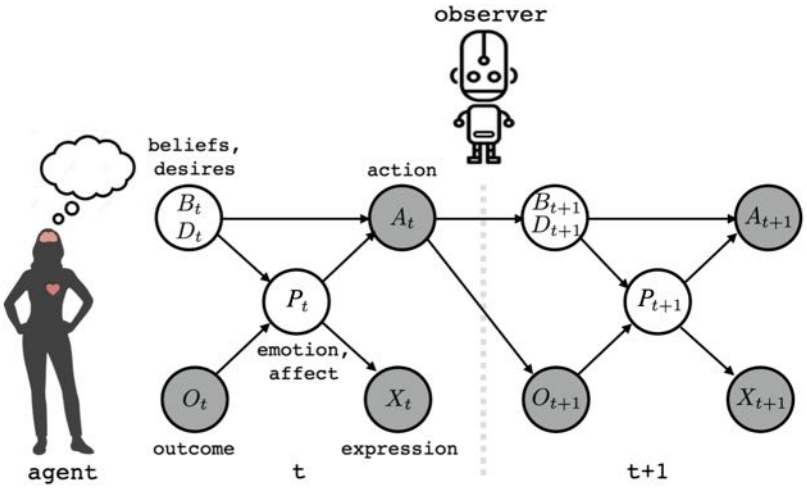


Figure 5.7: *A model of a folk theory of emotion in the form of a PGM, unfolded in time, where the time slice [t, t+1] is shown. Shaded circles represent observable variables, unshaded circles represent latent variables and arrows express a directed causal relationship. Beliefs, desires and the outcome of an event together result in an emotion, which later elicits an expression or an intentional action. The action at time* t *then leads to new outcomes at time* $t+1$*, so that the whole cycle repeats itself. Redapted from (Ong et al., 2018).*

Such input/output function is mediated through a latent state $S_t = \{C_t, P_t\}$ summarising Agent's hidden higher level perceptual/conceptual processing.

Moment by moment, O depends on S and on the input control U (pain stimulus).

Define the following stochastic processes, $U = \{u_t, 0 \leq t \leq T\}$, $S = \{s_t, 0 \leq t \leq T\}$, $O = \{o_t, 0 \leq t \leq T\}$, taking values in $\mathbb{R}^u$, $\mathbb{R}^z$ $\mathbb{R}^o$, respectively, over a time horizon T. Observations are made at discrete time points t_k, $k = 0, 1, ..., T$, so $o_k = o_{t_k}$ denotes the observable variable at time t_k. Similarly, $s_k = s_{t_k}$ denotes the latent state at time t_k, consistent with the assumption about the Agent's internal model dynamics. In this case, S can be defined as a first-order Markov process. The State Space Model (SSM) is then the stochastic dynamical system:

$$s_k = a(s_{k-1}, u_k, \xi_k) \quad \text{(hidden state equation)} \tag{5.7}$$

$$\mathbf{o}_k = h(s_k, u_k, \eta_k) \quad \text{(measurement equation)}. \tag{5.8}$$

Here, a and h are generic nonlinear, potentially time-varying functions, and ξ_k and η_k represent process (dynamical) and measurement noise, respectively. The state equation characterises the state transition probability $P(s_k \mid s_{k-1}, u_k, \theta_s)$ with parameters θ_s, while the measurement equation describes the probability $P(\mathbf{o}_k \mid s_k, u_k, \theta_o)$ related to the measurement noise model with parameters θ_o. The state-space dynamics can be equivalently written as:

$$u_k \sim P(u_k) \tag{5.9}$$

$$s_k \mid s_{k-1} \sim P(s_k \mid s_{k-1}, u_k) \tag{5.10}$$

$$\mathbf{o}_k \mid s_k \sim P(o_k \mid s_k, u_k). \tag{5.11}$$

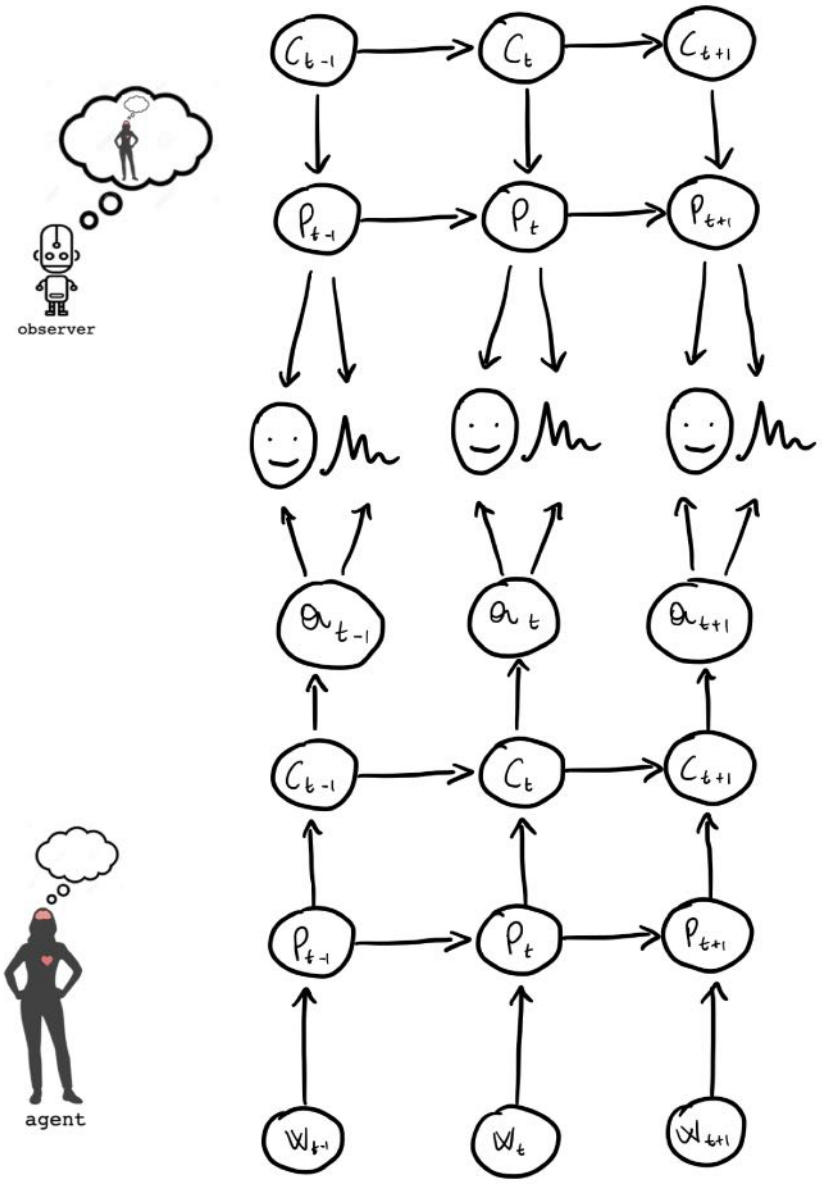


Figure 5.8: *The lower part of the figure illustrates the generative process in action during the experiment. The subject, exposed to a potentially painful stimulus, gathers interoceptive and exteroceptive evidence which is used to generate a perception that, in turn, may be conceptualised and thus categorised as a painful experience. Based on this experience, the subject will respond with an action (facial expression and autonomic regulation). The observer, as depicted in the upper part of the figure, is equipped with a model analogous to that of the experiment's main agent. This model enables the observer, from the agent's visible and/or measurable reactions, to infer an approximation of their perception and thus categorise the agent's experience as either painful or non-painful.*

where sampling $u_k \sim P(u_k)$ ideally denotes nature's choice of providing a painful stimulus to the agent.

Note that, by assuming a probability distribution over a discrete state dynamical

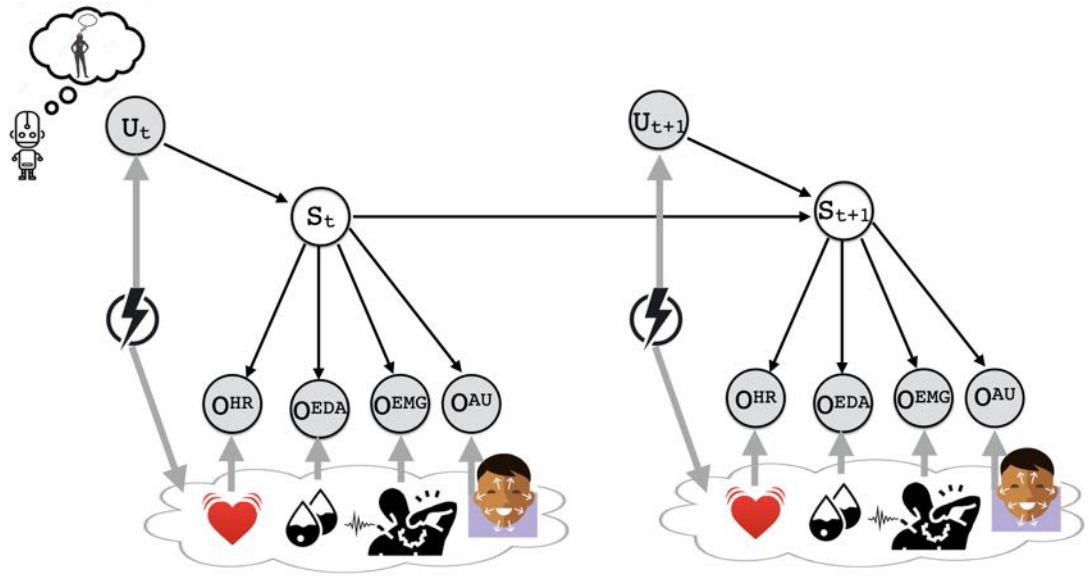


Figure 5.9: *The Observer conjectures that a painful stimulus u_t will be provided to the agent, which will affect the Agent's physiological and behavioural (facial AUs) responses $\mathbf{O}_t = \{O_t^{EDA}, O_t^{HR}, O_t^{EMG}, O_t^{AU}\}$ that can be directly observed or at least measured. Such input/output function is mediated through a latent state S_t summarising Agent's hidden higher level perceptual/conceptual processing.*

system, the above model can be easily shaped in the form of an Input-Output Hidden Markov Model (IO-HMM, but see Bengio and Frasconi (1996) for details), wherein the stimulus is treated as input and the ensuing reaction as output. The objective is to deduce the hidden state that represents the patient's perceived pain. Operating in an unsupervised framework precludes the provision of a genuine accuracy metric. However, by examining the transition matrix, emission matrix, and Viterbi path, various insights can be gleaned.

For the scope of this thesis, biological and video signals are considered. Detailed explanations of the data processing techniques for both types of signals are provided in their respective sections.

Video processing

The video dataset comprises recordings from 87 patients, each providing 100 videos. Various methods for extracting pertinent facial features were explored, including facial landmarks. However, Action Units (AUs) emerged as the most effective, offering a compact yet comprehensive representation of facial expressions. For enhanced accuracy in extracting action units, OpenFace software (version 2.2.0) was selected for its robustness and precision (Amos et al., 2016). This tool analyses each frame (up to 138 frames per video), ensuring maximal retention of information. OpenFace facilitates the identification of 18 distinct action units, each scaled from 0 to 5, along with a corresponding confidence level. During the video analysis, one file was found to be corrupted, preventing the processing of subject 101916 (a 40-year-old male). The

subsequent stages involved filtering and selection of AUs for aggregation and further analysis. To pinpoint the AUs most relevant to pain-related facial activity for this thesis, we drew upon research by Kunz and Lautenbacher (2014), which identifies four distinct activity patterns. Specifically, the AUs selected for inclusion were AU01, AU04, AU06, AU07, AU09, AU10, AU23, AU25, and AU26 (refer to Figure 5.10 for a visual guide).

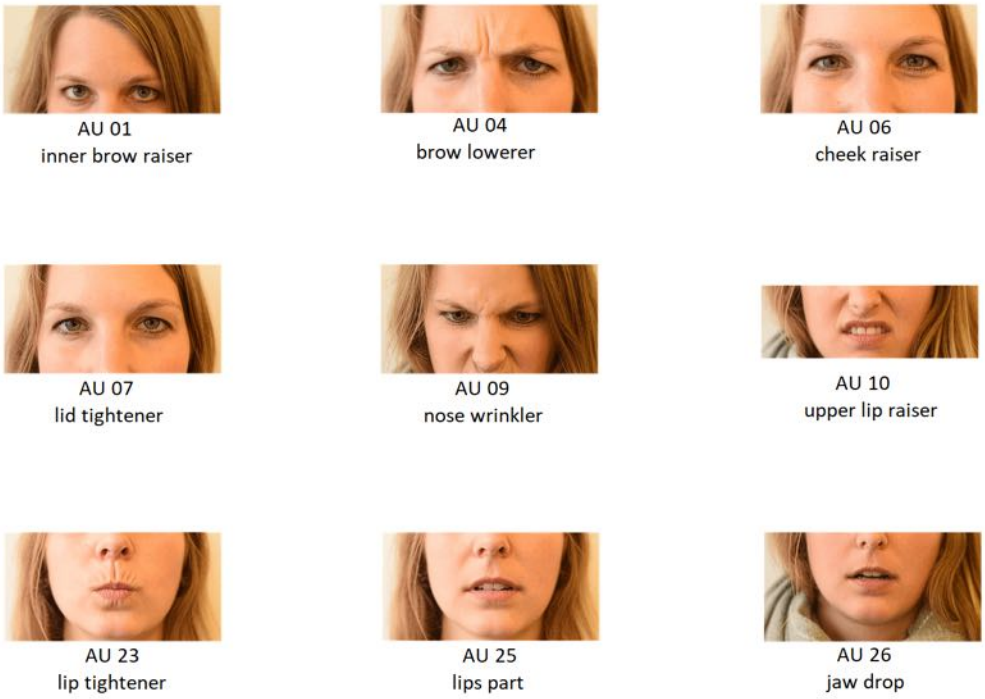


Figure 5.10: *Visual reference to the AUs selected. From: imotions.com*

To enhance the analysis of patients' reactions to the heat stimulus, which begins one second into each video (see Fig. 5.2), the initial 30 and final 15 frames were omitted to focus on the most relevant reactions and minimise potential noise. A confidence threshold was also applied during the filtering process, ensuring only frames with a confidence level above 80% were used. The measurements for the Action Units (AUs), except for AU04, were then averaged across all remaining frames for each video. For AU04, known for its high activation across various stimuli levels, the range (maximum minus minimum value) was calculated to capture its activation variability more effectively.

The preprocessing concluded with the encoding of stimulation labels into integer values for simpler model integration. The original labels represented five levels of heat intensity, from no stimulus (BL1) to maximum intensity (PA4). In the binary approach, BL1, PA1, and PA2 were classified as 0, and PA3 and PA4 as 1, based on observed patterns indicating minimal reaction difference within these groups. For a ternary classification, BL1 and PA1 were encoded as 0, PA2 and PA3 as 1, and PA4 as

2. Consequently, the final dataset for each patient includes video number, stimulation level, and average or range values for selected AUs, ready for model application.

Biosignal processing

The biomedical signal dataset comprises 100 CSV files for each of the 87 patients, with each signal sampled at 512 Hz and lasting 6 seconds. Following Werner's preprocessing approach, signals have been prefiltered for analysis. The aim is to extract key features from the ECG signals, such as BPM (beats per minute) and HRV (heart rate variability), with HRV identified as a crucial marker for detecting responses to pain. These features are processed using the Biosppy library (Carreiras et al., 2015), with R-peaks detection for BPM and rMSSD for HRV.

For EDA signals, which measure skin conductance in microsiemens, the focus is on capturing the phasic component related to pain stimuli responses. A high-pass filter isolates this component, and the SCR amplitude is measured to assess pain. This process, simplified due to the brief signal duration, relies on a single maximum peak detection with the NeuroKit2 tool (Makowski et al., 2021).

The EMG signal, specifically from the trapezius muscle, is analysed to detect stress levels indicative of pain. After excluding initial noise, the clean signal's amplitude, denoting muscle activity, is calculated using the Teager-Kaiser Energy operator and a Butterworth filter for the linear envelope. This method enhances muscle activation detection, with the final dataset encoding stimulation levels similarly to video data, thus providing a comprehensive overview of each patient's pain response through HRV, SCR amplitude, and EMG amplitude measurements.

5.3.2 Clustering

As outlined by the reference theoretical model, the conceptualisation process follows a hierarchical logic, whereby sensory data (both exteroceptive and interoceptive) are collected and hierarchically categorised through a process that transitions from perception to a more explicitly cognitive concept. In the implementation of this model, this step of perceptual categorisation is achieved through a clustering phase, essentially a procedure for category construction. This process, based on the collected data, defines (outlines the boundaries of) a category. Specifically, from the subject's reaction, including facial expression and physiological response, a more abstract representation is constructed that can denote the data's membership, or lack thereof, in the pain category.

Hence, following the preprocessing of our data, the next step is to establish how the extracted features will be represented, which will act as the observation or output in the IO-HMM.

A variety of algorithms were evaluated for the clustering phase, yet no significant differences were observed among them, leading to the selection of the k-means algorithm with k set to 2. The consideration of more than two clusters was explored, but the specific requirements of the study and characteristics of the dataset limited the depth of analysis that could be achieved. Specifically, the variation in temperature levels during heat stimulation did not justify a move towards a non-binary classification approach, as indicated in Table 5.1. Typically, patients showed a constrained range of responses, with reactions to PA1 and PA2 stimuli often being indistinguishable from the

no-stimulus condition, BL1. This issue was apparent in the clustering efforts, where a third potential cluster did not distinctly represent a separate response from the patients. Thus, the analysis was simplified to a binary framework, identifying two clusters: cluster 0, indicating no response, and cluster 1, representing a noticeable reaction to the stimulus. This binary classification was applied consistently across both video and physiological data to eliminate any confusion.

In the process of video clustering, we took into account all the characteristics identified during the initial preprocessing. Following this, we generated heatmaps from the resulting confusion matrices to evaluate the clustering outcomes. The procedure proved to be effective, particularly with younger patients, achieving a well-balanced distribution of clusters as depicted in the heatmaps in Figure 5.11. However, the method faced challenges with older participants, where the algorithm's results were less consistent. This inconsistency arose due to the presence of wrinkles, which caused an increase in Action Units (AUs) activation. This in turn introduced additional noise, impacting the clarity of the clustering outcomes, a phenomenon clearly observable in the heatmaps shown in Figure 5.12.

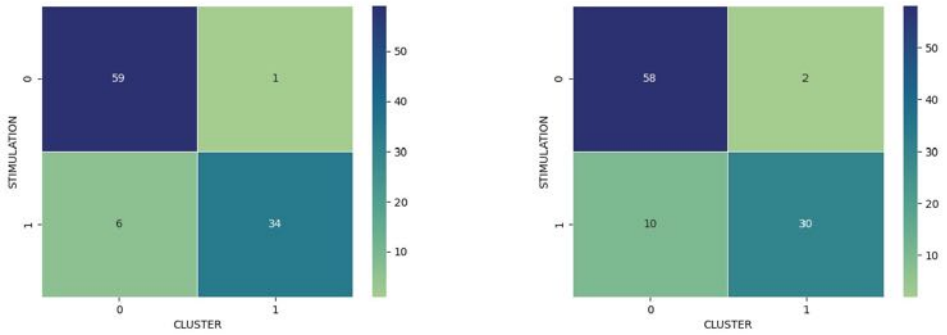


Figure 5.11: *Heatmap of confusion matrix for younger patient: a man (left) and a woman (right).*

To evaluate the accuracy of the clustering process, we visualised the central points of some clusters utilising `OpenFACS` (Cuculo and D'Amelio, 2019), an open-source software based on the Facial Action Coding System (FACS) that allows for the simulation of facial expressions by adjusting specific Action Units (see Figure 5.13). Through this method, we expected the centroid of the first cluster, labelled as 0, to depict a neutral facial expression, whereas the centroid of the second cluster, marked as 1, should illustrate an expression suggestive of pain. In the instance of a younger participant (illustration a), there was a clear shift from a neutral to a pained expression, particularly noticeable in the activation of AU 04, which is responsible for furrowing the brow. Conversely, the portrayal of an older participant (illustration b) showed a more pronounced painful expression due to increased activation of AUs, highlighting the robustness of this approach in capturing varied facial expressions across unbalanced clusters.

In our exploration of psychological data, several key challenges emerged that impacted the processing approach. Initially, the short duration of the six-second signals

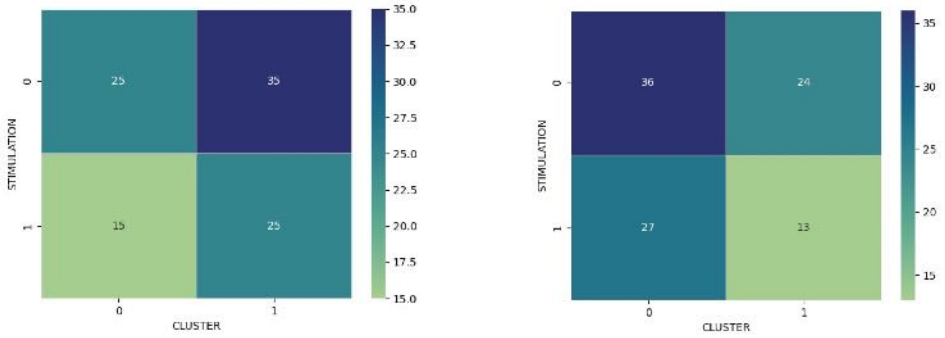


Figure 5.12: *Heatmap of confusion matrix for older patient: a man (left) and a woman (right). An increase in AUs activation noise can be detected.*

introduced a latency challenge, making it difficult to extract significant features, particularly from the ECG signals where changes within such a brief period were hard to distinguish. This issue was compounded by potential carryover effects into subsequent trials, adding ambiguity and potential noise. Another challenge was the unrecorded randomised pause between trials, which obstructed a thorough analysis of the signal progression over time. Furthermore, without access to pre-stimulation signal data, comparing activation levels before and after stimulus application was not possible, limiting insights into signal behavior and noise detection. The inclusion of 20 neutral stimuli (BL1) at the end of the series also posed a problem, as patient movements could introduce noise into the EMG signals from the trapezius muscle. Consequently, both ECG and EMG data were excluded from the clustering analysis due to these issues. Practical testing of various feature configurations underscored these theoretical concerns, revealing that incorporating all features led to inconsistent and less informative clustering results. Notably, isolating the SCR amplitude for analysis yielded more balanced clustering, supported by both sensitivity and specificity measures, as detailed in Table 5.4. This finding underscores the SCR amplitude's resilience to the issues identified, attributable to its ability to rapidly capture responses to stimuli via its phasic component.

Configuration	Sensitivity	Specificity
SCR + EMG + BPM + HRV	0.80	0.25
SCR + EMG + HRV	0.74	0.37
SCR + EMG	0.94	0.16
SCR	0.92	0.40

Table 5.4: *Cluster balancing for different configuration of extracted features.*

In Figure 5.14 the difference between considering all features (left) and only the SCR amplitude (right) is shown — considering the same patient. In the first image, the

Figure 5.13: *Clusters results for videos processed via OpenFACS:* ***(a)*** *younger patient* ***(b)*** *older patient.*

heatmap is strongly unbalanced and only one trial is classified as “positive reaction", while in the second image the heatmap depicts a more balanced situation.

After performing the clustering step, the label of the assigned cluster becomes the label of the outcome.

5.4 Implementation details

For our simulations, we explored the IO-HMM architecture (see Fig. 5.9), in particular focusing on the unsupervised framework, employing an open-source IO-HMM pack-

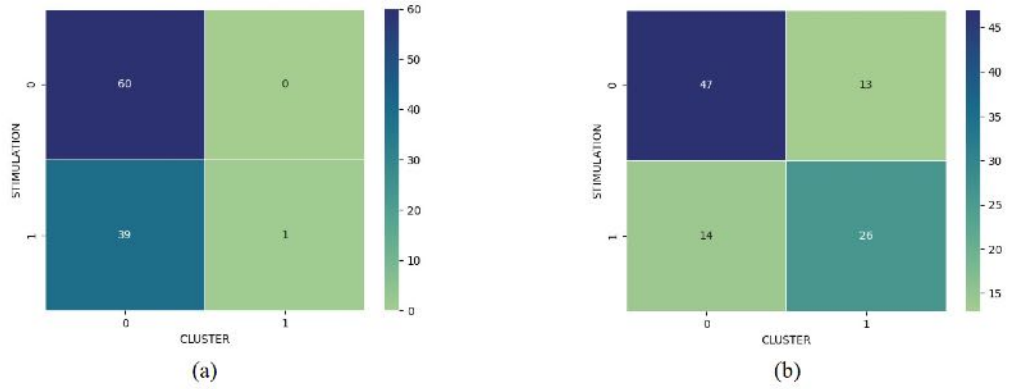


Figure 5.14: *Heatmap evolution of a 23 years-old woman.* ***(a)*** *heatmap resulting from a clustering with all the features considered.* ***(b)*** *heatmap resulting from a clustering with only SCR amplitude considered.*

age[2].

Two latent states were selected to represent the presence or absence of pain, configuring a unique model for each of the 87 patients, reflecting the 100 trials outlined in the dataset. Each model was developed independently to capture the distinct pain experiences of each patient, culminating in a suite of 87 unique models. The inputs for these models incorporate the two-level encoded stimulations as detailed earlier, while the outputs are based on the clustering of patient behaviours. To ensure the reliability of the models, a 5-fold cross-validation approach was applied, with outcomes averaged to enhance accuracy. The training phase was dedicated to refining the models' parameters, including the initial state probabilities, transition, and emission matrices. A variety of configurations were tested, with the most effective setup involving a cross-entropy loss in conjunction with multinomial logistic regression with an L-BFGS solver for the transition and initial state models, alongside a discrete multinomial logistic regression for the emission model.

The sequence of states for each patient, indicative of their pain experience, was deduced using the Viterbi algorithm. This algorithm employs the initial state probabilities, transition, and emission matrices to predict the most probable sequence of states. Here, a 0 in the Viterbi output designates a state of no pain, while a 1 indicates a state of pain. An illustrative example is provided in Fig. 5.15 for a 63-year-old female patient, showcasing her predicted pain state sequence throughout the experiment.

Given the unsupervised nature of our analysis, selecting the appropriate criteria for evaluating the model's performance emerged as a significant challenge. With each model representing a unique pain experience for individual patients, identifying a metric that could encapsulate the collective performance of all models was far from straightforward. Indeed, conducting an in-depth analysis of each patient and their respective model would have posed considerable difficulties. Thus, we explored various intuitive and efficient methods to summarise our findings.

Our initial method focused on verifying the accuracy of the constructed models by

[2] `https://github.com/Mogeng/IOHMM/`

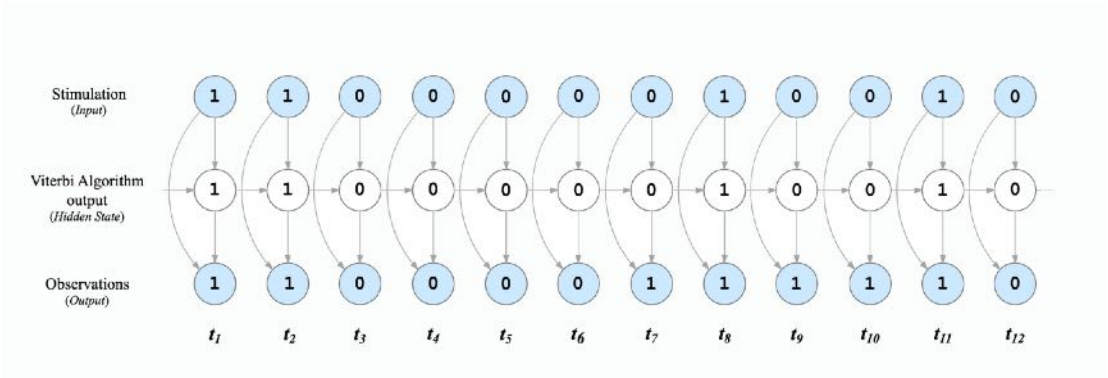


Figure 5.15: *Example of the Viterbi path for a patient (63-year-old woman), where 0 encodes the no pain state and 1 the pain state.*

examining the emission matrix. This matrix outlines the probability of a hidden state producing a specific observation at time t, based on the same hidden state at that time. It provides valuable insights into the model's learning efficacy. Specifically, a model is considered to have learned effectively when it can clearly differentiate between generating two possible observations (reaction, no reaction) based on the corresponding states. For example, if the "reaction" observation is produced with a probability of 0.8 by the "pain" state and 0.7 by the "no pain" state, it suggests that both states are likely to produce the "reaction" observation, without a clear distinction. Consequently, the "no reaction" observation would not be uniquely linked to any particular state. Figure 5.16 presents more detailed examples, where O_1 and O_2 denote the two possible observations, and S_1 and S_2 represent the two possible latent states. Three primary scenarios can be identified, two of which indicate incorrect configurations: the first scenario (a) illustrates the previously mentioned situation, while the second (b) shows a case where the hidden state S_2 ambiguously generates both observations O_1 and O_2 with similar probabilities. The third scenario (c) is considered acceptable, as each observation is more likely to be distinctly produced by its corresponding hidden state.

Under the assumptions of correctness, the outcomes were not entirely satisfactory, as only 43% of the models derived from video processing were found to be accurate. However, a more promising result was observed within biosignal processing, where the percentage of accurate models increased to 53%. This discrepancy could be attributed to the fact that many individuals do not exhibit a distinct facial expression for pain, often maintaining a neutral expression whether in the presence or absence of stimulation. Therefore, it can be deduced that employing biological signals offers a more effective analysis than relying on video signals alone. Figure 5.17 offers further insights into these findings, particularly highlighting the distribution of accurate models across different age and gender groups to explore potential variances between categories. Three age clusters were analysed, adhering to the original categorisation outlined in Section 5.1. Upon examination, it emerges that both younger and older women are less likely to express pain through facial expressions, whereas middle-age women are more prone to exhibiting distinguishable facial reactions. Conversely, no significant patterns or conclusions can be drawn regarding men.

Given the unsupervised nature of the experimental setup, a conventional accuracy

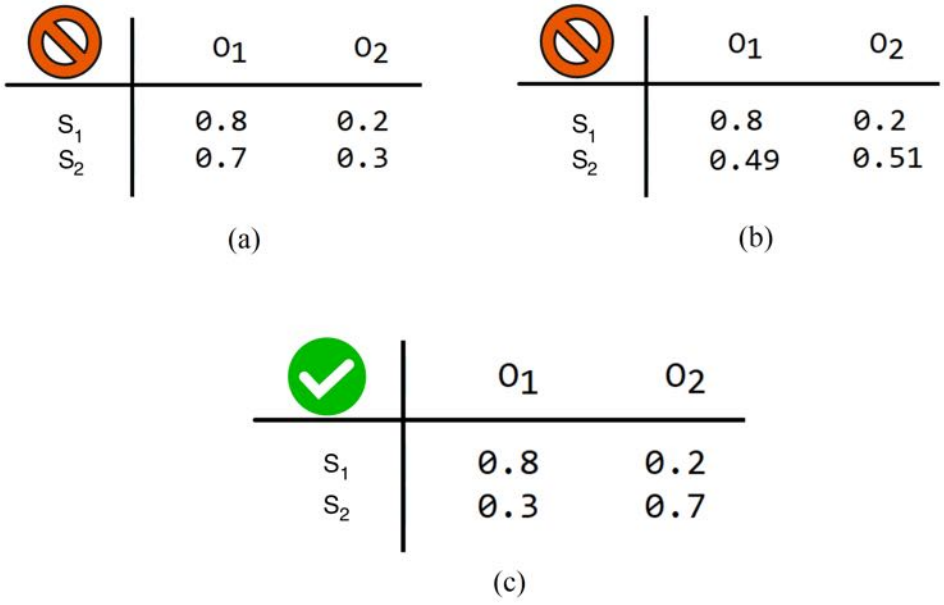


Figure 5.16: *Correctness check via emission matrix.* ***(a)*** *both hidden states are more likely to generate the same observation, without an actual distinction.* ***(b)*** *the hidden state S_2 generates the observation ambiguously, since both O_1 and O_2 are generated with similar probabilities.* ***(c)*** *the only acceptable configuration, as both observations are more likely to be generated distinctly by their respective hidden state.*

metric cannot be applied. However, to explore the relationship between stimuli and perceived pain, we introduce the concept of Hamming distance between the input vector, which encodes the stimuli, and the output vector generated by the Viterbi algorithm, indicative of perceived pain. Since the hidden states are not inherently meaningful—wherein 1 or 0 do not correspond to specific labels but are assigned arbitrarily across models—it's essential to methodically match hidden states to labels when calculating the Hamming distance. Given that the concluding 20 trials uniformly classify as BL1, suggesting an absence of pain, we infer that the final trial consistently reflects no pain as its hidden state. This premise ensures the proper encoding of hidden states: if the final hidden state is 0, then 0 represents the absence of pain; conversely, if it's 1, a swap is made so 1 always signifies pain, and 0 indicates its absence. This approach allows for an accurate computation of the Hamming distance for each patient.

Illustrated in Figure 5.18, the findings indicate that biosignal and video-based models demonstrate comparable outcomes, with an average Hamming distance of 0.396 for video and 0.35 for biosignals. Notably, video models display a broader distribution of distances, whereas biosignal distances tend to cluster more narrowly. A few outliers show distances under 0.1, with one exceptional instance registering a zero distance. These instances likely reflect specific input configurations that closely match the algorithm's method for deducing the hidden states' sequence. It's crucial to recognise that this measure does not serve as an accuracy indicator; thus, the goal is not to minimise these distances. In fact, achieving a universal distance of zero would imply a direct one-to-one mapping between stimulus and perceived pain, contradicting the premises of this

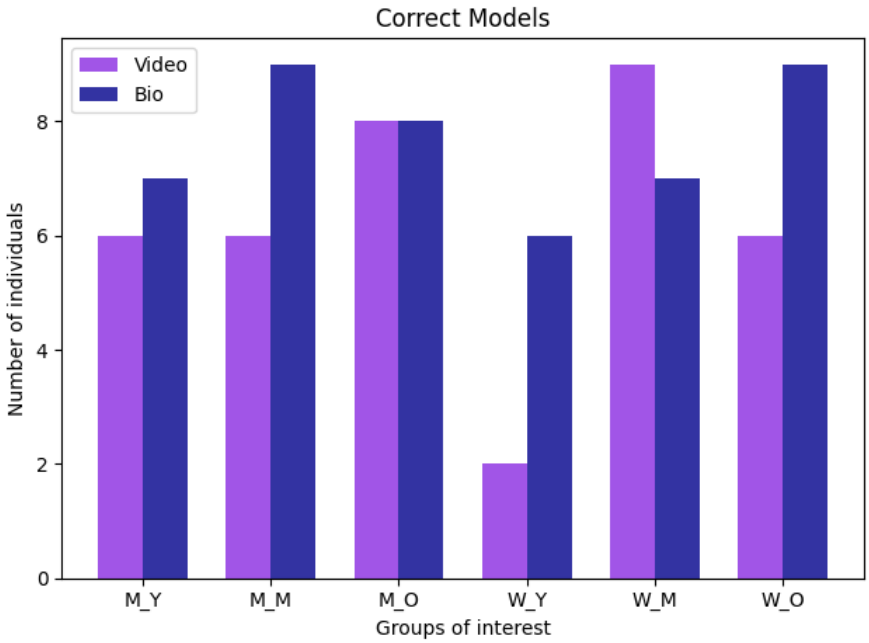


Figure 5.17: *Corrected models based on their reference age and gender clusters. M stands for man, while W stands for woman. Y represents younger patients, M is associated with middle-age patients and O refers to older patients.*

investigation. Conversely, a significant distance would suggest a complete disconnection between stimulus and perceived pain, which is clearly not true. These results, therefore, affirm a correlation between the noxious stimulus and perceived pain, while also acknowledging that other factors may influence this perception, as suggested by the pain literature.

The concluding segment of our in-depth analysis focuses on understanding the impact of individual patient experiences and the dynamics of pain perception through an examination of the transition matrix. Illustrated in Figure 5.19, which showcases the aggregated experiences of patients, we observe that the patterns emerging from both video and biosignal models are remarkably consistent. This observation was anticipated, given that a significant deviation would imply a stark contrast in the portrayal of pain, even though both models draw from the same experimental data. We infer that, generally, patients demonstrate a robust understanding of their pain condition: when experiencing a certain pain state, there's a pronounced tendency to remain within that state. This tendency for the persistence of pain could be consistent with the onset of chronic pain. Yet, the likelihood of this happening is moderate—marked at 0.67 for both models—indicating a level of variability and fluidity in the experience of pain.

5.5 Discussion

This chapter begins by examining the first discriminative model, which incorporates dynamics and temporality in a simplistic yet direct manner, focusing purely on sensory and perceptual aspects. Although this approach greatly simplifies the complexities in-

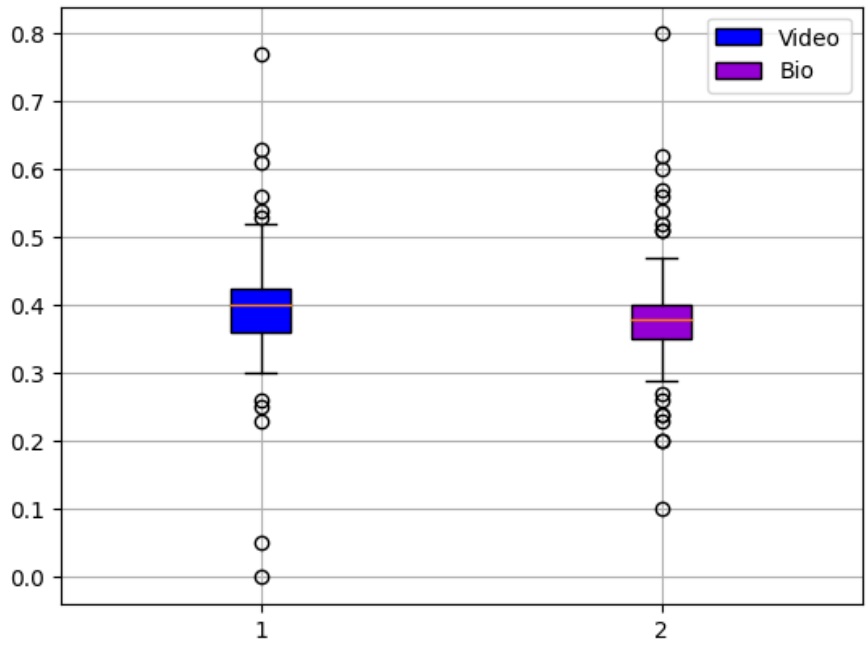


Figure 5.18: *Box plot of the Hamming distance between stimulation and hidden state (less is better).*

	no pain	pain
no pain	0.77	0.23
pain	0.33	0.67

(a)

	no pain	pain
no pain	0.81	0.19
pain	0.33	0.67

(b)

Figure 5.19: *Average transition matrix for all patients.* ***(a)*** *video models* ***(b)*** *biosignals models.*

volved, it proves to be effective in certain specific contexts. The chapter then transitions to a generative model that, while still maintaining a focus on perceptual aspects, introduces a Bayesian approach. This model not only addresses the perceptual dynamics more comprehensively but also lays the groundwork for subsequent models (Classes 2 and 3), which integrate actions within the context of active inference.

These developments highlight the evolution from purely perceptual models to more complex frameworks that incorporate dynamic interactions within biological systems.

Building on this foundation, we can explore the specific merits and applications of Class 1 models, particularly where the intricacies of cognitive processing and social frameworks do not predominantly influence the outcomes. This approach proves crucial in scenarios where the simplicity of direct sensory processing aligns better with the immediate needs of the system or organism involved, such as in early developmental stages or certain neurological states. Such models bypass the potential distortions that

advanced cognitive processes might introduce, thus safeguarding the integrity of the sensory information from the noise of over-interpretation.

For instance, scenarios involving the assessment of pain in neonates clearly demonstrate the essential role of perceptual models. These models effectively manage raw sensory data, providing quick and straightforward responses to stimuli. This tailored approach ensures that computational models remain both practical and relevant; a practical necessity in real contexts where the introduction of cognitive superstructures or other complex hierarchical levels would be unnecessarily complicated and not add significant value.

In essence, the use of Class 1 models accesses a fundamental layer of biological response, facilitating applications where fidelity to pure sensory input is paramount.

CHAPTER 6

Interlude: What is an emotion?

In this chapter, we introduce emotion theories to lay a clear foundation for our discussion. This groundwork is essential for the subsequent chapter, where we will explore a specific scenario in which affective aspects and pain closely interact. By understanding the theoretical landscape of emotions, we will be better equipped to analyse the intricate relationship between fear and pain, and how these emotional experiences influence and inform each other.

6.1 Emotion perspectives and theories

Emotion theories have a long history and have given rise to numerous competing theories. Providing a concise summary is essential to highlight the underlying assumptions, whether explicitly or implicitly, in computational theories of emotions, particularly within the affective computing field, from which pain research stems.

The endeavour to address the fundamental question raised by James (1884) has a long and tumultuous history in Western culture. Philosophers have been preoccupied with the nature of emotion since Socrates and his predecessors. Despite the development of rational thought as a discipline, emotions have always lingered in the background. Plato's depiction of the human psyche as consisting of three parts - rational thoughts, passions (now known as emotions), and appetites like hunger and sex - highlights the dominance of rational thought in controlling the passions and appetites, much like a charioteer steering two winged horses.

The essentialist perspective can be regarded as the genesis of the classical outlook on emotion, positing that emotions are biologically innate and universally experienced across different cultures. In contrast, Aristotle appears to have anticipated many of the contemporary appraisal theories. His examination of emotions involves a unique

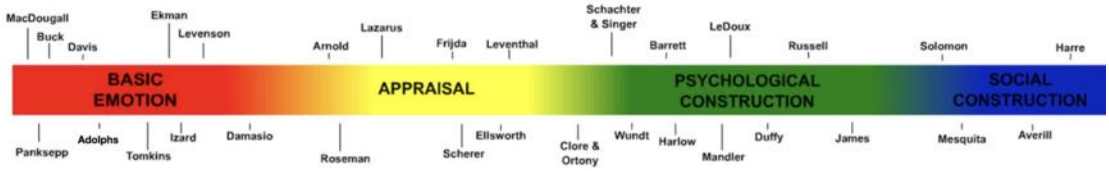


Figure 6.1: *Theories are loosely arranged along a spectrum, which could be defined in terms of a gradient of essentialism, the highest degree located at the left. Four "zones" are distinguished: (1) Basic Emotion Theories (BET); (2) Appraisal Theories; (3) Psychological Construction Theories; (4) Social Construction Theories. Adapted from Barrett (2016); Gross and Feldman Barrett (2011).*

cognitive aspect, a particular social setting, a behavioural inclination, and an acknowledgment of physical arousal, as stated by Solomon (2008).

The classical view

The *classical view* defines emotion as a distinct and autonomous faculty that arises from separate processes. Emotions are deemed to be fundamentally different phenomena from perceptions and cognitions, and each emotion, such as anger, sadness, or fear, is believed to be distinct from every other emotion, triggered by a unique mechanism.

This essentialist perspective holds that each emotion faculty has its inherent physical essence or "fingerprints" that set it apart from all other emotions. The modern versions of the classical view, including Basic Emotion Theory (BET) approaches and causal appraisal approaches, share a similar hypothesis concerning these emotional fingerprints.

The construction view

The psychological or social *construction view* does not view emotion as a distinct faculty with its separate mechanism. Emotion categories are not considered natural types. Instead, the primary assumption shared by all constructionist theories is that an emotion word, like "happiness", refers to a group of highly diverse instances, each adapted to a particular situation or context (Barrett, 2016; Gross and Feldman Barrett, 2011). Consequently, an emotion is not an entity with rigid boundaries in nature, but rather a category of instances. Instances within a category differ because each one is tailored to its environment, and thus there are no fixed Platonic emotion essences or fingerprints.

Appraisal theories

Figure 6.1 illustrates how appraisal theories can be positioned between the two opposing views. The figure is divided into four "zones" or color bands: (1) Basic Emotion Theories (BET), shown in red; (2) Appraisal Theories, shown in yellow; (3) Psychological Construction Theories, shown in green; and (4) Social Construction Theories, shown in blue. The theories in the red band and the leftmost part of the yellow band are more essentialist than those in the rightmost part of the yellow band and the green/blue bands, which are non-essentialist theories.

The most diverse range of essentialist beliefs can be found in the appraisal area, where classical appraisal theories (such as Arnold, 1960 and Lazarus, 1991) share several similar assumptions with BET. On the other hand, constitutive appraisal theories (for instance, Ortony and Turner (1990)) have more similarities with psychological construction theories (Barrett et al., 2016; Gross and Feldman Barrett, 2011) than with causal appraisal approaches. The latter approaches are, to some extent, more comparable to BET approaches.

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

Table 6.1: *Core assumptions of four emotion perspectives. Adapted from Gross and Feldman Barrett (2011).*

6.2 Fear and the theory of constructed emotion: a brief tour

Indeed, the experience of fear is a complex phenomenon (cfr. left panel of Fig. 6.2) binding together the external perception of the world, the internal perception of the body, conceptual knowledge about fear emotion itself, and experience (e.g., previously

experienced situation or episodes of fear). Differently from previous work, we first frame the problem addressed here in a solid framework, namely the theory of constructed emotion (which is summarised at a glance in the right panel of Fig. 6.2). In such theory, terms like "emotion" and "affect" (and thus possibly related measurements) have a clear meaning at different levels of explanation (conceptual level and core affect/interoceptive level, respectively). The setup of the experiment and measured subjects' behaviour can thus be accordingly interpreted.

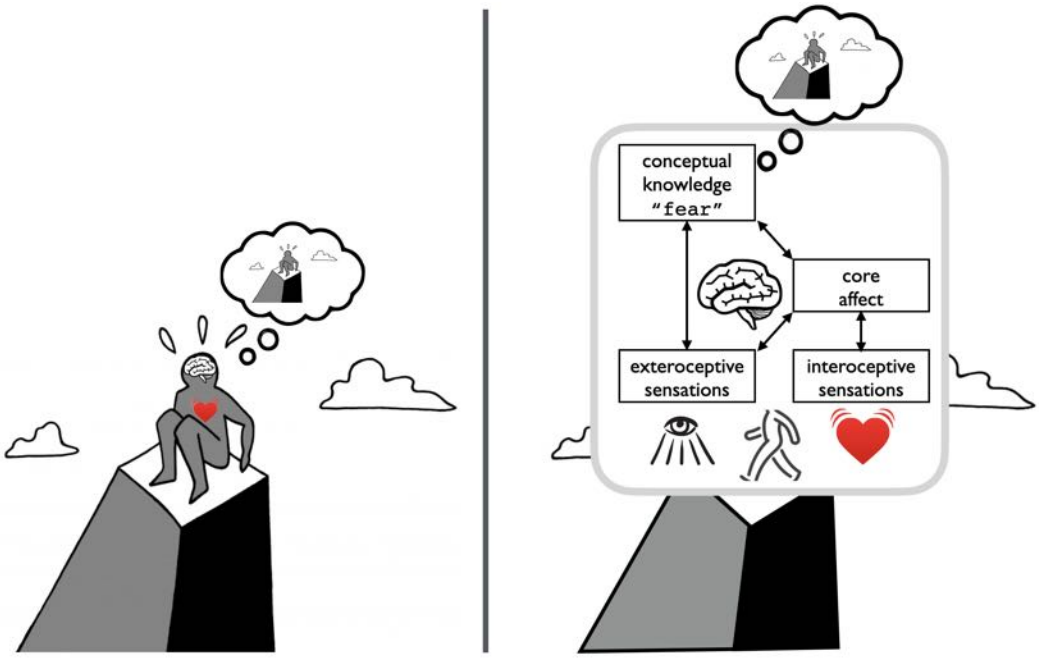


Figure 6.2: *Fear as situated conceptualisation.* Left panel*: the actual experience of fear bounding together the external perception of the world, the internal perception of the body, conceptual knowledge about emotion, and past experience.* Right panel*: the theoretical view of fear as a categorical emotion constructed through a conceptual act, the result of the brain's active, ongoing predictive processing endeavour (see text for details)*

In the perspective that motivates the present work, emotions (e.g., fear) are the result of situated conceptualisations (whose overall process is depicted in the right panel of Fig. 6.2) constructed from affect. In this view, emotional events are specific instances of affect that are linked to the immediate situation and involve intentions to act (Barrett, 2017b; Barrett et al., 2015).

Here, differently from what proposed in BET (Tomkins, 1962; Izard, 1993; Ekman et al., 1969; Panksepp, 2004), emotions labelled by words such as "fear", "anger", and "surprise" are not natural kinds, Platonic essences wired in the brain, but abstract *ad hoc* categories (Barrett and Satpute, 2019). A *category* can be defined as a population of events or objects that are treated as similar because they all serve a particular goal in some context (Barrett, 2017a). For example, perceived fear is the result of categorising the current situation (sensations, context) as fearful when it contains features similar to those of previous situations that have been experienced as fearful (Lindquist and Barrett, 2008).

A category has a mental representation, a *concept*, namely the population of rep-

resentations that correspond to the category's events or objects. Moment by moment, conceptual representations are tested against incoming sensory evidence - from the external world and from the body - to categorise them according to experience, anticipate body needs and prepare to satisfy those needs before they arise, a mechanism defined allostasis (Schulkin and Sterling, 2019). This dynamics is represented in the right panel of Fig. 6.2 by bidirectional arrows connecting the different components: when the information flows from concept to sensations, the agent is generating predictions; otherwise, the information flows from sensations to concept, representing agent's inference or categorisation.

Fear, for instance, as any other emotion is nothing but an abstract, *ad hoc* category, mentally represented by a situated conceptualisation, or more precisely a set or a manifold of situated conceptualisations as summarised in Figure 6.3.

Just like any other form of categorisation it is a construction across many levels of abstraction (Figure 6.4).

In such endeavour, *interoceptive sensations* from the body play a cogent role because weighing which parts of the world are worth caring about at the moment. Without them, an actual agent would not appraise relevant features of the physical surroundings. Interoception is fundamental to construct the pivoting psychological primitive (Barrett and Bliss-Moreau, 2009) named "core affect". *Core affect* can be described as a state of pleasure or displeasure, named *valence*, with some degree of *arousal*. Together, valence and arousal form a unified, continuous state-space. It is referred to as "core" because it is grounded in the *internal milieu*, an integrated sensory representation of the physiological state of the body: the somatovisceral, kinesthetic, proprioceptive, and neurochemical fluctuations that take place within the core of the body.

Core affect is realised by integrating incoming sensory information from the external world, i.e. *exteroceptive sensations*, with internal, interoceptive information from the body (see Fig. 6.2).

In modern psychological usage, "affect" refers to the mental counterpart of internal bodily representations associated with emotions and actions, that involve some degree of motivation, intensity, and even personality dispositions. In the science of emotion, "affect" is a general term that has come to mean anything emotional. A cautious term, it allows reference to something's effect or someone's internal state without specifying exactly what kind of an effect or state it is. This way researchers can talk about emotion in a theory-neutral way. Under such circumstances, if one observes the neural reference space of core affect (as presented in Figure 6.5) this might be considered as the neural underpinning of emotion.

However, this is not the case. This neural reference space can be subdivided into two related functional networks (Barrett and Bliss-Moreau, 2009).

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

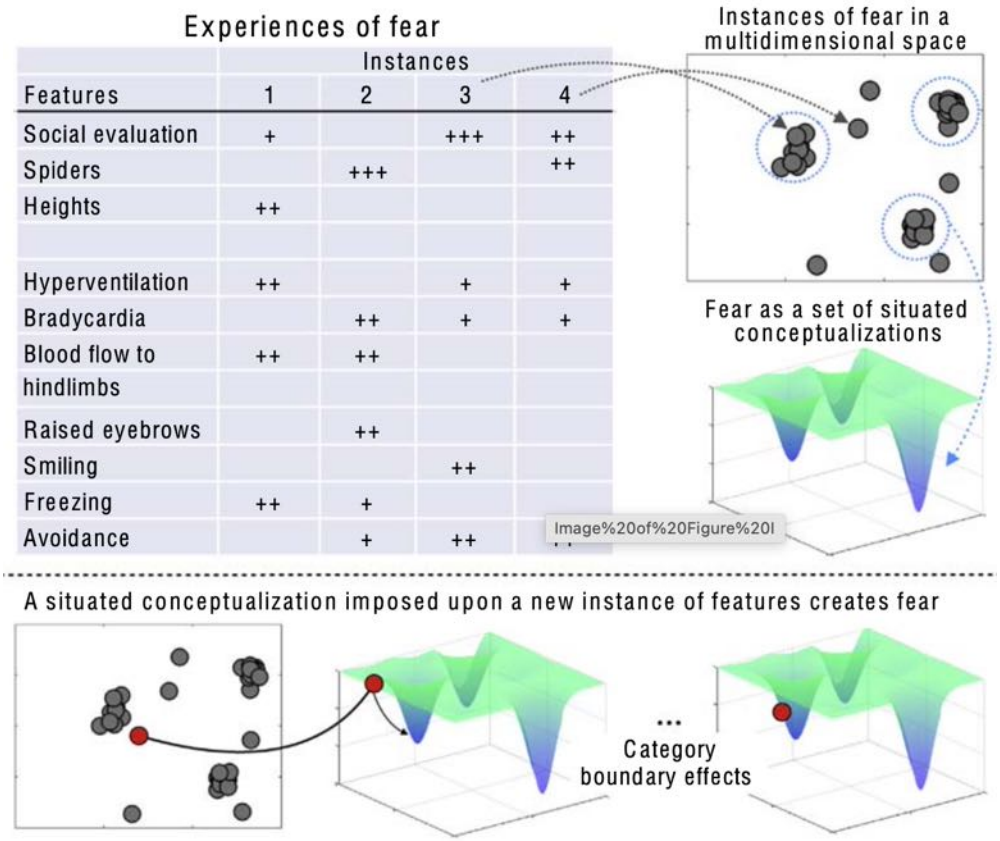


Figure 6.3: *Conceptualisation of fear. The table outlines four hypothetical instances of fear that involve a set of features that vary in kind (along rows) and intensity (number of +s). For example, Instance 1 may involve rock climbing (heights), being watched (social evaluation), and physiological and behavioural responses (hyperventilation and freezing). Instance 2 may involve encountering a tarantula while hiking, bradycardia, redistribution of blood to the legs, and eye widening to increase visual input. Instances can be represented in a high-dimensional feature space (simplified to two dimensions for the sake of illustration). Situated conceptualisations are modeled as a landscape of attractor basins. Grouping together the full collection of variable instances as fear is, by definition, an abstract category that refers to the representational space of fear. The abstract representation of those instances as all belonging to the same category of fear may differ between individuals and may be uniquely human. Adapted from Satpute and Lindquist (2019).*

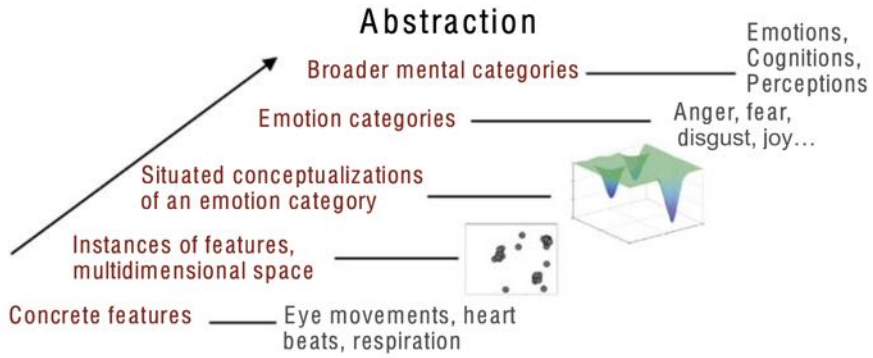


Figure 6.4: *The nested hierarchy of increasing abstraction from broad and abstract categories of mental experience to concrete sensory and motor features that are associated with those mental states. An emotional experience manifests when there is resonance across levels, that is, concrete features are made meaningful as a conceptualisation of a discrete emotion category, in a given context. Without higher levels making meaning of lower levels, elemental concrete features (e.g., tachycardia), or combinations of features (e.g., tachycardia and hindlimb locomotion), are not necessarily a manifestation of an emotional experience. Without top-down categories and conceptualisations an instance of features may be experienced in alternative ways, for example, as merely a behavior (e.g., running), visceral sensation (e.g., stomach sinking), or general affective feeling (e.g., displeasure). Adapted from Satpute and Lindquist (2019).*

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

To sum up, some parts of affective circuitry are strongly interconnected with sensory cortical areas. Others are strongly interconnected with areas that direct the autonomic and hormonal responses to regulate the homeostatic state of the body. The strongly re-entrant nature of neural activity makes it difficult to derive simple cause and effect relationships between the brain and the body, or between sensory and affective processing Barrett and Bliss-Moreau (2009).

The key concept here is that the circuitry within the neural reference space for core affect binds sensory information from the external world to sensory information from

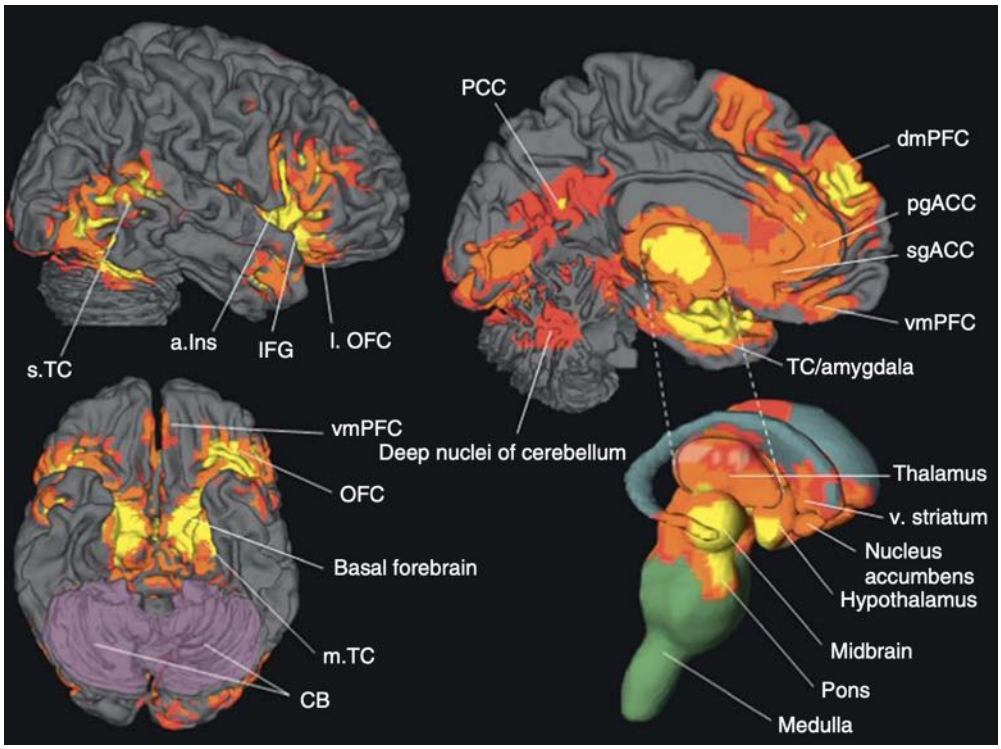


Figure 6.5: *Neural reference space for core affect*

165 neuroimaging studies of emotion (58 using PET and 107 using fMRI) summarised in a multilevel meta-analysis to produce the observed neural reference space for emotion. These areas include (from top left, clockwise) anterior insula (aIns), lateral OFC (lOFC), pregenual cingulate cortex (pgACC), subgenual cingulate cortex (sgACC), ventral medial prefrontal cortex (vmPFC), temporal cortex/amygdala (TC/Amygdala), thalamus, ventral striatrum (v Striatum), nucleus accumbens, hypothalamus, midbrain, pons, medulla, OFC, and basal forebrain. Other areas shown in this figure (e.g., inferior frontal gyrus (IFG), superior temporal cortex (sTC), dorsal medial prefrontal cortex (dmPFC), posterior cingulate cortex (PCC), medial temporal cortex (mTC), and cerebellum (CB)) relate to other psychological processes involved with emotion perception and experience. From Barrett and Bliss-Moreau (2009).

the body so that every mental state is intrinsically infused with affective content.

When core affect is in the background of consciousness, it is perceived as a property of the world, rather than as the person's reaction to it. It is under these circumstances that scientists usually refer to affect as "unconscious" (we have another sip of Barolo because it tastes so good). When core affect is in the foreground of consciousness, it is experienced as a personal reaction to the world. It is at these times that feelings that can be described as pleasant or unpleasant content with some degree of arousal can serve as information for making explicit judgments and decisions. In this case, such experience might be categorised as that of feeling an emotion.

The result is a mental state that can be used to safely navigate the world by predicting, for instance, reward and threat. Thus, by no means affect can be equated to emotion.

In this perspective, in every waking moment, brains function as predictive machines that run on concepts - predictive internal models - to give sensations, either exteroceptive or interoceptive, meaning. Note that, under such circumstances, there is no specific difference between emotion, vision, or audition: when we focus on some of those sensations that are exquisitely interoceptive, the resulting experience can be an instance of emotion (Barrett et al., 2015; Barrett, 2017b).

CHAPTER 7

Affective Alarms: Unpacking Fear Generalisation in Pain Contexts

Given the challenges outlined in the previous chapter regarding the comprehensive implementation and validation of a complete pain model — one that encompasses all proposed dimensions — we introduce a set of minimal model implementations. These models specifically focus on the affective and emotional aspects of pain, tested against data collected from a fear generalisation experiment. This approach allows for an exploratory examination of pain's complex nature within the constraints of current research methodologies and available data.

To achieve this, we will begin with a brief introduction to the methodologies employed in affective computing and discuss how our approach fits within the broader landscape of the literature. This context will set the stage for our explorations and contribute to a nuanced understanding of our minimal model implementations, particularly in their capacity to address affective and emotional dimensions of pain in the context of fear generalisation experiments.

Subsequently, we will delve into the experimental setting that served as the foundation for data collection, effectively outlining the boundaries within which the model implementation was developed. This exploration will provide critical insights into the conditions and parameters that shaped our approach, ensuring a coherent integration of the model with the empirical data gathered.

We will also explore how the setting from which the data were derived allowed us to model another component often overlooked in computational models of pain: action. Specifically, this is represented by eye movements, indicating the choice to focus on one element of the scene over another. This aspect introduces a dynamic layer to our understanding of pain, bridging the gap between internal experiences and observable, external actions.

More precisely, in this chapter we

- propose a simple latent variable model accounting for fear for pain and its associated variables incorporating both interoceptive and exteroceptive signals, and the elementary subject's action of communicating their shock expectancy (see Fig. 7.4);
- we discuss how the more complex Active Inference framework could be adopted for devising a Class 2 model;
- eventually, we analyse some consequences on the agent's visual/attentive behavioural actions, when striving to balance exploitation/exploration as posited by Active Inference; we show that on such basis a link between attention and social anxiety can be detected .

Data-based and simulation-based implementation strategies are detailed as well as the outcomes achieved.

7.1 How fear spreads: the mechanisms of generalisation

If a vicious dog has painfully bitten you, you might have acquired fear of all dogs therefrom conceptualised as those harmful barking animals with sharp teeth and claws, four legs, and a tail. Under such circumstances, regrettably, you are contending with a fear (over)generalisation problem.

Fear, a primal emotion that persists throughout the animal kingdom (see Carew et al. (1981), LeDoux (2012), Adolphs (2013)), requires flexible assessment of threatening stimuli for survival in the wild. This assessment involves processing, integrating, and synthesising information gathered from multiple sensory modalities. Animals, faced with non-identical aversive experiences, must generalise fear from past encounters to future situations that share a reasonable degree of similarity with the original event. Like other memory-related processes, generalisation is subject to modulation by various intrinsic factors such as internal states (estrous and circadian cycles, prior experiences, genetic background, and gender differences). External factors, including the type and intensity of aversive stimulation, early-life stress, as well as the saliency of specific elements in the environment, also exert influence on generalisation. Finally, generalisation is sensitive to the passage of time, as memories naturally degrade in both precision and strength over time. While numerous variables impact the generalisation of fear, developing a comprehensive neurobiological framework with robust explanatory power has proven to be a formidable challenge. Nevertheless, recent studies have started to provide intriguing new insights into this intricate process.

The process of fear memory generalisation is a neurobiological adaptation that enhances survival in intricate and ever-changing surroundings. When an animal encounters a potential threat, it needs to decide on an appropriate defensive reaction using prior experiences that are not identical. This involves evaluating various cues and contextual details that could indicate safety or danger.

Fear generalisation (FG) describes the phenomenon that learnt fear is not restricted to those exact stimuli with which an aversive experience was originally paired (the

Figure 7.1: *Fear learning and generalisation. An aversive episode pairs the perceptual stimulus (conditioned stimulus CS) of a black dog with a painful bite (unconditioned stimulus, US). Learnt fear can subsequently spread over a gradient of harmless stimuli (generalisation stimuli, GSs) more or less similar to the original CS*

specific wicked black beast) but it spreads to perceptually or conceptually similar ones (all black dogs or even all dogs, see Fig. 7.1) (Dymond et al., 2015).

Clearly, the ability to learn which stimuli in the environment signal threat has an important adaptive advantage (to initiate appropriate defensive responses); a hallmark of human cognition is the ability to extract conceptual knowledge from a learning episode (Tenenbaum et al., 2011). Yet, excessive generalisation may become maladaptive and pathological. Crucially, the overgeneralisation of fear to harmless stimuli or contexts might even turn into a burden to daily life and characteristic of several anxiety disorders (Dunsmoor and Paz, 2015).

7.2 Adaptive vs Maladaptive Fear Generalisation

Adaptive fear generalisation refers to the ability of an organism to extend its fear response from a specific, aversive experience to similar situations or stimuli. This ability is crucial for survival because it allows individuals to anticipate and respond to potential threats, even if they have not encountered them. For example, if an animal learns to fear a particular predator in a specific context, adaptive generalisation allows it to be cautious in similar environments or when encountering animals with similar characteristics.

While adaptive fear generalisation is critical for survival, maladaptive fear generalisation occurs when the fear response extends to situations or stimuli that are not genuinely threatening. This phenomenon can lead to anxiety disorders, phobias, and other mental health issues. In cases of maladaptive fear generalisation, the individual's brain fails to differentiate between safe and genuinely threatening situations, resulting in excessive and inappropriate fear responses.

For example, consider a person who experienced a car accident on a rainy day. If they develop a phobia of rain and refuse to leave their home when it's raining, this represents maladaptive fear generalisation. Their fear has extended to a situation (rain) that is not inherently dangerous, and it impairs their ability to lead a normal life.

Understanding the balance between adaptive and maladaptive fear generalisation is a complex and critical aspect of psychology and neuroscience. Researchers are continually investigating the neural mechanisms and psychological factors that contribute

to these processes. By gaining a deeper understanding of fear generalisation, we can develop more effective treatments for individuals struggling with anxiety disorders and phobias, helping them regain control over their lives and their emotional responses.

From an ethological perspective, behaviours that contribute to an organism's survival are considered adaptive, while those that go against the imperative of self-preservation are labelled as maladaptive (Johnson et al., 1992; McEwen, 1998; Cooper and Blumstein, 2015). Nevertheless, it's important to recognise several important nuances in this classification.

First and foremost, the environmental context in which a generalised fear response occurs plays a pivotal role. A behaviour that proves advantageous in one environment might become detrimental in another. Take, for instance, rodents in regions with a high prevalence of predators. In such areas, it's adaptive for them to exhibit increased defensive behaviours and reduce their foraging activities. However, deploying an exaggerated defensive response in environments where the threat level is low becomes maladaptive. In such cases, it needlessly compromises their ability to acquire resources and maintain allostasis, the array of adaptive processes that uphold homeostasis (Fanselow, 1994; McEwen, 1998; Blanchard and Blanchard, 2008).

This same principle applies not only to laboratory mices but also to humans. In experiments, subjects conditioned to fear a particular context or cue may generalise that fear response to different contexts or cues (Kaczkurkin et al., 2017), although there are exceptions (Elzinga and Bremner, 2002). The perception of cues and environments exists on a continuous spectrum, and maladaptive fear generalisation occurs when an abnormal stimulus-response pattern emerges, remember the example of having a fear of rain after an accident on a rainy day, leading to defensive behaviours in environments or in response to cues that have never been directly associated with danger or threat.

Fear learning and its generalisation effects are usually investigated using the fear-conditioning paradigm (Dymond et al., 2015; Ghirlanda and Enquist, 2003). Learning is fostered by pairing a perceptual stimulus with an aversive one (e.g., an electric shock) and then the extent of subject's generalisation to stimuli perceptually similar to the original one is assessed via subject's shock expectancy and behavioural/physiological measures. To such end, extensive research has been conducted by using geometric shape stimuli (see Figure 7.2).

Fear learning *in vivo*, however, hardly involves such simple sensory cues; thus, in order to increase the ecological validity of FG studies, recent works (Ahrens et al., 2016; Roesmann et al., 2020; Reutter and Gamer, 2022) have considered faces as suitable targets. Notably, Reutter and Gamer (2022) have provided experimental evidence that the extent of explicit fear generalisation is related to individual patterns of attentional deployment. Precisely, by using visual facial stimuli, eye-tracked participants who dwelled on distinguishing facial features faster and for longer periods were likely to exhibit less fear generalisation. Their analyses though were based on classic measurements (latency of the first fixation, dwell time, etc.); gaze dynamics of viewing behaviour was only indirectly considered.

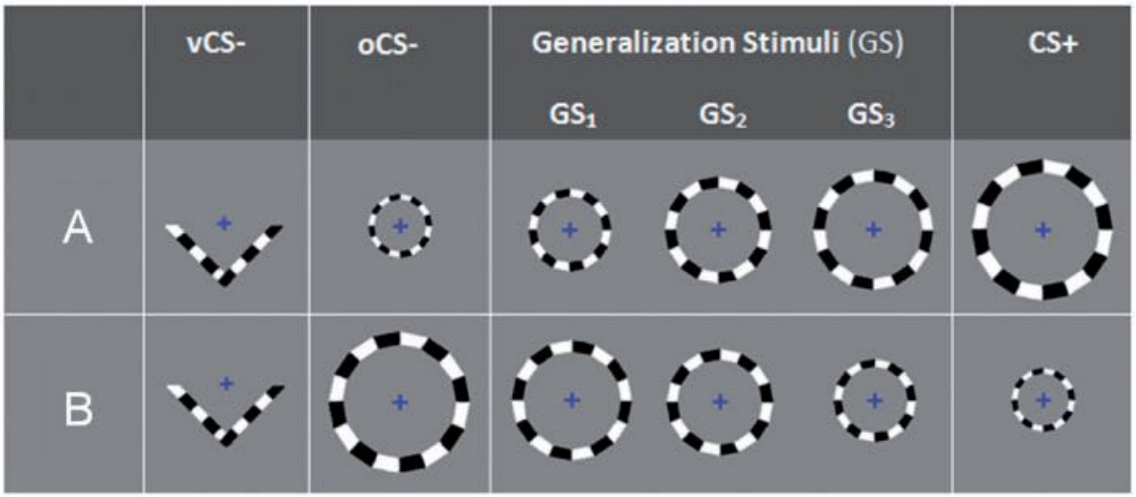


Figure 7.2: *Example of geometric shapes used for fear generalisation tasks (Figure from (Lissek et al., 2014))*

7.3 Description of the experimental setting

The experimental setting for this work relies on data collected by Reutter and Gamer (2023). The experiment focused on diagnosing facial features and exploring fear generalisation.

More in detail, to assess that the visual preferences for distinguishing stimulus aspects are associated with reduced fear generalisation and to investigate the fact that selective attention to distinguishing facial features may also increase discriminability between faces and thus reduce generalisation of fear, Reutter and Gamer (2023) developed a set of facial stimuli such that pairs of faces could either be distinguished by looking into the eyes or the region around mouth and nose, respectively. These pairs were then employed as *conditioned stimuli* (CS): those associated with a subsequent shock are referred to as CS+ while those that are not followed by a shock are denoted as CS-. After an initial phase of fear acquisition, four different intermediate morphs from CS+ to CS- were presented to test generalisation of fear (Figure 7.3).

Throughout the whole experiment, shock expectancy ratings, heart rate, electrodermal activity, and eye movements were measured; to analyse how a person generalises, they calculated a linear deviation score (LDS) that provides an efficient way to compare different results between people. Their main assumption was that individuals who allocate greater attention towards diagnostic facial features would exhibit fear generalisation gradients with a more pronounced curvature, thus meaning less generalisation.

7.3.1 Experiment description

To experiment, a total of 44 individuals took part (37 females and 7 males). Before the study, all participants confirmed that they did not have any mental or neurological conditions and scored low on the Social Interaction Anxiety Scale (SIAS), indicating a low level of social anxiety.

All participants in the experiment were asked to view images of human faces dis-

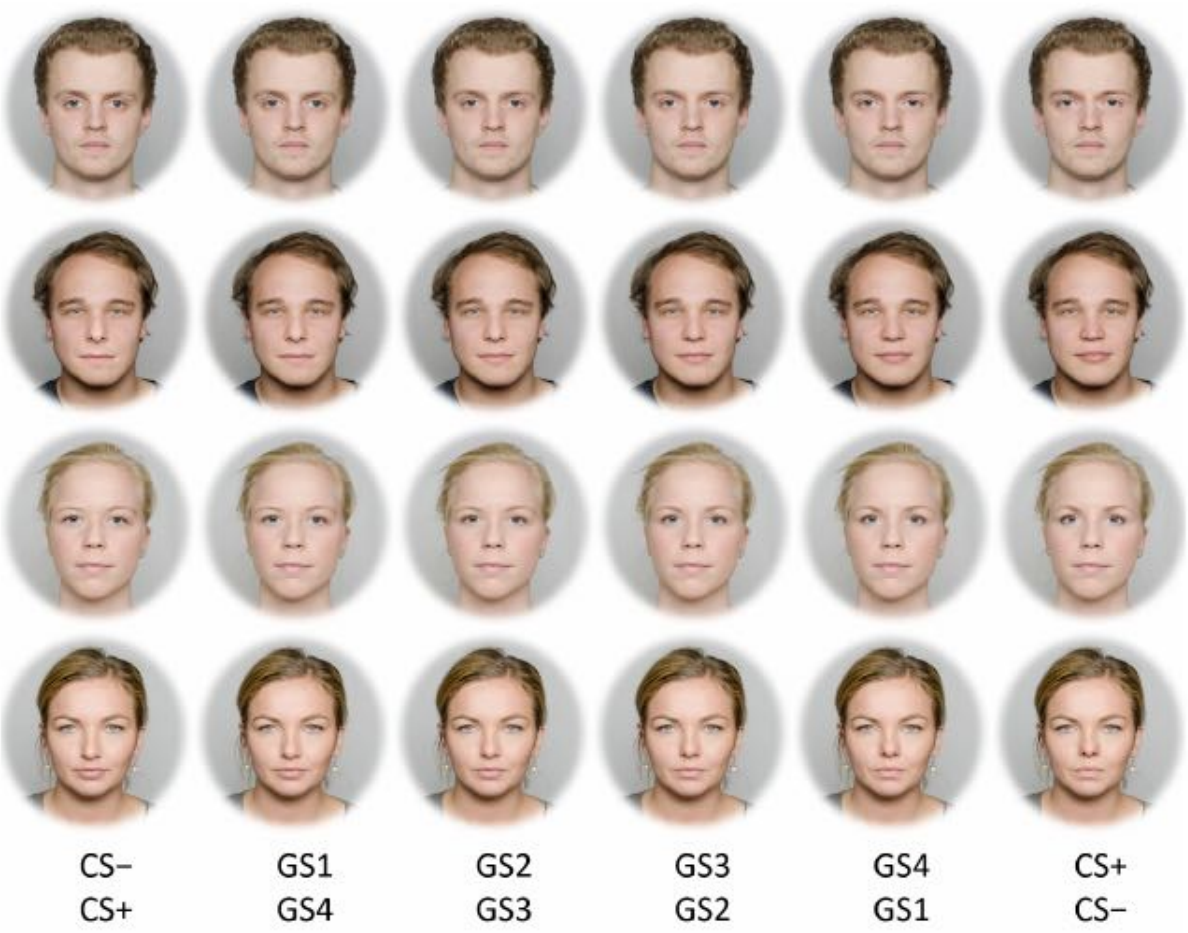


Figure 7.3: *Visual stimuli, from CS+ to CS- with all GSs intermediate simuli*

played on a computer screen and subsequently an appropriate feedback. In particular, the images of faces, described in the previous section, may or may not be followed by an electric shock (*unconditioned stimulus* - US) applied to the back of the left hand. After each stimulus presentation, participants were asked to rate on a scale from 1 to 5 whether they expected the shock (US). To provide further clarification, the participants completed a discrimination task to ensure that they were capable of distinguishing between the visual stimuli that were utilised as conditioned stimuli in the fear generalisation paradigm.

Fear generalisation task

The fear generalisation experiment consisted of 3 phases:

- habituation;
- fear acquisition;
- fear generalisation

During the initial phase, each of the four chosen facial stimuli (consisting of one male and one female pair that differed either in the eye region or in the mouth/nose region) was displayed four times without being followed by any shocks. In the second phase, the two stimuli that were assigned to denote the CS+ were reinforced in 75% of the cases (meaning that only in 75% of the cases, the stimuli were followed by a shock), for a total of 32 trials. In the fear generalisation phase, four intermediate stimuli between CS+ and CS- were also presented for each pair of stimuli.

The anxiety generalisation task included a total of 160 trials. Each test consisted of presenting a face for 6 seconds. After 4 seconds, a rating indicator will appear at the bottom of the screen for 2 seconds. Subjects were asked to indicate the perceived likelihood of electrotactile stimulation at the end of the test on a 5-point scale (1 = no shock, 3 = uncertain, 5 = shock certain), but were not informed of the different phases or eventualities. The painful stimulation was applied or omitted 5.85 seconds after the onset of the stimulus. If subjects pressed too early or too late, the answer was marked as invalid. A fixation cross was drawn in the center of the screen during the interval between trials, the duration of which followed an even distribution between 2 and 4 seconds to minimise anticipation effects.

The main hypothesis alongside the experiment is that visual exploration patterns are related to individual differences in fear generalisation. In particular, people are expected to pay more attention to diagnostic regions (e.g. regions of the faces that morph) and they should show less fear generalisation compared to those that can't really focus on the right patches and instead look over faces more sparingly. The generalisation of fear is determined by evaluating individual expected discharge values and by monitoring pupillary responses, electrocardiography, and skin conductance. In particular, they observed an increase in pupil dilation and larger decelerative heart rate responses as a function of threat level.

7.3.2 Data recording

Shock expectancy ratings

During each trial, participants' behavioural responses were recorded using a QWERTY keyboard with only five response keys available (Space = 1, L = 2, Semicolon = 3, Apostrophe = 4, and Right Shift = 5). The shock expectancy data for two subjects were not included in the analysis because over 25% of their ratings were absent during the generalisation phase.

Eye-tracking

An EyeLink 1000 system was used to record eye movements and the change in pupil diameter. Participants' right eye was monitored at a rate of 1000 Hz, except for one individual who had their left eye monitored due to calibration problems. Calibration was carried out before the habituation and generalisation stages. However, eye-tracking was not conducted for one participant due to inadequate calibration accuracy.

Pupillary Responses

Pupil diameter and eye movement were simultaneously measured and recorded. Blinks were interpolated linearly, and the trial-level time series were down-sampled to 100 Hz. A low-pass filter with a roll-off of 12 dB/octave and a cutoff frequency of 2 Hz was used to filter the data. The values were then converted from arbitrary units to millimeters using the method described by Hayes and Petrov (2016). Baseline correction was performed by referencing the last second before stimulus onset. Finally, pupillary responses were calculated by averaging the change in pupil diameter between 500 and 4000 ms after stimulus onset.

Electrocadiography

To track the heart activity, three ECG electrodes were attached underneath the right clavicle and the left costal arch, with a ground electrode at the right costal arch. The data was recorded at a sampling rate of 500 Hz. Using R 3.5.1, the maxima of cyclic electrocardiac activity (R-peaks) were detected semi-automatically, with an option for manual editing, based on their amplitude. The heart rate in beats per minute was calculated from the resulting R-R intervals. However, data from one participant were excluded due to excessive extrasystoles. The remaining artifacts were interpolated using the average heart rate change of adjacent R-R intervals (occurring 6 times in the remaining data). To account for inter individual differences in tonic heart rate, change scores were calculated relative to the last second before stimulus onset, in bins of 0.5 seconds, up to 5.5 seconds after stimulus onset (similar to pupillary responses).

Social Anxiety

All participants in the study were required to complete the Social Interaction Anxiety Scale (SIAS) questionnaire (Heimberg et al., 1992). This questionnaire is designed to measure a person's fear of social interaction and evaluate the emotional aspects of their anxiety response, rather than their general social apprehensiveness or concern

for others' opinions. The SIAS originally consisted of 19 questions, but an additional question has been added, and participants have to rate how much each question relates to them on a 5-point scale. The final scores range from 0 to 80.

Usually, the general threshold for identifying social phobia/anxiety is around 30. It is worth noting that with this approach, we identified only six individuals with anxiety, so we decided to not apply a division in non-anxiety subjects vs anxiety subjects but we use this data as a value.

Shock expectancy rating

The subject labelled each trial with a shock expectancy score ranging from 1 to 5, indicating their perception of how likely they were to receive an electrotactile stimulus. For classification techniques, a threshold of 3 was used to identify scores of 3 or greater as "high likelihood" and scores less than 3 as "low likelihood". However, it should be noted that this approach results in an imbalanced dataset, and thus, it's crucial to consider the imbalance during model training.

Electrotactile stimulus

In addition to the shock expectancy score, each trial was marked with a binary value indicating whether the stimulation was administered or not. The stimulation was delivered using a Wasp Electrode on the back of the left hand and consisted of three sets of seven trains of 2 ms pulses, with a 48 ms interval between them.

7.4 Bridging pain and fear

Individuals construct their concept of pain through both internal and external sensory signals. Recent cognitive science frameworks, particularly those focused on the Bayesian brain and predictive coding, presented in the previous chapter, suggest the brain functions as a predictive machine. This concept, traditionally applied to external perceptions and movements, has been expanded through interoceptive inference (Seth, 2013) (Seth and Friston, 2016), proposing the brain predicts internal states from sensory data to understand bodily sensations and emotions. This perspective aligns with appraisal theories on emotion origins but offers new insights into bodily awareness and self-perception.

An important factor to take into account is modelling attention, as the subjects participating in the experiment described above, place their attention in different areas of the visual stimulus, changing their interoceptive sensations.

The idea is to model the concept of pain using a Probabilistic Generative Model like the one described in Figure 7.5. Starting with interoceptive cues, such as heart rate, electrodermal activity, and change in pupil diameter size, and exteroceptive cues, such as eye movement and where the subject turns its gaze, the goal is to create a model that, by making inferences about the interoceptive or exteroceptive manifestations (or only one of them), can generate the concept of pain, which can be seen as a latent state. In all of this there is a prior belief within each subject, which is based on their level of social anxiety. The concept of pain modelled as latent space relies on Conceptual Act Theory (Barrett, 2014), whereby our perceptions are an active process of continuously

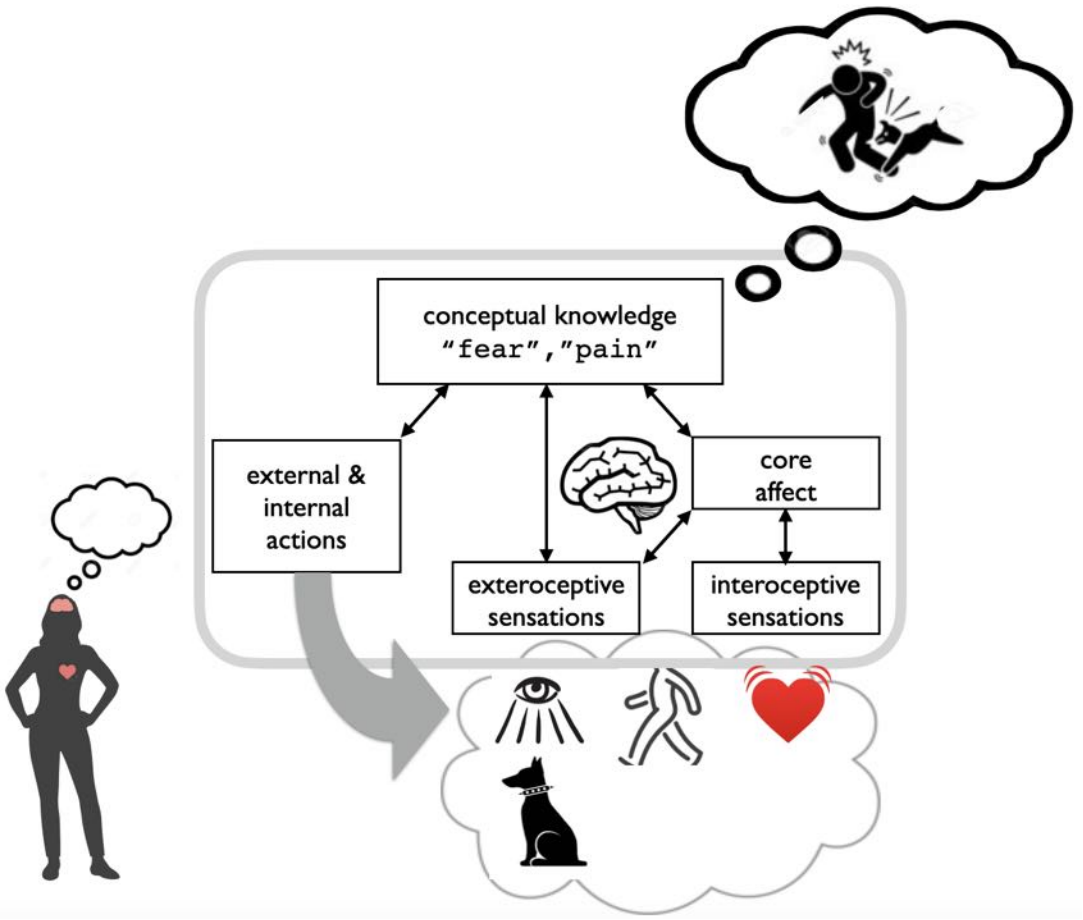


Figure 7.4: *The agent process model in the context of fear generalisation. As posited by situated conceptualisation theory, the fear concept interacts with pain at the conceptual level: both can be represented as categories constructed through a conceptual act, the result of brain's active, ongoing predictive processing endeavour. The kind of interaction and the pain avoidance tendency, constrain and trigger agent actions/controls towards their body and the external environment.*

constructing meaning from sensory data, thanks to our prior experiences and interaction with the external world.

Another aspect to consider is how the subject categorises their experience of pain; to do this they shift their focus of attention to different areas of the visual stimulus, relying on their interoceptive stimuli. In particular, a subject's level of social anxiety acts as an attentional bias in that information about interoceptive stimuli gains more importance than exteroceptive stimuli (Pineles and Mineka, 2005).

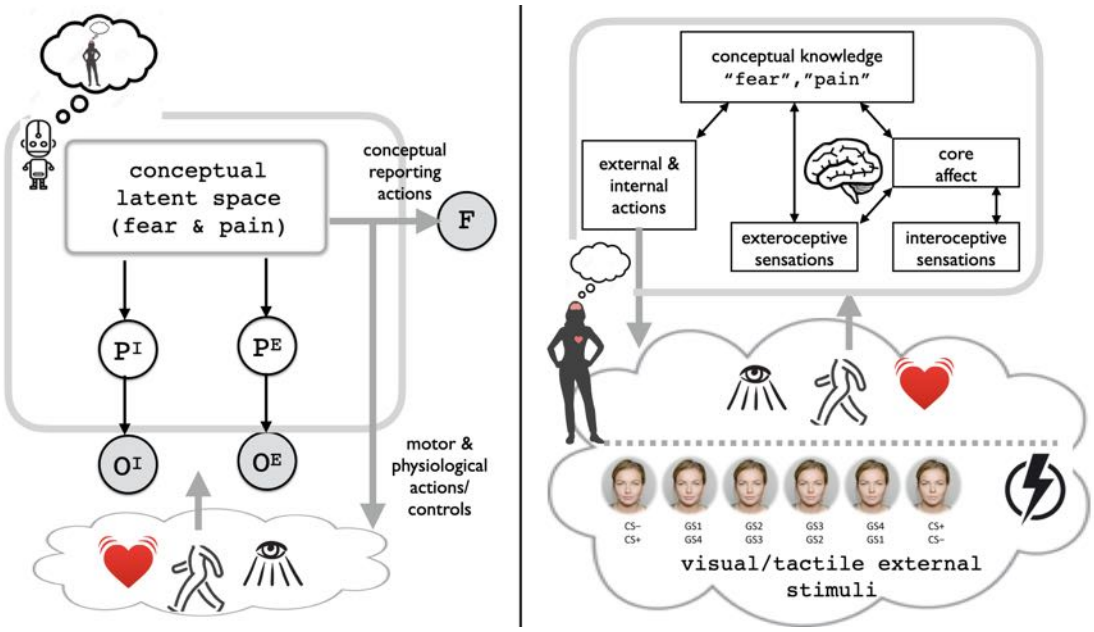


Figure 7.5: *Similar to Figure 5.8, the agent's process model (right) and the observer's model of the agent (left) are presented. The agent's model contextualises the general FG model in Figure 7.4, to the Reutter and Gamer (2023) experimental setting. The observer's relies on a simplified model of the agent where a latent space shapes the manifold where fear and pain conceptualisations interact. After learning, the manifold feeds on the perceptual (exteroception and interoception) representation of the agent and provides the basis for agent's judgement on the expected shock based on the input stimuli. F is the pain expectation (fear), and P^I and P^E are interoceptive and exteroceptive representation of the corresponding observations O^I and O^E.*

One may ask what kind of abstraction can provide such a latent space representation. Interestingly enough, it is possible to highlight a number of shared regions between a minimal core affect network and the pain network (see Fig.7.6). However, for a more detailed discussion of FG neurobiology, see Appendix C.

Indeed, at the most abstract level, this shared manifold can be conceived as a joint latent space for pain and affect. This abstraction will be utilised for modelling throughout the rest of this chapter.

7.5 A simple Class 2 model of fear generalisation

Technically, we let the observer's rely on a simplified model of the agent where a latent space shapes the manifold where fear and pain conceptualisations interact. Once learnt,

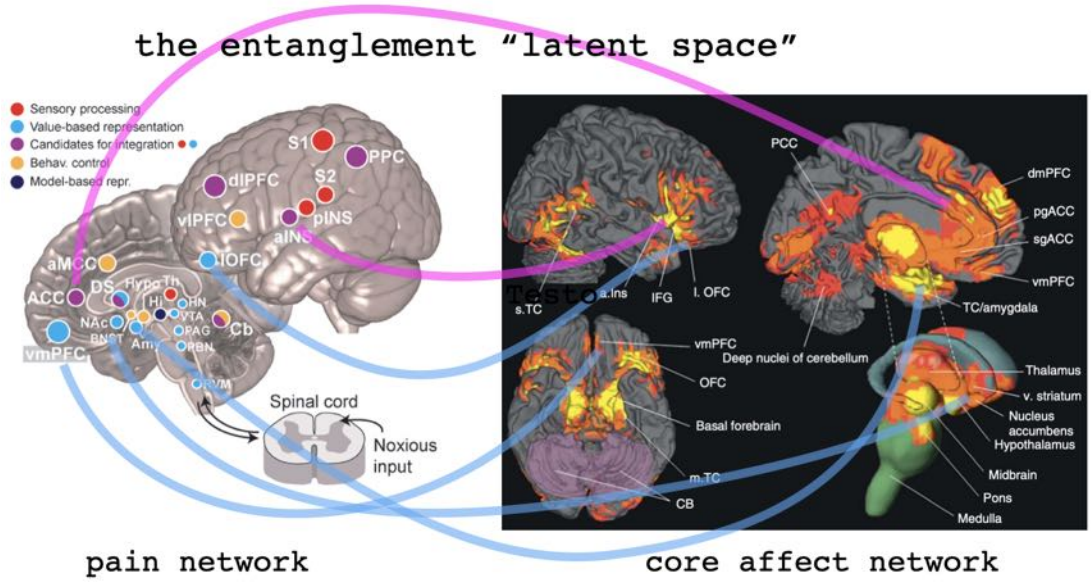


Figure 7.6: *A minimal representation of the entanglement of pain and the core affect basic network. Thick lines link shared brain regions. All in all this shared manifold can be abstractly seen as a sort of joint pain/affect latent space as summarised in Figure 7.5 (left).*

such manifold feeds on the perceptual (exteroception and interoception) representation of the agent and provides the basis for categorisation and action.

Minimally, agent's actions in this experimental setting are:

- internal actions aimed at controlling the body *internal milieu*, which can be gauged by the observer in terms of resulting physiological observations O^I;
- external actions, such as eye movements sampling visual/exteroceptive information O^E and agent's responses scoring shock expectancy in the trial.

In particular, let S be the latent random variable, $\mathcal{O} = \mathbf{o}^I$ the set/vector of stimuli features derived from physiological measurements.

Denote F the (binary) RV representing the low/high expected shock as reported by the agent based on the input stimuli (equivalently F can be understood in terms of pain expectation or, simply, fear). Then P^I and P^E denote interoceptive and exteroceptive representations of the corresponding observations O^I and O^E.

The generative approximation considers a latent conceptual space summarising information of interoceptive and exteroceptive signals. For this purpose, we employed a latent variable model (LVM).

For $n = 1, \cdots, N$ data points:

$$s_n \sim \mathcal{N}(s_n \mid 0, I_L) \tag{7.1}$$

$$\mathbf{o}_n \mid s_n \sim \mathcal{N}(o_n \mid W_o s_n, \sigma_o^2 I_{D_o}) \tag{7.2}$$

$$f_n \mid s_n \sim Bern(f_n \mid \sigma(w_f^T s_n)) \tag{7.3}$$

Here the realisation $S = s_n$ represents the shared latent subspace, tying both measurements $O = \mathbf{o}_n$ and reported shock expectancy rate $F = f_n$. Equation 7.3 approximates the rating action as a kind of logistic regression categorisation, based on the latent state space as affected by perceived stimuli. Put simply, the “input" data for this model, $\mathbf{o}_n$, are in this case interoceptive signals represented by the vector of wavelet coefficients extracted from each physiological modality.

The variance terms σ_o and σ_f control how much emphasis the model puts on the two different signals.

In the training phase, a latent space is built, and it abridges information about the signals. In the second phase, inference can be made on a subset of signals using the remaining as input to the model. This flexible model supports varied tasks in its inference phase, allowing for diverse simulations by altering input signals and deducing others. This versatility enables the application of a single model across multiple simulations, enhancing its utility in understanding complex datasets.

7.5.1 Simulation and results

Here we considered HR, EDA and pupil size variation (dilation) signals. Differents physiological signals have different latency. The idea is to consider for each signal its latency (5000 ms for EDA, 250 ms for HR, and 1000 ms for the pupil). Heart rate and EDA signals were recorded throughout the whole experiment, even between trials; this allowed us to extract the portion of the signal corresponding to the duration of the individual trial (6 seconds), also taking into account the latency of the signal. Unlike the previous signals, we do not have all the recording of signals related to the change in pupil diameter: considering signal latency, we kept a signal duration of 5 seconds (one less than the trial duration). For a group of subjects the heart rate and the EDA data signals of the last trial were not valid because the recording of the signals was interrupted before we could consider the latency of it: in these cases we decided to consider only 159 trials instead of 160.

To extract O^{HR}, O^{EDA} the HR and EDA signals features, respectively, we have used the approach proposed by Boccignone et al. (2018) exploiting 4-s windows with a 2 seconds overlap and then pass the data to a *discrete wavelet transform* (DWT). As suggested, we have chosen Daubechies 3. We empirically selected a suitable level of Daubechies 3, following the rule $Lmax = (\log 2N) - 1$, where N is the signal length; we retained only the approximate coefficients as feature vector both for heart rate and electrodermal activity.

A wavelet is a function $\psi \in L^2(\mathcal{R})$ that yields a basis in $L^2(R)$ by means of translations and dyadic dilations of itself, i.e.

$$g(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_{j,k} \psi \left(2^j x - k\right)$$

for all $g \in L^2(R)$ Such a decomposition is called the discrete wavelet transform. For O^{Pupil} feature extraction, related to the dilation of the diameter of the pupil, we used the same method by changing the size of the sliding window to 3 seconds with 1 second of overlap.

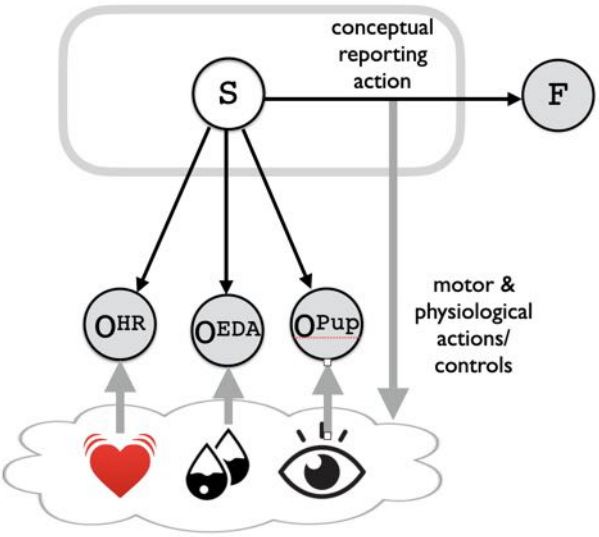


Figure 7.7: *A simplified PGM of a model of shock-expectancy prediction (represented by the binary RV F) based on physiological perceptual features extracted via wavelets from HR, EDA and pupil diameter measurements. The state-space spanned by S, here simply denoting a latent RV, might be more generally conceived as a complex latent space $\mathcal{S} = \{S^{(l)}\}_{l=1}^{L}$.*

Initially, we created a distinct model for each participant, referred to as the unpooled model approach. This method involved analysing the physiological signals from each individual's trials within the original study.

The main components of the model presented in Figure 7.7 are the latent space S and the perceptual features derived from wavelet-based analysis of HR, EDA, and pupil diameter variation physiological signals.

The input data considered have a shape $D_O \times N$, where D_O is different for all the signals (D_{eda}, D_{hr} and D_{pupil}) and depends from the feature extraction method (the length of the wavelet coefficient) and N represents the number of trials in the experiment.

One crucial aspect is determining the optimal dimensionality of the latent space, which consists of two dimensions: K and N. K can be altered as a parameter to compress the data while minimising information loss.

Each observed data type has an associated set of weights that are used to combine the latent variable with the observed data. The weights are modelled using the W_{eda}, W_{hr}, and W_{pupil} variables which are $K \times D_{eda}$, $K \times D_{hr}$, and $K \times D_{pupil}$ matrices of Normal random variables, respectively. Also for shock expectancy, there is a set of weights, combining data with the latent variable, represented as a $K \times D_F$ matrix of Normal random distribution.

The observed physiological data are modelled using Normal distributions with mean $W_{eda} \times P$, $W_{hr} \times P$, and $W_{pupil} \times P$, and variance σ_{eda}, σ_{hr}, and σ_{pupil}. For the binary shock expectancy label, the observed data are modelled using a Bernoulli distribution with probability $\sigma(W_F \times S)$.

The whole model was computationally instantiated by resorting to PyMC, a Python Probabilistic Programming framework (Patil et al., 2010). Figure 7.8 presents the actual

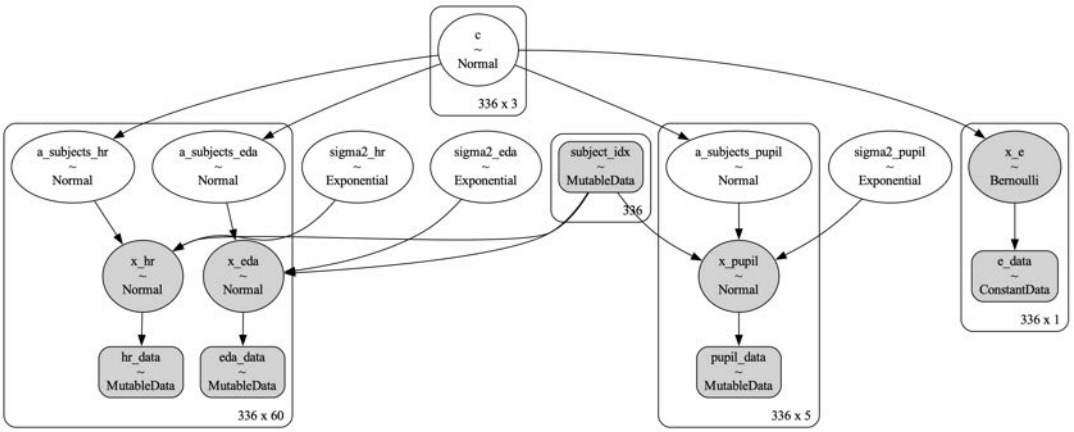


Figure 7.8: *The full PyMC implementation model of the PGM of shock-expectancy prediction outlined in Fig.7.7. The model has been automatically generated from PyMC code.*

Latent space categorisation with fear generalisation data		
	Latent Space dimension	Test accuracy mean
EDA	6	0.56
Heart rate	10	0.57
Pupil	15	0.57
EDA + HR + Pupil	10	0.68

Table 7.1: *Binary categorisation accuracy for shock expectancy rating prediction.*

Bayesian model that specifies the PGM sketched in Fig.7.7, and based on the sampling Eqs. 7.1,7.2,7.3. In addition, appropriate priors have been adopted for parameter sampling following the canonical Bayesian approach.

During training, the model's weight matrices were refined by analysing physiological data alongside shock expectancy labels. In the testing phase, both the latent space and physiological signals informed the prediction of shock expectancy, resulting in binary labels. For simplicity, user's ratings from a scale of 1 to 5 were reduced to these binary outcomes.

A summary of results is reported in table 7.1. The main and best result is provided in the bottom row. Others refer to single modality performance.

7.6 The active inference model: a simulative approach

If we consider the Class 2 model definition we considered in Chapter 4 and summarised in Figure 4.10, one distinctive feature of such class was in the possibility of accounting for agent's action and the feedback provided by the environment (comprising either the external world and the agent's body *internal milieu*) as a consequence of agent's action on the environment.

Indeed, in the observer's model of the agent's behaviour (namely, the participant's behaviour), the *internal milieu* feedback is implicitly accounted for through the phys-

iological signal observations (see Figures 7.5 and 7.7) while the reporting action is reduced to the inference of the categorical state formalised via F. As to this reporting action the model might be classified as a kind of $1\frac{1}{2}$ Class model.

A further step towards a plain Class 2 model, should consider the feedback effect that the agent's shock expectation guessing/reporting might have on the agent's inferential process (thus closing the action-perception loop).

In this section, we introduce a simulative approach using an Active Inference Partially Observale Markov Decision Process (AI-POMDP) to analyse the experimental data collected from subjects tasked with evaluating the likelihood of receiving a shock given a specific visual stimulus. The application of active inference in an POMDP framework is particularly suitable in this context as it allows the agent to make informed decisions based on probabilistic inferences from the morphed images they are presented with, simulating the human process of pain anticipation linked to visual cues.

Figure 7.9 outlines at a glance a minimal FEP-AI architecture suitable for POMPD realisation. For a detailed and formal examination of Active Inference, please see Appendix B.

The choice to employ an active inference approach is driven by its robust framework for modelling perception and decision-making processes under uncertainty, which mirrors the experimental setting where participants are required to gauge potential pain from visual stimuli. In this model, the type of image morphing serves as a critical factor in influencing the agent's prediction about the presence or absence of pain, aligning with how participants in the study might interpret and respond to varying degrees of stimulus ambiguity.

To facilitate a more streamlined modelling process, the experimental setting was scaled down while preserving its essential characteristics. A POMDP agent is defined by observations, hidden states, policy and related distribution. With regard to observations, it can rely on three outcomes:

- a binary morphing level `morph_level_obs = {0, 1}`
- an observation of shock reception (whether they receive it or not), in the variable `shock_obs = {shock, no_shock, null}`
- and the agent's level of surprise concerning the shock, that is `surprised_obs = {not_surprised, surprised, null}`

The latter serves somewhat as a reward because the goal is to eliminate surprise, indicating an effective prediction. The hidden state, on the other hand, represents whether the current trial is associated with a CS+ (conditioned stimulus positive) or CS- (conditioned stimulus negative), `s={cs+, cs-}`.

This model simulates the behaviour of a rational agent participating in the experiment. Based on the observations about the morphing level of visual stimuli and shock reception, the agent is capable of predicting shock expectancy and managing beliefs over states within the active inference framework. The algorithm 2 presents a simulation cycle with active inference, thus a single trial.

To validate the active inference model, we crafted a simplified environment resembling the original experiment. We limited the simulation to 100 trials, with the first 50 involving CS- images not followed by an electric shock and the last 50 involving CS+

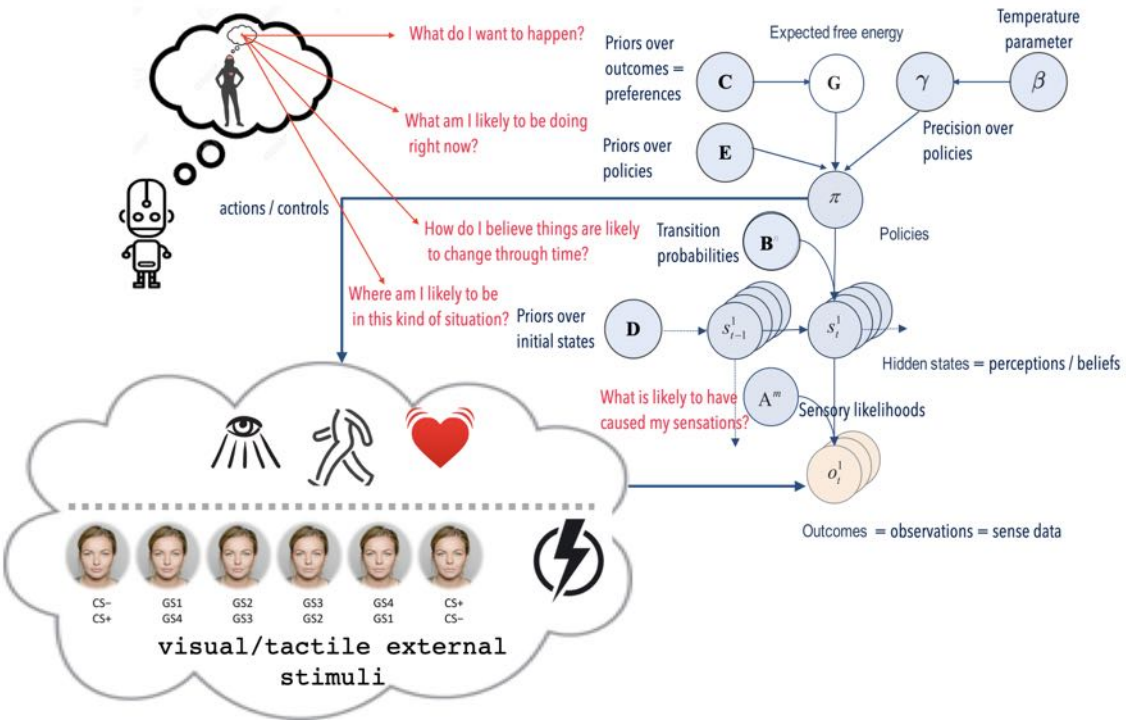


Figure 7.9: *An AI-POMPD agent, in a nutshell. The PGM models the data generation process via active inference, which in this case takes the form of selecting policies (π) as sequences of actions which result in influencing various world states (external and internal to the agent's body). Selected policies take the form of specifying sequences of state transitions, chosen according to whichever actions are expected to minimise the overall prediction-error PE (free energy) and the expected free energy G. These states are hidden, or latent, and must be probabilistically inferred based on sensory observations. Perceptual inference flows in an opposite direction to that indicated by the arrows (inverse probability/inferential process). The action selection resulting from the generative process constitutes a kind of inference, in that it reflects expectations as to what is likely to minimise the divergence between predictions and observations. The precision with which the agent probabilistically select among actions is reflected by "inverse temperature" parameters (γ and β) to specify thresholds for policy deployment, so influencing* exploration/exploitation *trade-offs. The parameter/prior matrices* $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}$ *can be seen as encoding answers to the fundamental questions (highlighted in red) that the agent "has in mind".*

Algorithm 2 Algorithm of active inference loop in fear generalisation task

```
while t ≤ T do
    morph_level_obs = [0, 1]
    while s ≤ S do
        states ← infer_states(morph_level_obs)
        π ← infer_policy(states)
        action ← sample_action(π)
        shock_obs ← step(action)
        states ← infer_states(shock_obs)
        reset_agent()
    end while
end while
```

images followed by an electric shock. This setup allows us to explore how the model's beliefs about states and actions adjust based on observations from a specific trial and the beliefs formed in previous trials.

Particularly, the first two panels in Figure 7.10 display the data observed by the agent, while the third panel shows the model's shock expectancy rating. After conducting a trial with random choices, the agent learnt the correlation between observations and shocks, using this knowledge to anticipate the likelihood of a shock in future trials.

The model was then tested in a more experiment-like environment, using the same data. As illustrated in Figure 7.11, the agent maintained predictable behaviour, aligning with expectations. This demonstration underscores the model's capability to adapt and accurately predict outcomes based on the integrated active inference approach, mirroring the decision-making processes observed in the experimental participants.

7.7 Pain and fear in the eyes: theoretical and empirical consequences of the exploitation/exploration dilemma

Embracing the active inference approach has the advantage of making clear from first principles the nature and the course of the actions the agent engages in under the context of their beliefs and the material circumstances of the environment. Agent's behaviour in a fear generalisation experiment makes no exception.

In the active inference framework (cfr. Appendix B), from a general standpoint, minimising free energy ensures expectations encode posterior beliefs, given observed outcomes. However, beliefs about policies rest on future outcomes. This means that policies should, a priori, minimise the free energy of beliefs about the future. This can be formalised by making the log probability of a policy π proportional to the free energy expected in the future $G(\pi) = \sum_{\tau>t} G(\pi, \tau)$

This is summarised, as detailed in Appendix B, by specifying $G(\pi, \tau)$ via Eq. B.24, which we rewrite here for the reader's convenience:

$$G(\pi, \tau) = \underbrace{D_{KL}\left[Q(o_\tau|\pi)||P(o_\tau)\right]}_{Expected\ cost} + \underbrace{\mathbb{E}_{Q(s_\tau|\pi)}\left[H(P(o_\tau|s_\tau))\right]}_{Entropy} \tag{7.4}$$

where $H[P(o_\tau|s_\tau)] = \mathbb{E}_{P(o_\tau|s_\tau)}\left[-\ln P(o_\tau|s_\tau)\right]$ is the entropy.

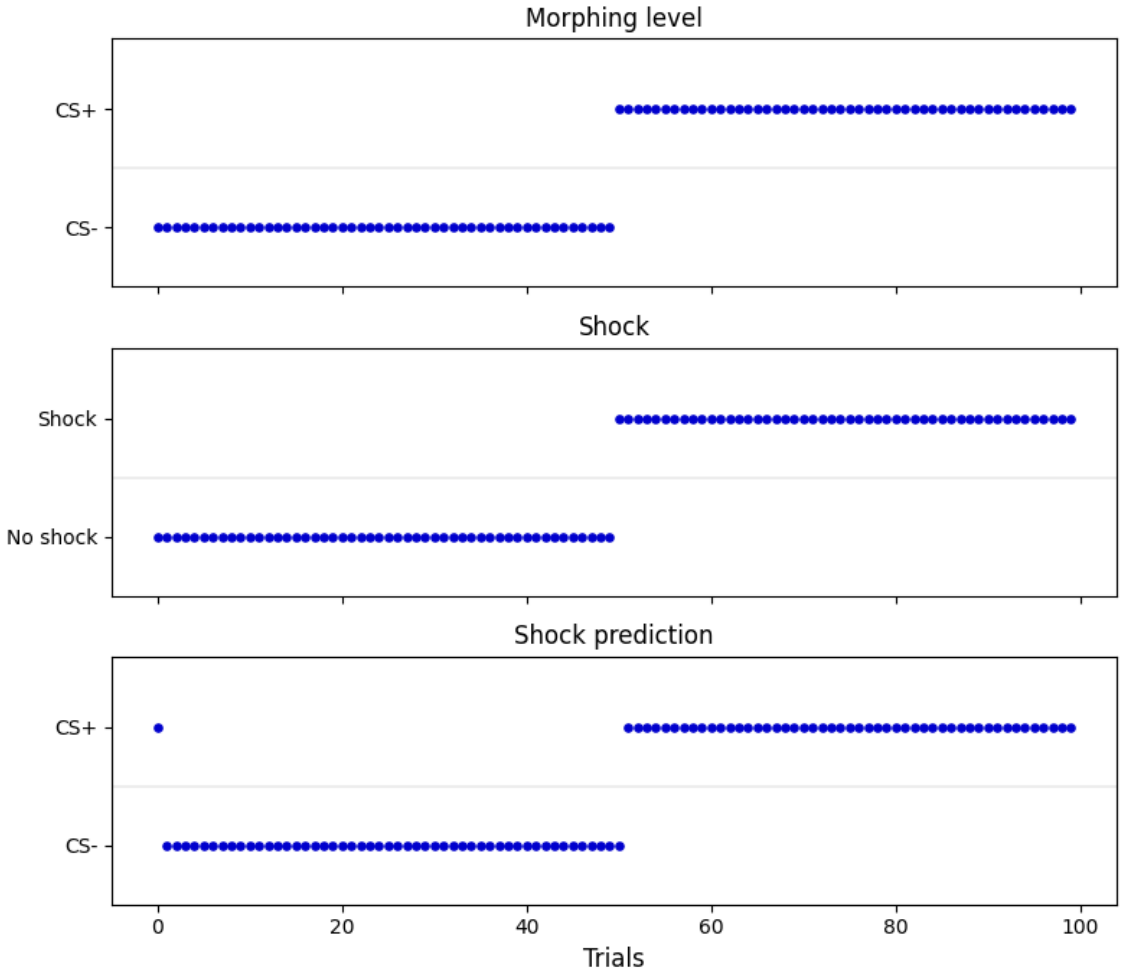


Figure 7.10: *Simulation of an agent on emulated trials.*

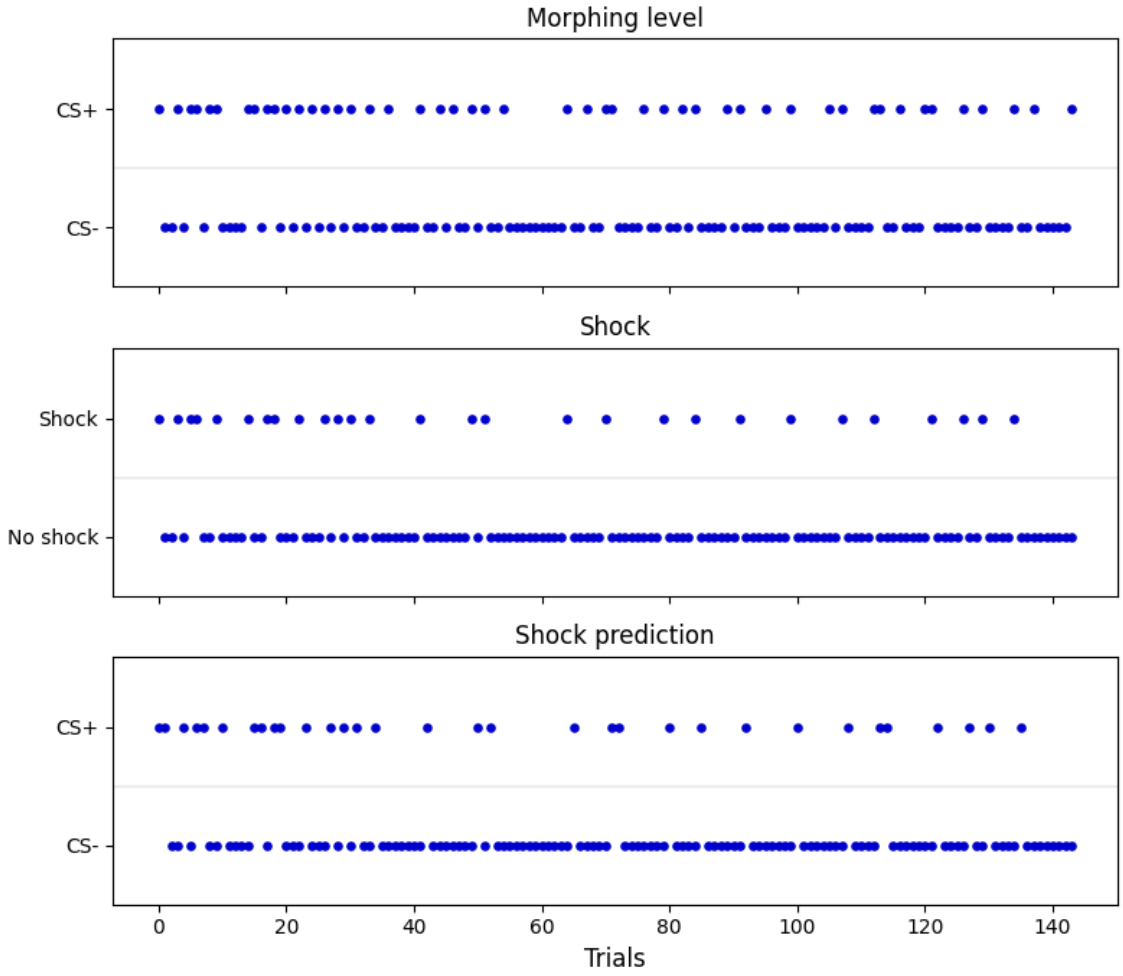


Figure 7.11: *Simulation of an agent provided with experimental data.*

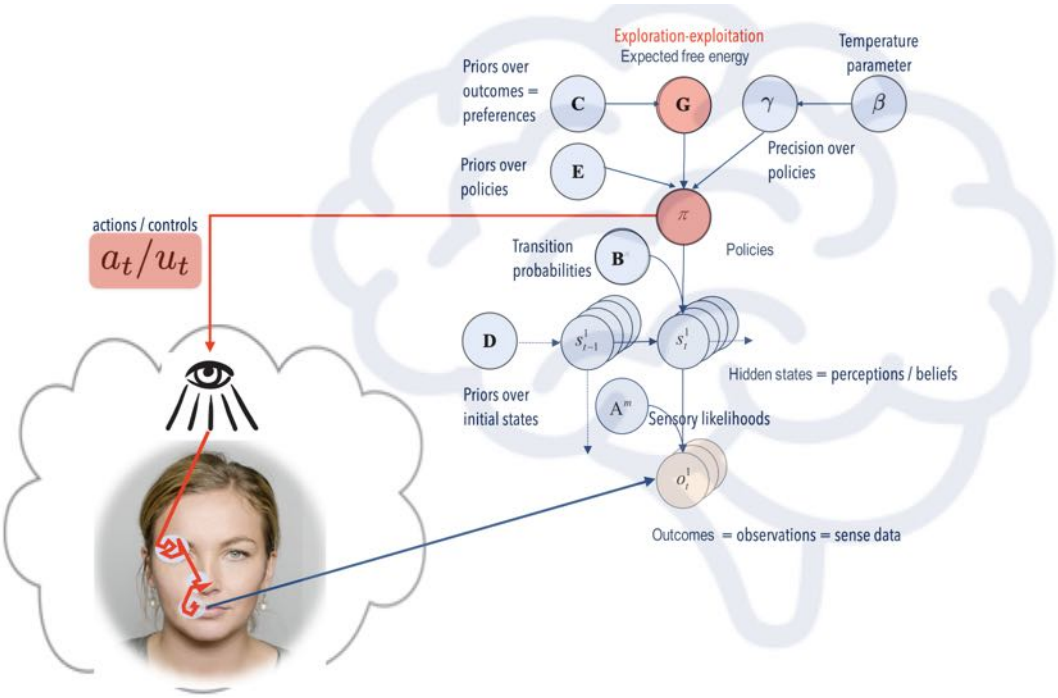


Figure 7.12: *Visual foraging actions unfolded through subject's gaze behaviour (local inspection of the stimulus by fixations and coarse relocations via saccades) are the consequence of the actual process representing state transitions in the real world, The latter generates observations or outcomes that are used to update the internal states of the agent. Keeping with the general assumption that such agent acts as an AI-POMDP agent, then their gaze behaviour as determined by actions and motor control parameters instantiates the* exploration/exploitation *entailed by the optimisation of the expected free-energy G. In turn, the G function clearly depends on parameter/priors $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}$ together with the "inverse temperature" parameters (γ and β) specifying thresholds for policy π deployment. Overall, such prior and parameters setting (formally, the agent's model) are the "signature" of agents individuality and traits.*

Equivalently, $G(\pi, \tau)$ can be formulated as

$$G(\pi, \tau) = \underbrace{\mathbb{E}_{Q(o_\tau, s_\tau|\pi)}\left[\ln Q(o_\tau|\pi) - \ln Q(o_\tau \mid s_\tau, \pi)\right]}_{Epistemic\ value} + \underbrace{\mathbb{E}_{Q(o_\tau, s_\tau|\pi)}\left[\ln P(o_\tau)\right]}_{Extrinsic\ value} \quad (7.5)$$

Equation 7.4 shows that plan (i.e. action sequence) selection aims at minimising the expected cost and ambiguity. The latter relates to the uncertainty about future observations given hidden states. In a sense, plans tend to bring the agent to future states that generate unambiguous information over states. On the other hand, the cost is the difference between predicted and prior beliefs about final states. Plans are more likely if they minimise cost, and lead to observations that match prior desires. Minimising G leads to both exploitative (cost minimising) and explorative (ambiguity minimising) behaviour. This results in a balance between goal-oriented and novelty-seeking behaviours.

Equivalently, the formulation in Eq. 7.5, shows that, in a visual context, for instance (Mirza et al., 2016a, 2018; Heins et al., 2020), the agent should act to maximise epistemic value or resolve uncertainty about the unknown context (the scene and its spatial transformations / morphings), until the uncertainty about the scene is reduced to a minimum. At this point, it should maximise extrinsic value by sampling the choice location it believes will provide feedback that endorses its beliefs. This again speaks to the trade-off between exploration and exploitation.

Essentially, in the case of eye movement actions, actions associated with the exploration of the scene have no extrinsic value—they are purely epistemic. In contrast, actions associated with the choice locations (locations that are used to identify the scene's category) have extrinsic value, because the agent has prior preferences about the consequences of these actions, namely reduce pain.

As to the specific case of attentive visual behaviour, individual differences, crucially the differences between individuals' internal model parameters, priors and preferences, will reflect in the individual's foraging behaviour, which is eventually apparent in the "feed-and-fly" pattern characterising gaze dynamics.

More precisely, as summarised in Figures 7.9 and 7.12, the agent's model

$$\mathcal{M} = \langle \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}, \beta, \gamma \rangle$$

can be represented via matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}$ together with the "inverse temperature" parameters (γ and β) specifying thresholds for policy π deployment

Thus, here we consider the following:

Research hypothesis Under the general assumption of individual exploitation/exploration processes (active inference internal model $\mathcal{M}$), the deployment of visual attention through gaze offers an effectual window on the individual's information-seeking, foraging behaviour (see Bella-Fernández et al. (2022) for an in-depth discussion). Thus, under the visual foraging hypothesis, we assume that a foraging-based analysis of gaze deployment over ecological stimuli such as those we are dealing here is an appropriate approach to characterise the individual's visual information-seeking behaviour within a fear conditioning setting.

More precisely, in such a setting we expect that, by inferring gaze behaviour parameters $\widetilde{\mathcal{M}}$ from participants' eye-tracking data, a compact set of descriptors is provided

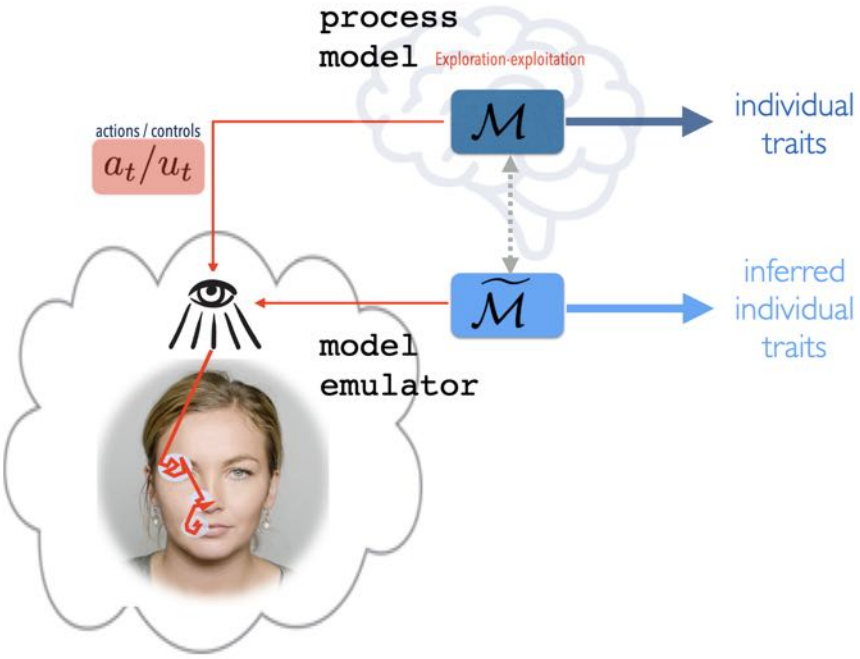


Figure 7.13: *Visual foraging in the FG experimental setting unfolding under agent's process model $\mathcal{M}$, can be phenomenologically accounted for by the statistical emulator $\widetilde{\mathcal{M}}$. Agent's individual traits such as social anxiety traits playing a role in FG are grounded in and causally linked to the process model $\mathcal{M}$. Here we work under the hypothesis that such traits can be phenomenologically inferred from the emulator model $\widetilde{\mathcal{M}}$.*

that can be exploited to weigh latent factors behind the individual's FG process, in particular their social anxiety traits.

In other terms we can think of $\widetilde{\mathcal{M}}$ as an *emulator*. An emulator is a statistical approximation of the simulator $\mathcal{M}$, which allows for simpler computations than using the simulator itself (Insua et al., 2012).

In a nutshell, we explore the use of $\widetilde{\mathcal{M}}$ as an emulator of $\mathcal{M}$ under the assumption that the relationship

$$\mathcal{M} \leftrightarrow \widetilde{\mathcal{M}}$$

holds.

Figure 7.13 depicts at a glance the rationale of the hypothesis.

Below we further expand on and motivate such hypothesis.

7.7.1 Foraging signatures in eye movements: a premise

The uniqueness of behaviour hiding behind the individual's gaze dynamics can be effectively reconsidered in the light of theories and models adopted in animal ecology; for an in-depth discussion cfr. D'Amelio et al. (2023).

Individual-based approaches in ecology seek a mechanistic understanding of how variation among individual organisms generates or contributes to patterns at population, community and ecosystem levels (but for an in-depth discussion, see Trappes (2022)).

What makes an organism unique? Why does this animal have this unique set of traits? Clearly, answering such questions involves describing highly specific, idiosyncratic properties and causal histories.

Individuality in behavioural ecology is defined as the phenotypic and ecological uniqueness of single individuals; namely, what makes an individual phenotypically unique is generally not a single phenotypic property, but rather the possession of a whole set of phenotypic traits together with unique sets of ecological relations (Trappes, 2022). Features like behaviour, habitat use, feeding preference, social relations, and so on, are core to what makes individuals unique.

In this broad perspective, one remarkable example is foraging behaviour (Stephens, 1986; Viswanathan et al., 2011). Foraging is a term that includes where animals search for food and which sorts of food they eat. According to optimal foraging theories (OFT), the forager continuously faces four problems (Bartumeus and Catalan, 2009):

1. *patch-choice*: the forager has to make a decision concerning what type of patch (a clump of food) to search for (e.g. good-bushes full of berries);

2. *patch-finding*: the forager wanders through a landscape where certain resources are available and makes a choice on how to move between patches (optimal movements);

3. *target/prey-finding*: once a patch is located, the forager should decide what prey to take and handle (optimal diet choice);

4. *patch-leaving*: the available resources in the patch decrease over time spent handling preys; eventually, a forager makes the decision of leaving the current patch to search in a more profitable patch, or just to finish the task (giving up or departure times from patches).

All in all, the forager, in order to solve the above problems, engages in a recurrent cycle of successive local exploitation of the patch alternated with exploration, i.e. relocation between patches. While engaged in the foraging cycle, the forager also learns about the profitability of the individual patches and the environment after visiting each patch. Such exploitation/exploration pattern is the genuine fingerprint of animal's foraging behaviour.

In recent research, consistent individual differences in behaviour (animal personality) and food resource use (individual specialisation) have emerged. Animal personality and individual specialisation are patterns of individual-level phenotypic variation, both describing similar concepts of temporally consistent individuality in behavioural and food web ecology (Toscano et al., 2016). Personality is a motif which accounts for consistent behavioural differences or traits in animals (tipically: exploration, activity level, sociability, boldness and aggressiveness); indeed, research in behavioural ecology provides evidence that personality traits can be detected in a vast collection of invertebrate and vertebrate taxa (Toscano et al., 2016). Individual specialisation involves significant individual differences in the animal's diet arising within a population. Most important here, differences in both personality and specialisation remain consistent over time and across ecological contexts; namely a steady variation is observed among individuals, as opposed to a strikingly low variability within the individual (Toscano et al., 2016).

Given that foraging behaviour provides an important marker of animal individuality, how does then foraging link to cognition and, in turn, to eye movement behaviour?

In a nutshell, it has been argued that ancestors' foraging for material resources in a physical landscape evolved over time in foraging for information in cognitive space (Todd and Hills, 2020; Budaev et al., 2019; Rosati, 2017; Pirolli, 2007; Hills, 2006); when the individual specifically engages in a cognitive visual task, then gaze foraging over a visual scene is by far the fundamental action to sample and gauge the visual input (Liversedge and Findlay, 2000; Wolfe, 2013; Ehinger and Wolfe, 2016; Mirza et al., 2016b; Cain et al., 2012)

At the psychological level, goal-directed foraging has been related to relevant cognitive skills (Todd and Hills, 2020): spatial memory (the ability to recall the location of resources and navigate efficiently between them), value-based decision-making (to choose the best course of action given all the available alternatives, what is), and executive control (the versatile guidance of behaviour, overriding reflexive, automatic responses). Even self-awareness, deliberation, and free will can be subsumed under this view.

At the brain level, the neuro-molecular architectures and mechanisms that support the foraging mind are nothing but the evolutionary response to the exploration/exploitation dilemma that must be faced in the external, physical environment (Todd and Hills, 2020). For instance, it has been surmised by Hill (Hills, 2006) that in both animal and cognitive foraging, dopamine acts as a modulator between two behavioural extremes: when dopaminergic activity is high, behaviour is focused, and eventually, stereotypic; by contrast, when dopaminergic activity is low, behaviour becomes unfocused and fails to endure (Hills, 2006). Noticeably, Hill's dopamine hypothesis fits comfortably in recent perspectives on visual foraging based on the active inference account (Mirza et al., 2016a; Friston et al., 2012), an influential framework in current theoretical neurobiology. Here, in a Bayes-optimal behavioural view, dopamine controls the precision or salience of (external or internal) cues that engender actions, either exploitative (i.e., risk minimising) and explorative (i.e., ambiguity minimising) (Mirza et al., 2016a).

To sum up, foraging as exploration/exploitation trade-off is an appealing and principled framework for dealing with gaze at different explanation levels (Wolfe, 2013; Ehinger and Wolfe, 2016; Mirza et al., 2016b; Cain et al., 2012)

This statement can be intuitively appreciated at a glance by considering the single trace in the left panel of Fig. 7.14: gaze dynamics unfolds by alternating explorative, long relocations between semantically relevant image patches (e.g., the face), and local movements to "exploit" the selected patch. Importantly, these trajectories represent the microscopic, fine resolution scale of gaze dynamics. At a coarser scale, this explore/exploit pattern can be parsed into a "saccade and fixate" discrete sequence (Land, 2006) or scan path (right panel of Fig. 7.14), which is most often chosen as the input for further analyses in either the visual attention/eye movement and the behavioural biometrics fields. Saccades are thus functionally defined as the fast movements that redirect gaze to a new part of the surroundings; each fixation summarises fixational movements that occur within intervals between saccades, in which gaze is held almost stationary, so to keep onto the fovea (the central part of the retina) the circumscribed region of interest (RoI) within the viewed scene.

To make the connection clear beyond the intuition, Table 7.2 lays down the relation-

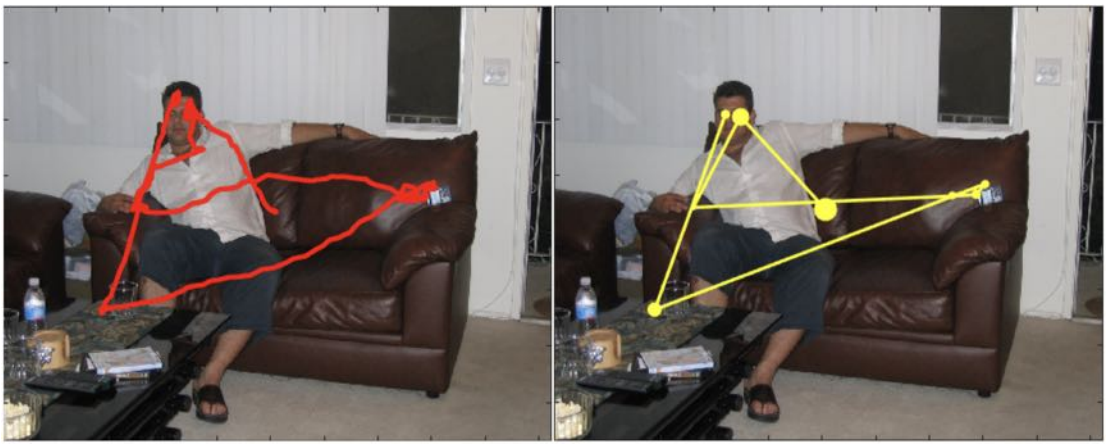


Figure 7.14: *The dynamics of one observer's gaze while visually foraging on the image landscape as recorded through an eye-tracking device.* Left*: the "raw data" represented as a time-sequence of spatial coordinates (red dots) displayed as the red trace overlapped on the viewed image.* Right*: the observer's scan path, namely, the continuous raw data trace parsed into a discrete sequence of fixations (yellow disks) and saccades (segments between subsequent fixations); disk radius is proportional to fixation time. Image and eye-tracking data are publicly available from the Cerf's dataset (Cerf et al., 2007).*

ship between deployment of attention through gaze over a scene and animal foraging behaviour.

Table 7.2: *Relationship between Visual Attention and Foraging*

Visual attentive processing of a scene	Patchy landscape foraging
Perceiver	Forager
Perceiver's gaze shift	Forager's relocation
RoI/object	Patch
RoI/object selection	Patch choice
Deploying attention to RoI/object items	Patch/prey handling
Disengaging from RoI/object	Patch leave or giving-up

It is important to note that the given task and the stimulus exteroceptive properties are not the only constraints to the visual exploration/exploitation pattern exhibited by the perceiver. Interoception, the observer's own perception of their internal state of the body, matters too, and, consequently, the perceiver's affective state and feelings: it is no surprise that in everyday life, gaze is of service to cogently capitalise on visual information that includes social information; markedly, others' emotions and intentions (Shepherd and Platt, 2007; Guy et al., 2019; Boccignone et al., 2019).

All in all, the exploration/exploitation pattern springing from the gaze sampling endeavour provides a signature of the individual's plans, goals, interests, likely sources of rewards and expectations about future events, (Kowler, 2011; Henderson, 2017), social traits, and personality (Guy et al., 2019; Cuculo et al., 2018).

Further, in addition to the exteroceptive/interoceptive properties of the stimulus coupled with the pursued goal/task, one should consider biomechanical factors and motor

uncertainty (affecting, for instance, oculomotor flight time and landing accuracy) as relevant sources of systematic tendencies or "biases" (Tatler and Vincent, 2009). The latter, albeit often neglected in the literature, can be conceived as those regularities that are common across all instances of, and manipulations to, behavioural tasks (Tatler and Vincent, 2008, 2009). One remarkable example is the amplitude distribution of saccades and microsaccades (ocular movements occuring within a fixation) that typically displays a positively skewed, long-tailed shape (Tatler et al., 2011; Dorr et al., 2010; Tatler and Vincent, 2008, 2009).

As a result of such a multifactorial drive, there is a tiny probability that two subjects will fixate exactly the same location of the visual scene at precisely the same time. Such variability is easily detected in a variety of experimental conditions, especially under a free-viewing or a general-purpose task: e.g., when looking at natural images, movies (Dorr et al., 2010), or even dynamic virtual reality scenes (Hu et al., 2020). Interestingly enough, this effect is more pronounced when free-viewing static images (but see Tatler et al. (2011), for a discussion), as it has been shown in Figure 7.14: consistency in fixation locations selected by observers decreases over the course of the first few fixations after stimulus onset. Crucially, variations in individual scan paths (as regards chosen fixations, spatial scanning order, and fixation duration) still keep on when the scene embeds semantically rich "objects" and get to be idiosyncratic (Tatler et al., 2011).

On the other hand, recent studies considered the variability of eye movements between observers marking off the characteristics that are stable and reliable, which should be conveniently handled as a trait of the observer rather than be downgraded to "noise" (Henderson and Luke, 2014; Bargary et al., 2017). For example, a novel gaze-related trait has been reported by Guy et al. (2019), who have shown that the amount of time subjects fixate on others' faces differs between individuals in a consistent manner.

In these terms, gaze behaviour is apt to unveil individual's characteristics tantamount to gait and speech. However, different from gait and speech, gaze behaviour ineluctably controls the visual input processed by the brain, hence making such variability an important factor in determining and reflecting the individual's inner world.

Given that here the general problem is that of modelling the probability that in the fear conditioning experiment, the individual's gaze is in state $S_t = s_t$ at time t, representing the outcome of the gaze-deployment stochastic process, where s_t denotes a spatial position (i, j) or region (e.g., nose, eyes, etc.), within the viewed image (face), such problem can be analysed at two levels or scales: the macroscopic and the microscopic levels.

Consider each foraging trajectory (scan path) as the trajectory of a particle or a random walker. In this view, the probability $P(\mathbf{s}, t)$ can be interpreted as the density $\rho(\mathbf{s}, t)$ (number of particles per unit length, area or volume) at point $\mathbf{s}$ at time t. In fact, the density $\rho(\mathbf{s}, t)$ can be recovered by multiplying the probability density $P(\mathbf{s}, t)$ by the number of particles.

On the other hand, the finest grain of representation of a many-particle system is the individual particle, where each stochastic trajectory becomes the basic unit of the probabilistic description of the system. This is the **microscopic level**. In its modern form, it was first proposed by the French physicist Paul Langevin, giving rise to the notion of random walks where the single walker dynamics is governed by both regular

and stochastic forces (which resulted in a new mathematical field of stochastic differential equation, briefly SDE. At this level, $P(\mathbf{s}, t)$ can be be obtained by considering the collective statistical behaviour as given by the individual simulation of many individual particles (technically a Monte Carlo simulation)

In the opposite way, we could straightforwardly consider, in the large scale limit, the equations governing the evolution of the space-time probability density $P(\mathbf{s}, t)$ of the particles. This coarse-grained representation is the **macroscopic level** description.

A useful analogy for visualising both levels is provided by structure formation on roads such as jam formation in freeway traffic. At the microscopic scale, one can study the motion of an individual vehicle, taking into account many peculiarities, such as motivated driver behaviour and physical constraints. On the macroscopic scale, one can directly address phase formation phenomena collectively displayed by the car ensemble.

For a preliminary evaluation of participants' visual foraging behaviour we first conducted a coarse-grained assessment.

7.7.2 Preliminary analysis at the macroscopic level

We know that each visual stimulus is characterised by a morphing region, which is the significant area where the subject must look in order to disambiguate aggressive images (thus followed by shock) from nonaggressive ones. A preliminary analysis of Reutter's study (Reutter, 2022) was conducted on the regions of interest, calculating for each fixation where it fell and whether that region was diagnostic or not, to see if the subject was looking in the right area. From an initial analysis, it was found that longer observation times in diagnostic regions (eyes and nose/mouth) were associated with less fear generalisation by the subject.

A gaze data analysis was conducted using Variational hierarchical Hidden Markov Models (VHMMs) to investigate whether it is feasible to group fixations together and derive attentional properties of a group of subjects.

The VHMM model and VHEM learning For this analysis we have exploited the variational HEM (VHEM) algorithm (Coviello et al., 2014). The hidden Markov model (HMM) is a widely-used generative model that copes with sequential data, assuming that each observation is conditioned on the state of a hidden Markov chain. The VHEM is a novel algorithm to cluster HMMs based on the hierarchical EM (HEM) algorithm. The VHEM algorithm i) clusters a given collection of HMMs into groups of HMMs that are similar, in terms of the distributions they represent, and ii) characterises each group by a "cluster center", i.e., a novel HMM that is representative for the group, in a manner that is consistent with the underlying generative model of the HMM. To cope with intractable inference in the E-step, the HEM algorithm is formulated as a variational optimization problem, and efficiently solved for the HMM case by leveraging an appropriate variational approximation. The benefits of the VHEM have been demonstrated on several tasks involving time-series data, such as hierarchical clustering of motion capture sequences, and automatic annotation and retrieval of music and online hand-writing data, showing improvements over current methods.

Assume a sequence of τ observations $O_{1:\tau} = \{o_1, \ldots, o_\tau\}$ to be generated via a double embedded stochastic process, where each observation (or emission) o_t at time t

depends on the state of a discrete hidden variable S_t, and the sequence of hidden states $S_{1:\tau} = \{s_1, \ldots, s_\tau\}$ evolves as a first-order Markov process. The hidden variables can take one of S values, $\{1, \ldots, S\}$, and the evolution of the hidden process is encoded in a state transition matrix $B = [b_{\beta,\beta'}]_{\beta,\beta'=1,\ldots,S}$, where each entry, $b_{\beta,\beta'} = p(s_{t+1} = \beta' | s_t = \beta, \widetilde{\mathcal{M}})$, is the probability of transitioning from state β to state β', and an initial state distribution $\pi = [\pi_1, \ldots, \pi_S]$, where $\pi_\beta = p(s_1 = \beta | \widetilde{\mathcal{M}})$.

Each state β generates observations according to an emission probability density function, $p(o_t | s_t = \beta, \widetilde{\mathcal{M}})$. Here, we assume the emission density is *time-invariant*, and modeled as a Gaussian mixture model (GMM) with M components:

$$p(o | s = \beta, \widetilde{\mathcal{M}}) = \sum_{m=1}^{M} c_{\beta,m} p(o | \zeta = m, s = \beta, \widetilde{\mathcal{M}}), \tag{7.6}$$

where $\zeta \sim \text{Multinomial}(c_{\beta,1}, \ldots, c_{\beta,M})$ is the hidden assignment variable that selects the mixture component, with $c_{\beta,m}$ as the mixture weight of the mth component, and each component is a multivariate Gaussian distribution,

$$p(o | \zeta = m, s = \beta,) = \mathcal{N} o; \mu_{\beta,m} \Sigma_{\beta,m}, \tag{7.7}$$

with mean $\mu_{\beta,m}$ and covariance matrix $\Sigma_{\beta,m}$. The emulator HMM model is thus specified by the parameters

$$\widetilde{\mathcal{M}} = \{\pi, B, \{\{c_{\beta,m}, \mu_{\beta,m}, \Sigma_{\beta,m}\}_{m=1}^{M}\}_{\beta=1}^{S}\}, \tag{7.8}$$

which can be efficiently learned from an observation sequence $O_{1:\tau}$ (e.g. via Baum-Welch algorithm Murphy (2012), which is based on maximum likelihood estimation).

A hidden Markov mixture model (H3M) models a set of observation sequences as samples from a group of K hidden Markov models, each associated to a specific sub-behavior. For a given sequence, an assignment variable $z \sim \text{Multinomial}(\omega_1, \cdots, \omega_K)$ selects the parameters of one of the K HMMs, where the kth HMM is selected with probability ω_k. Each mixture component is parametrised by

$$\widetilde{\mathcal{M}}_z = \{\pi^z, B^z, \{\{c^z_{\beta,m}, \mu^z_{\beta,m}, \Sigma^z_{\beta,m}\}_{m=1}^{M}\}_{\beta=1}^{S}\}, \tag{7.9}$$

and the H3M is parameterised by $\widetilde{\mathcal{M}} = \{\omega_z, \mathcal{M}_\ddagger\}_{\ddagger=\infty}^{\mathcal{K}}$, which can be estimated from a collection of observation sequences using the EM algorithm.

Briefly, denote $\widetilde{\mathcal{M}}_b$ a base hidden Markov mixture model with K_b components. The goal of the VHEM algorithm is to find a reduced hidden Markov mixture model $\widetilde{\mathcal{M}}_r$ with $K_r < K_b$ (i.e., fewer) components that represents $\mathcal{M}\widetilde{\mathcal{M}}_b$ well. The likelihood of the sequence $o_{1:\tau} \sim \widetilde{\mathcal{M}}_b$ is given by

$$p(O_{1:\tau} | \widetilde{\mathcal{M}}_b) = \sum_{i=1}^{K_b} \omega_{i_b} p(o_{1:\tau} | z_b = i, \widetilde{\mathcal{M}}_b), \tag{7.10}$$

where $z_b \sim \text{Multinomial}(\omega_{1b}, \cdots \omega_{K_b b})$ is the hidden variable that indexes the mixture components. $p(o_{1:\tau} | z = i, \mathcal{M}_b)$ is the likelihood of $o_{1:\tau}$ under the ith mixture component, as in 7.9, and ω_{i_b} is the mixture weight for the ith component. Likewise, the

likelihood of the sequence $o_{1:\tau} \sim \widetilde{\mathcal{M}}_r$ is

$$p(o_{1:\tau}|\widetilde{\mathcal{M}}_r) = \sum_{j=1}^{K_r} \omega_{j_r} p(y_{1:\tau}|z_r = j, \widetilde{\mathcal{M}}_r), \tag{7.11}$$

where $z_r \sim \text{Multinomial}(\omega_{1_r}, \cdots, \omega_{K_r r})$ is the hidden variable for indexing components in $\widetilde{\mathcal{M}}_r$.

At a high level, the VHEM-H3M algorithm estimates the reduced H3M model $\widetilde{\mathcal{M}}_r$ and provides

1. a soft clustering of the original K_b components into K_r groups, where cluster membership is encoded in assignment variables that represent the *responsibility* of each reduced mixture component for each base mixture component, i.e., $\hat{z}_{i,j} = p(z_r = j|z_b = i)$, for $i = 1, \ldots, K_b$ and $j = 1, \ldots, K_r$;
2. novel cluster centers represented by the individual mixture components of the reduced model in eqn:mixmodel$_r educe, i.e.$,p(o$_{1:\tau}|z_r = j, \mathcal{M}_\nabla)$ for $j = 1, \ldots, K_r$.

Details of the variational procedure can be found in Coviello et al. (2014). The VHMM can provide a suitable tool to investigate gaze deployment over spatial regions, and to automatically detect regions of interest (ROIs) (Coutrot and Guyader, 2017; Chuk et al., 2014). In this case each HMM is used to directly model the eye movement patterns of each participant. The hidden states are assumed to be the ROIs and the eye movement data is assumed to be emissions of the ROIs. Hierarchical clustering can then be used to infer, at a higher level, an HMM describing the gaze behaviour over regions of a specific group of participants, each participant being represented by a single HMM at a lower level (Coutrot and Guyader, 2017; Chuk et al., 2014).

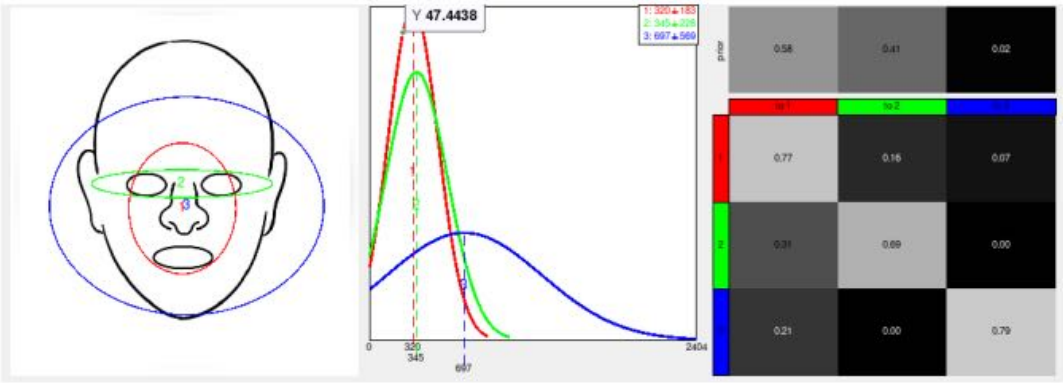
Analysis Specifically, HMMs were estimated based on the eye-gaze data of individual participants, and these individual HMMs were then clustered to identify shared attentional strategies. One significant result is shown in Figure 7.15.

Results were obtained as follows. First, all subjects that participated in the original experiment were divided into three groups, based on their Linear Deviation Score value: low generalisation, medium generalisation and high generalisation. Then, the algorithm has been run on these groups to see if there were any differences in how the various groups analysed the visual stimuli. From Figure 7.15 (a) it was observed that subjects belonging to the discriminative group (i.e., those who exhibit less fear generalisation) initially focus their attention on the diagnostic regions of the face (eyes and nose/mouth) as can be seen from the red and the green clusters. On the other hand, subjects who exhibit higher levels of generalisation tend to concentrate less on the specific diagnostic regions and instead direct their attention towards the entire face (green cluster in Figure 7.15 (b)).

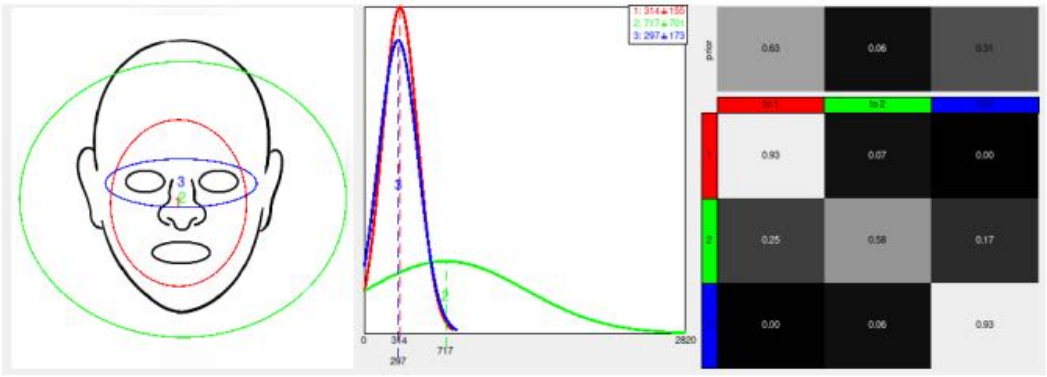
Given the results obtained via preliminary macroscopic analysis, a fine-grained analysis was undertaken.

Analysis at the microscopic level

To such end, use the the time-varying location state vector $\mathbf{s}(t)$ (screen coordinates) to denote the gaze position at time t. In the following, to simplify the analysis, we assume



(a) *HMM analysis on fixations of subjects with low fear generalisation.*



(b) *HMM analysis on fixations of subjects with high fear generalisation.*

Figure 7.15: *Macroscopic analysis via Hierarchical HMM analysis.*

that the state variable is observed; more precisely that observation $\mathbf{o}(t) \sim p(\mathbf{o}(t) \mid \mathbf{s}(t))$ is sampled via $p(\mathbf{o}(t) \mid \mathbf{s}(t)) = \delta(\mathbf{o}(t), \mathbf{s}(t))$ that is a Dirac-delta distribution. Thus, we shall directly refer to $\mathbf{s}(t)$

Each observed trajectory $\{\mathbf{s}(t),\ t = 0 \dots T\}$ is a realisation of a stochastic process $\{S(t),\ t = 0 \dots T\}$, with $S(t) = \mathbf{s}(t)$. A compact description of both local fixational gaze movements, occurring within a selected region of the visual field, and saccadic relocations between regions can be given in terms of a particle randomly wandering but being pulled towards an attractor (the center of a region of interest, RoI) can be formalised as a switching Ornstein-Uhlenbeck (O-U) process. To such end, denote $u_t \in [fix, sac]$, a switching state control random variable, indicating whether at time t either a fixational movement within a RoI or a saccadic relocation between RoIs is performed by the observer (D'Amelio and Boccignone, 2021; Boccignone et al., 2020). Then, the stochastic differential equation (SDE)

$$d\mathbf{s}(t) = \mathbf{B}^{u_t}(\boldsymbol{\mu}^{u_t} - \mathbf{s}(t))dt + \boldsymbol{\Gamma}^{u_t} d\mathbf{W}^{u_t}(t) \tag{7.12}$$

accounts for the switching O-U process dynamics within and between RoIs sequentially selected by the observer as foci of attention. The term $\mathbf{B}^{u_t}(\boldsymbol{\mu}^{u_t} - \mathbf{s}(t))$ represents the drift towards the attractor point $\boldsymbol{\mu}^{u_t}$, where the 2×2 matrix $\mathbf{B}^{u_t}$ $\mathbf{B}^{u_t} = \begin{bmatrix} B_{ii}^{u_t} & B_{ij}^{u_t} \\ B_{ji}^{u_t} & B_{jj}^{u_t} \end{bmatrix}$

controls the magnitude of the attraction effect; B_{ii} and B_{jj} represent the drift of the process towards the attractor in the i (horizontal) and j (vertical) dimensions, respectively, while the off-diagonal elements $B_{ij} = B_{ji} = \rho_B \sqrt{B_{ii} B_{jj}}$ encode the cross-correlation between drift in both dimensions. The stochastic term $\mathbf{\Gamma}^{u_t} d\mathbf{W}^{u_t}(t)$ accounts for diffusion. Akin to $\mathbf{B}^{u_t}$, the 2×2 matrix $\mathbf{\Gamma}$ is the control parameter (variances and covariances) of the two driving white noise processes (horizontal and vertical) described by $d\mathbf{W}(t)$. Higher values of variances/covariances generate noisier/more anisotropic gaze trajectories. Given the set of parameters $\boldsymbol{\theta} = \{u_t, \mathbf{B}^{u_t}, \mathbf{\Gamma}^{u_t}, \boldsymbol{\mu}^{u_t}\}$, the simulation of a sequence of eye movements $\mathbf{s}(t) \rightarrow \mathbf{s}(t')$, with $t' > t + \delta t$, δt being an arbitrary time step, can be obtained by solving Equation 7.12. In generative form, the solution can be written as the conditional sampling of $\mathbf{s}(t')$ given $\mathbf{s}(t)$, i.e., $\mathbf{s}(t') \mid \mathbf{s}(t) \sim P(\mathbf{s}(t') \mid \mathbf{s}(t))$, where the distribution $P(\cdot)$ is the Normal distribution $\mathcal{N}(\cdot)$ (see e.g. Kloeden and Neuenkirch (2013)):

$$\mathbf{s}(t') \mid \mathbf{s}(t) \sim \mathcal{N}(\boldsymbol{\mu}^{u_t} + e^{-\mathbf{B}^{u_t}\delta t}(\mathbf{s}(t) - \boldsymbol{\mu}^{u_t}), \mathbf{\Psi}^{u_t}), \tag{7.13}$$

where $\mathbf{\Psi} = \boldsymbol{D}^{u_t} - e^{-\mathbf{B}^{u_t}\delta t}\boldsymbol{D}^{u_t} e^{-\mathbf{B}^{u_t T}\delta t}$; $\mathbf{B}^{u_t}$ and $\boldsymbol{D} = \frac{\mathbf{\Gamma}^2}{2}\mathbf{B}^{-1}$ are 2×2 matrices and the form $e^{-\mathbf{M}}$ denotes the matrix exponential.

The emulator $\widetilde{\mathcal{M}}$ is thus fully specified by the set of parameters

$$\widetilde{\mathcal{M}} = \{u_t, \mathbf{B}^{u_t}, \mathbf{\Gamma}^{u_t}, \boldsymbol{\mu}^{u_t}\}$$

which gives a complete description of gaze dynamics. Emulator's parameters can be inferred as follows.

First, the raw eye-tracking data of an individual's gaze trajectory is parsed via the NSLR-HMM algorithm (Pekkanen and Lappi, 2017), to provide a set of fixational events (saccades are here discarded).

Only the fixations dwelling inside the diagnostic regions, i.e. eyes vs. mouth/nose (cfr Section 7.3.1) are retained. Call $\mathbf{e}^{eye} = [e_1, ..., e_{F_1}]$ the ensemble of F_1 fixations inside the eyes diagnostic area and $\mathbf{e}^{mn} = [e_1, ..., e_{F_2}]$ the group of F_2 fixations inside the mouth/nose diagnostic area. Define $\boldsymbol{\xi} = [\mathbf{e}^{eye} | \mathbf{e}^{mn}]$.

Consider the slice $\mathbf{s}^e = [\mathbf{s}_m, ..., \mathbf{s}_q]$ of the sample $\mathbf{s}(t)$, with $m \geqslant 0$ and $q \leqslant n$; the e index represents a generic fixation $e \in \boldsymbol{\xi}$. The likelihood of the slice, given the parameters $\{\mathbf{B}^e, \mathbf{\Gamma}^e\}$ writes $P(\mathbf{x}^e \mid \mathbf{B}^e, \mathbf{\Gamma}^e) = \prod_{i=1}^{q-m-1} P(\mathbf{s}^e_{i+1} \mid \mathbf{s}^e_i, \mathbf{B}^e, \mathbf{\Gamma}^e)$. Then, the posterior probability of the O-U parameters of the event e given the gaze trajectory slice is recovered via Bayes' theorem

$$P(\mathbf{B}^e, \mathbf{\Gamma}^e \mid \mathbf{s}^e) = \frac{P(\mathbf{s}^e \mid \mathbf{B}^e, \mathbf{\Gamma}^e) P(\mathbf{B}^e, \mathbf{\Gamma}^e)}{P(\mathbf{s}^e)}, \tag{7.14}$$

where under the mean field approximation $P(\mathbf{B}^e, \mathbf{\Gamma}^e) \approx P(\mathbf{B}^e) P(\mathbf{\Gamma}^e)$, the LKJ distribution is adopted as the prior for the $\mathbf{B}^e$ and $\mathbf{\Gamma}^e$ matrices in order to ensure all positive eigenvalues. Next, the event parameter posterior in Eq. 7.14 is computed in approximate form via Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2017) and summarised through its sample average and uncertainty (Highest Density Interval, HDI). The distribution summaries are joined together, thus yielding the vector $\mathbf{v}^e_{(id)}$ for each subject $id \in [1, ..., ID]$, ID being the total number of subjects:

$$\mathbf{v}^e_{(id),k} = [B^{avg,e}_{ii}, B^{hdi,e}_{ii}, B^{avg,e}_{ij}, B^{hdi,e}_{ij}, B^{avg,e}_{jj}, B^{hdi,e}_{jj}, \Gamma^{avg,e}_{ii}, \Gamma^{hdi,e}_{ii}, \Gamma^{avg,e}_{ij}, \Gamma^{hdi,e}_{ij}, \Gamma^{avg,e}_{jj}, \Gamma^{hdi,e}_{jj}]. \tag{7.15}$$

Eventually, the sequence of events (fixations) - each event e being summarised by the vector $\mathbf{v}^{e}_{(id),k}$-, characterises the visual behaviour of observer id while scrutinising the stimulus k (image).

Denote:

- $\langle \mathbf{v}^{eye}_{(id),k} \rangle$ and $\langle \mathbf{v}^{mn}_{(id),k} \rangle$ the average fixation feature vector relative to either the eye or mouth/nose diagnostic region associated to the scan path (image) k:

$$\langle \mathbf{v}^{e^{eye}}_{(id),k} \rangle = \frac{1}{F_1} \sum_{a=1}^{F_1} \mathbf{v}^{e^{eye}_a}_{(id),k}, \qquad \langle \mathbf{v}^{e^{mn}}_{(id),k} \rangle = \frac{1}{F_2} \sum_{a=1}^{F_2} \mathbf{v}^{e^{mn}_a}_{(id),k} \tag{7.16}$$

- $\mathbf{v}_{(id),k}$ the descriptor of scan path k obtained by concatenating the two vectors above:

$$\mathbf{v}_{(id),k} = \left[\langle \mathbf{v}^{eye}_{(id),k} \rangle | \langle \mathbf{v}^{mn}_{(id),k} \rangle \right]; \tag{7.17}$$

- $\langle \mathbf{v}_{(id)} \rangle$ the summary descriptor of the visual behaviour of observer id, over the set of the K observed stimuli:

$$\langle \mathbf{v}_{(id)} \rangle = \frac{1}{K} \sum_{k=1}^{K} \mathbf{v}_{(id),k}. \tag{7.18}$$

The categorisation of each fixation to its eventual diagnostic region has been carried out utilising the pre-existing masks within the dataset, which were employed for the original study. The account for that was motivated by the potential diagnostic value of these facial regions. By dividing fixations into these two categories, we sought to capture the nuanced dynamics of visual attention within different facial regions and explore their specific contributions to the prediction of social anxiety levels. The extracted parameters and features were derived separately for fixations related to the eyes and fixations related to the region of the nose/mouth, enabling a more fine-grained analysis of the observer's gaze behaviour and its potential relevance for social anxiety prediction. The extracted gaze dynamics parameters were then used as inputs for a Random Forest (RF) classifier to predict the SIAS score. To simplify the prediction task and facilitate the interpretation of results, we transformed the social anxiety prediction problem into a binary classification task. Given that the SIAS scores range from 0 to 80, we decided to use the median score in our dataset, which was found to be 18, as the threshold for binarizing the scores. Individuals with SIAS scores equal to or above the median threshold were considered to have high social anxiety, while those below the threshold were classified as having low social anxiety. The transformation into a binary classification problem allowed us to use well-established classification algorithms, such as the Random Forest classifier, to predict social anxiety levels effectively. Before selecting the RF classifier, we conducted an evaluation of other classification algorithms, such as the Support Vector Machine (SVM) with a kernel of radial basis functions and the Linear Support Vector Machine Classifier (linSVM).

Results

We utilised data from a cohort of 43 participants out of the initial pool of 44 participants, as recordings from one participant resulted to be inoperable and had to be excluded from our study.

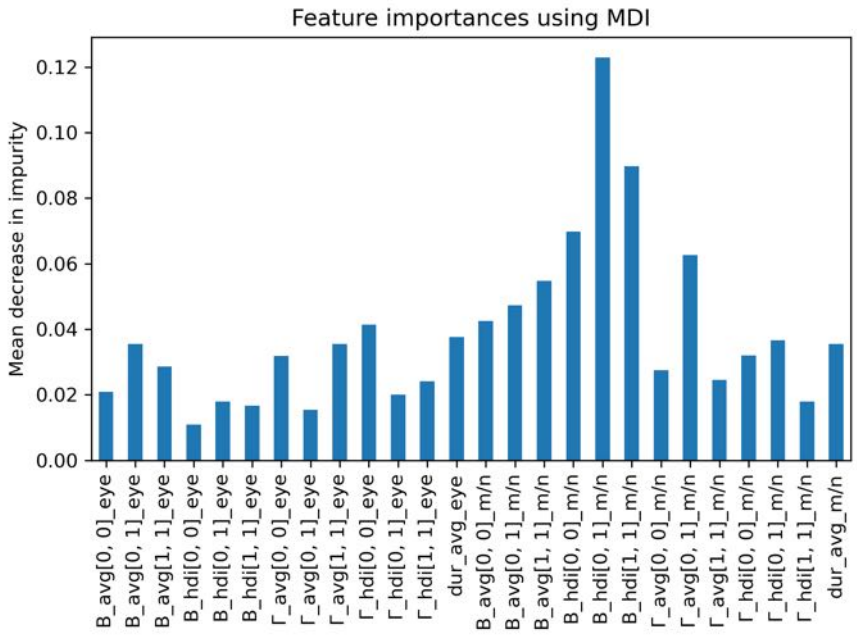


Figure 7.16: *Results of feature importance estimate, where indexes in brackets indicate the specific element of the matrix.*

The evaluation process involved 5-fold cross-validation to ensure robustness and mitigate any potential biases. Performance was assessed using accuracy.

The results of the evaluation revealed that the RF classifier outperformed the other algorithms with a significantly higher accuracy score of 0.73. In comparison, the linSVM and SVM algorithms yielded lower accuracy scores of 0.58 and 0.61 respectively 7.3.

Algorithm	Accuracy (5-fold cv)
linSVM	0.58
SVM	0.61
RF	**0.73**

Table 7.3: *Accuracy results of tested algorithms for SIAS classification.*

Additionally, we conducted an analysis to assess the feature importance, computed as the mean of accumulation of the impurity decrease within each tree (Fig. 7.16). This analysis provided valuable insights into the significance of different gaze dynamics features for the task at hand.

Notably, the results of this analysis revealed that features related to the gaze dynamics of the nose/mouth region hold greater importance in predicting social anxiety levels. Specifically, the feature that exhibited particularly high importance was the standard deviation of the magnitude of fixations drift towards the mean (B matrix).

Results Discussion

In this section, we have presented a systematic and principled approach for analysing eye movements in the context of fear generalisation, specifically focusing on the prediction of Social Interaction Anxiety Scale (SIAS) scores.

In the framework of foraging theory applied to eye movements, we have introduced a composite O-U process to operationalise social anxiety assessment. This phenomenological model captures the exploration-exploitation signature inherent in foraging eye behaviour. By inferring the relevant parameters of the composite O-U model through Bayesian analysis of eye-tracking data, we have identified a feature set that is suitable for predicting SIAS scores.

The results of our study demonstrate the effectiveness of the proposed approach. By utilising the inferred parameters from the composite O-U model of fixations as features, we have achieved promising performance in predicting SIAS scores using Random Forest (RF) as classification technique.

This research contributes to the body of knowledge by providing a novel perspective on the analysis of gaze for social anxiety assessment. By embracing a model-based approach and leveraging principles from foraging theory, this work can open novel avenues for future research in understanding and utilising gaze behaviour as a valuable modality for psychological and behavioural assessment.

7.8 Discussion

In this chapter, we explored how emotion and action can be incorporated into pain modelling, highlighting the theoretical and computational challenges that accompany these efforts. The models reviewed, especially those utilising active inference frameworks, effectively connect theoretical insights with practical scenarios in the field of pain research.

Particularly, exploring the complex interplay between pain and fear, we have examined the phenomenon of fear generalisation within pain contexts. The exploration of adaptive versus maladaptive fear generalisation offers significant insights into how pain-related fear can expand beyond the original stimulus, impacting patient outcomes in chronic pain conditions.

Introducing a simulation-based model that captures experimental setups and applying active inference paradigms helps concretise theoretical concepts into operational models. These approaches facilitate a nuanced understanding of pain beyond its immediate sensory interpretations, encompassing the broader affective responses and decision-making processes involved in pain experiences.

Moreover, our proposal of Class 2 models grounded in the presented theoretical framework highlighted the significant role of actions. This advances our understanding from static pain assessments to dynamic interactions involving actions and reactions within pain-related contexts. By recognising the importance of action, particularly through eye movement data that reflects decision-making processes in experimental settings, we align with the active inference theory which emphasises the balance between exploration and exploitation strategies in managing pain and fear responses.

This integration enhances our understanding of pain mechanisms and establishes a foundation for future research to investigate complex interactions in pain perception

and its modulation via cognitive and behavioural frameworks. The models discussed effectively align theoretical constructs with empirical data, thereby improving both predictive and explanatory capabilities of computational pain modelling.

However, it is crucial to heed a cautionary note regarding the limited agency of actions in the experimental context presented here, as participants rated their expectations without directly affecting the outcome. This scenario highlights how the absence of effective escape or control strategies can exacerbate stress and pain perception. Such conditions may lead to a negative anticipatory loop where the fear of pain intensifies the experienced pain, underscoring the importance of perceived control in therapeutic settings.

In scenarios where a Class 2 model is necessary, these models bridge the gap between simple stimulus-response mappings and those requiring complex social interactions. While Class 1 models focus solely on direct reactions to stimuli, Class 2 models incorporate the influence of actions and decisions, making them suitable for more dynamic contexts where pain experience is not only a reaction but also involves proactive strategies. However, they do not yet extend into the social dynamics of Class 3 models, which are necessary when the social context significantly alters pain perception or management strategies.

CHAPTER 8

Beyond the Individual: The Social Resonance of Pain

So far, we lingered on intrapersonal factors influencing pain experience, nevertheless, we live in social environments, so pain experience itself is connoted by interpersonal features that cannot be overlooked. That will be the focus of this chapter.

The recent surge of interest in the social relevance of pain within the scientific community signifies a noteworthy shift in the landscape of pain research and management. As previously discussed, the predominant focus in the field has traditionally leaned heavily toward sensory aspects, with affective dimensions receiving somewhat limited attention. However, it is only in recent times that the significance of social factors in understanding pain has gained substantial recognition.

This growing emphasis on the social component of pain reflects a broader acknowledgment of the complexity of human experiences related to health and illness. It underscores the idea that pain isn't solely a biological phenomenon, but rather a multifaceted experience intricately intertwined with an individual's social context. This shift has been prompted by the realisation that pain is not just a physiological response but a phenomenon deeply influenced by social determinants, cultural factors, and interpersonal relationships.

In essence, the importance of the social dimension in pain management and research has evolved into a central theme in contemporary discussions. It highlights the necessity of a more holistic approach to pain, one that recognises the interconnectedness of biological, psychological, and social factors in the experience of pain and well-being. This recent emphasis underscores the ever-evolving nature of scientific inquiry and the ongoing quest to comprehensively understand and address the complexities of pain in the modern world.

The first part of this section presents a supportive framework to elucidate the com-

plex interplay among psychological and social determinants in shaping pain experiences, with a particular focus on the dynamics of pain communication. Indeed, a comprehensive understanding of pain as a social phenomenon requires consideration of social or communicative (i.e. "expressive" and "receptive") features. The objective is to deepen our comprehension of immediate interactions during painful episodes within different social contexts, in particular clinical dimension. By delving into the intricacies of social transactions during pain, a more profound understanding emerges of the sociobehavioural dynamics involving both the individual experiencing pain and their caregivers. Pain communication, being an integral component of the pain experience, has evolved alongside language and non-verbal cues within human society. This section approaches pain communication through diverse perspectives, including the rational speech act (RSA) framework and an active inference viewpoint, to inform computational models of pain encompassing intrapersonal (sensory and psychological) and interpersonal (social) dimensions.

8.1 Biopsychosocial model of pain

The comprehensive model for understanding human health and illness, known as the biopsychosocial framework, emphasises the significance of considering biological, psychological, and social elements in the context of pain and well-being. In Figure 8.1, we can observe the various dimensions of this model, each playing a vital role in an individual's overall well-being. This approach has garnered support, particularly from proponents of multidisciplinary care for those grappling with chronic pain. It advocates for incorporating expertise from diverse healthcare professionals, including but not limited to medicine, nursing, physical therapy, psychology, social work, and rehabilitation.

The biopsychosocial perspective on pain has been recognised as necessary for research and practice if the full scope of pain is to be understood or care is to be effectively delivered to individuals in need (Engel, 1977; Robinson and Riley III, 1999). Figure 8.1 depicts the intersecting dimensions of biological, psychological and social determinants as they contribute to personal wellbeing. Substantial evidence supports this approach to pain (Turk and Okifuji, 2002; Gatchel et al., 2007), although attention to biological phenomena has outweighed the consideration of other factors.

In this work, we have incorporated interpersonal factors into the overarching model, recognising their crucial role in accurately reflecting that pain experiences always occur within a social setting. Here, the social context and the reactions of individuals witnessing the subject's experience are pivotal in shaping both the response to and the perception of the experience itself. The conceptualisation of pain, therefore, is deeply intertwined with sociality, highlighting its significance in understanding pain experiences.

Indeed, pain not only signals a threat, capturing the affected individual's focus (Eccleston and Crombez, 1999), but it also, via observable behaviours, garners the attention of those within the individual's social vicinity (Cano et al., 2008; Craig, 2004; Goubert et al., 2005; Hadjistavropoulos and Craig, 2002). The reactions of these observers play a crucial role in influencing the pain sufferer's experience and overall well-being, facilitating access to the benefits of social support and care. Therefore, to fully grasp pain as

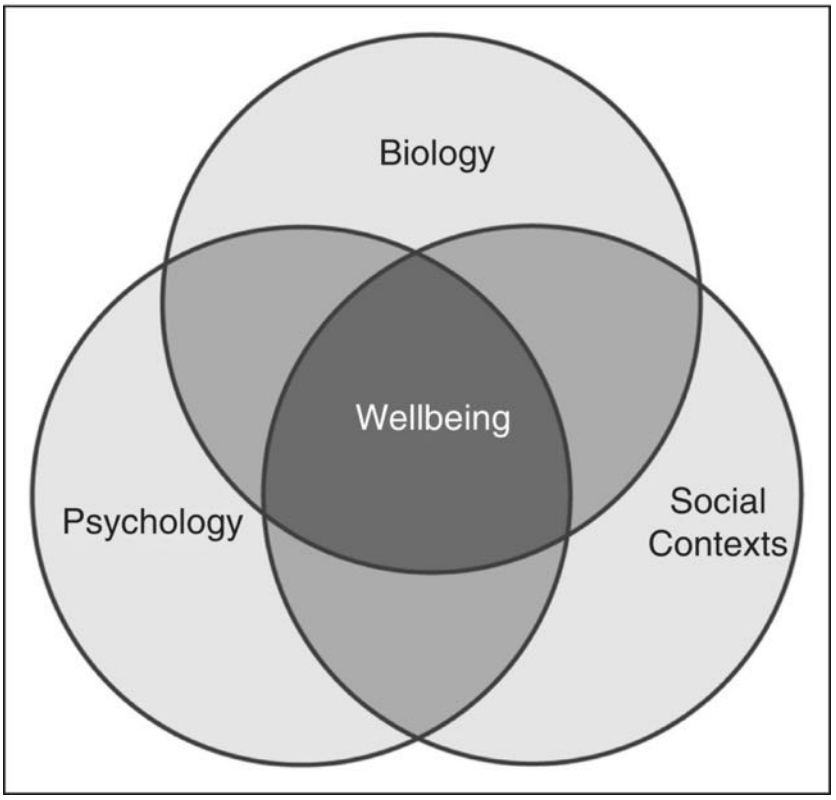


Figure 8.1: *The biopsychosocial model's approach to human illness and injury emphasises the importance of a holistic perspective. Achieving optimal well-being involves addressing not just the biological aspects, but also considering the psychological and social factors that play a significant role in an individual's experience of pain and associated disability. Figure from Craig and Versloot (2022).*

a socially embedded phenomenon, it's essential to explore its communicative aspects, both in terms of expression and reception.

8.2 Social communication model of pain

In general, communication theorists identify three distinct categories of communication, according to Duck and McMahan (2010):

- communication as action (or expression) (CaA), happening when a message is sent or received;
- communication as interaction (CaI), where the message is interpreted after sending and receiving it;
- communication as transaction (CaT), where messages are not only exchanged but also culminate in outcomes beyond the mere exchange itself.

Hence, CaT is conceptualised as a dynamic and interactive process where participants are simultaneously senders and receivers of information. This model emphasises the mutual influence and continuous exchange between involved agents, recognising

that communication shapes and is shaped by the context in which it occurs. It acknowledges that each participant in the communication process affects and is affected by the others, with the roles of sender and receiver constantly interchanging and evolving based on the feedback and interactions within the communication environment. This perspective views communication as an ongoing, fluid process rather than a one-way transmission of information, highlighting the importance of the relationship between communicators and the context of their interaction.

Applying this framework to the context of pain, whereas CaA is limited to the action of expressing pain (e.g. verbal reports and facial expressions), CaI implies a comprehension by the observer of the sufferer's expression, so that pain is serving as interactive communication and the process starts to be genuinely dyadic. When the context, beyond the understanding, allows some form of intervention: imagine a patient expressing their pain to a clinician, the latter having the capability of administer a treatment and, thus, the former can influence the clinician's decision. Indeed, during an interactive communication between a clinician and a patient, when the clinician deduces an underlying pathophysiological process, offers a diagnosis, and agrees to a course of treatment, pain plays a role in facilitating transactional communication.

Naturally, any dyadic communication can, in some way, become corrupted. This distortion, which can vary in degree, may be influenced by the intentions of the speakers. In the context of pain interaction, the sufferer might have an incentive to either exaggerate or understate their level of pain to affect the perception or response (in the case of transactional communication) of the observer. Similarly, the observer can also manipulate their estimation of the sufferer's pain, whether consciously or unconsciously. In addition to these potentially intentional distortions, unintentional distortions may also occur. These can stem from difficulties in encoding or decoding the message, such as through a vague report, a hard-to-interpret expression, or an inability to access adequate informational sources. Moreover, still in an unintentional manner, the interactants could be influenced by cognitive biases that prevent an appropriate evaluation of the message, whether being sent or received, misattributing the intent of the message.

In clinical environments, the propensity to ascribe intentions to communications may result in skepticism towards patients, especially in instances where clear pathophysiological causes for pain complaints are absent (Craig, 2007). This aligns with Burke (1966)'s insight that humans commonly interpret others' behaviours through the lens of motivation. However, even reflexive or automatic behaviours are indicative of pain, even if occur without conscious thought. For this reason, in this work, we accept a broader definition of communication. Our methodology encompasses both deliberately sent actions and spontaneous responses, along with both the deliberate and inadvertent reactions of recipients to these signals. Communication takes place within pairs or larger collectives, where the actions of the sender are tailored to the audience and influence the receivers. Customary communicative acts, such as speaking, which are purposefully intended and consciously perceived, often occur alongside unintentional behaviours.

Within the realm of pain communication, it is crucial to acknowledge that the experience and expression of pain straddle both innate biological predispositions and established social norms. While it is not entirely clear to what extent biological predispositions play a role in how we manage our own pain and that of others, the milieu of

social symbols and expectations plays a pivotal role in shaping an observer's response. Again, this backdrop of societal norms significantly informs our comprehension of how "communication" transpires among members of the same community or species (Hadjistavropoulos et al., 2011). For this reason, addressing these dimensions necessitates adopting a biopsychosocial approach.

Figure 8.2 unfolds this approach, mapping the intricate journey of pain from a personal sensation to a shared understanding. This model integrates multiple components that contribute to the experience, articulation, transmission, and interpretation of pain. Starting with a pain stimulus, which can be any noxious event that causes pain, the process then moves into *Internal Experience*, encompassing the personal and subjective experience of pain that's influenced by cultural, situational, intrapersonal, interpersonal, and social determinants. These determinants shape how the pain is felt and understood within the individual's unique social and cultural context. The internal experience leads to encoding, where the pain is converted into a communicable form (CaA). This involves behaviours, affects, cognition, and motivation which are rooted in basic neurological processes like nociceptive pathways from the spinal cord to the cortex, processing in the amygdala related to arousal, and appraisal mechanisms in the prefrontal cortex. Cognitive executive mediation, such as attention and memory, plays a role in ensuring message clarity.

On the receiving end, the communication is *decoded* involving the observer's interpretation, which is influenced by their own attitudes, abilities, and characteristics. This stage involves understanding the communicated pain as both interaction and transaction. The latter refers to what changes as a result of the communication—beyond the mere exchange of messages. For instance, in a clinical setting, the patient's expression of pain (encoding) may lead to specific actions by a healthcare provider, like providing pain relief (decoding). This might involve both automatic responses and more controlled responses informed by understanding the observed behaviour and its underlying affective and motivational states.

In summary, this model illustrates the complex interplay between biological processes, personal experience, and social context in pain communication, showing how various factors and processes influence each step of communication.

8.3 Dyadic interaction: a transactional perspective

Building upon the discussions presented so far, this section proposes a model of the interaction between an individual experiencing pain and an observer with different action prerogatives depending on their role. Specifically, as explored in various formulations herein, this dyadic interaction involves a subject likely in pain who can communicate their state to an observer through specific behaviours, expecting to elicit a reaction that improves their situation. For instance, if the observer is a healthcare professional, the sufferer expects appropriate treatment; if the observer is a family member, they anticipate empathetic support. Similarly, the observer relies on the communicative actions received from the sufferer to provide the assistance deemed appropriate.

In this model, communication is conceptualised as a transaction. As an agent experiencing pain engages in a communicative strategy, their goal is to influence the interlocutor to undertake an action believed to improve their condition. The same applies to

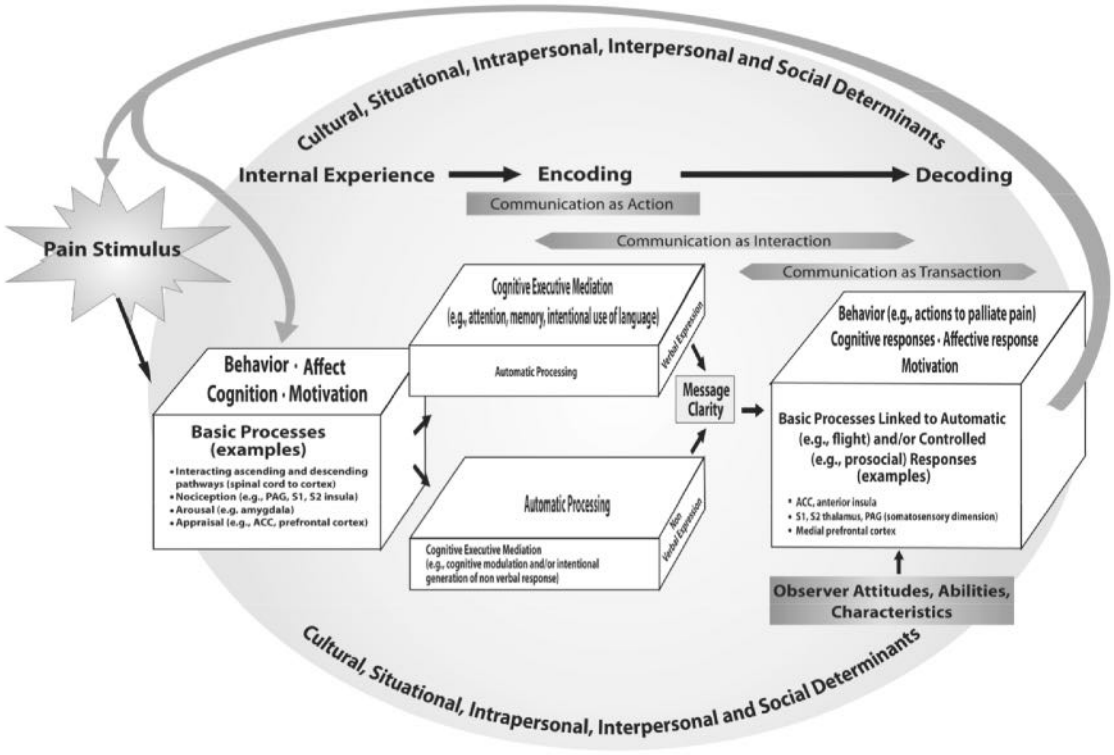


Figure 8.2: *The Communication Model of Pain. This image depicts a comprehensive model of pain communication, incorporating various elements that interact to form a full picture of how pain is experienced, encoded, communicated, and decoded. Figure from Hadjistavropoulos et al. (2011).*

the interlocutor, of course. It is clear, however, that this does not preclude the presence of the other two types of communication, which are actually assumed. In other words, this transactional perspective, rooted in the principles of social interaction, emphasises that both parties in the communication use their responses to either reinforce or modify each other's behaviour.

It is worth noting that, in a real-world scenario, each individual also communicates unintentionally through facial expressions, gestures, and prosody. These signals, even if not consciously sent, contribute to forming a message when a receiver is present. In the model we present, for simplicity, we have considered only facial expressions and verbal reports. Facial expressions can be categorised as a means of spurious communication: they convey both intentional and unintentional messages. Verbal reports, on the other hand, are the quintessential intentional medium, as verbal communication presupposes some level of planning.

Fig. 8.4 presents a model of dyadic interaction based on the theoretical model already presented in section 4.

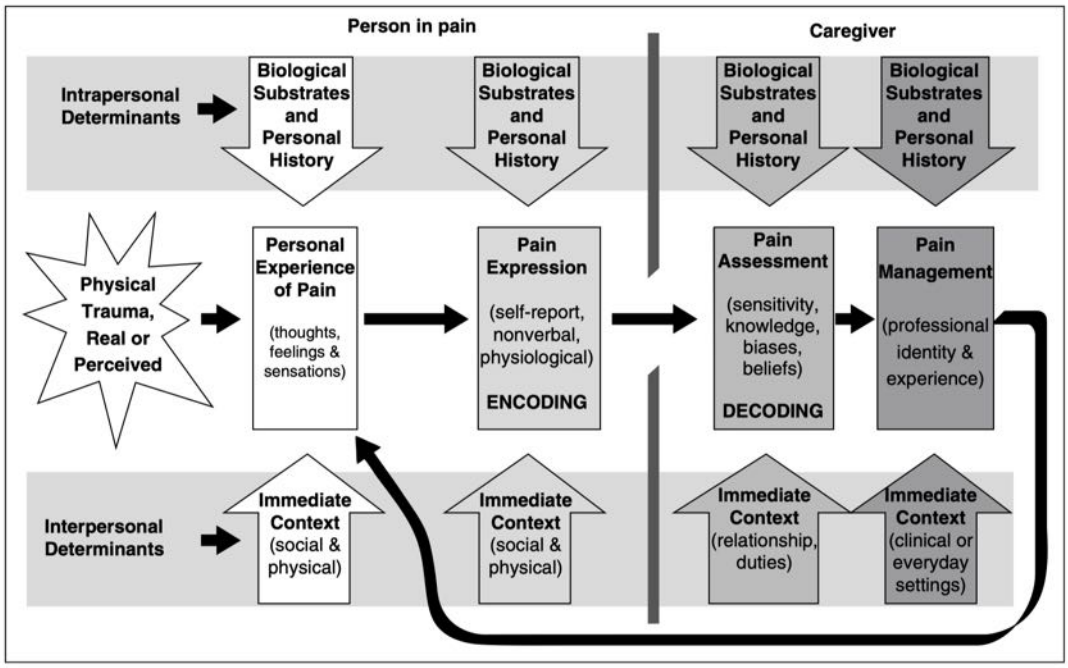


Figure 8.3: *The Pain Social Communication Model. This framework examines pain by considering its biological, psychological, and social aspects in the context of common injury and chronic pain escalation events. It also includes the reactions of caregivers and highlights both personal and interpersonal factors influencing each stage of the process. Figure from Craig and Versloot (2022).*

8.4 Modelling patient-clinician dynamics

8.4.1 Active inference POMDP model

Diving into the specifics of the implemented model, it is essential to understand the elements that define the various agents and the dynamics occurring between them. Different roles lead to different possibilities for action and perception. Thus, we present a specific illustrative scenario, involving the interaction between a patient experiencing pain and a clinician. Specifically, the model is inspired by the experimental setting presented by Kappesser et al. (2006) and, with the necessary adjustments and simplifications, it simulates the dynamics involved.

The experiment involved one hundred and twenty healthcare professionals observing videotaped facial expressions of pain patients and estimating their pain levels. The participants were divided into three groups: the first group was shown only the patients' faces, the second group also received patients' self-reports, and the third group received a contextual cue that primed them to anticipate deception, in addition to seeing the faces and receiving the patients' ratings. It aims to delve into an issue highly relevant in clinical settings, namely the underestimation of patients' pain by clinicians. This phenomenon, more pronounced in the absence of clinical signs, is often attributed to clinicians' overexposure to pain, leading to habituation (Choinière et al., 1990). However, the phenomenon becomes more pronounced when patients have certain char-

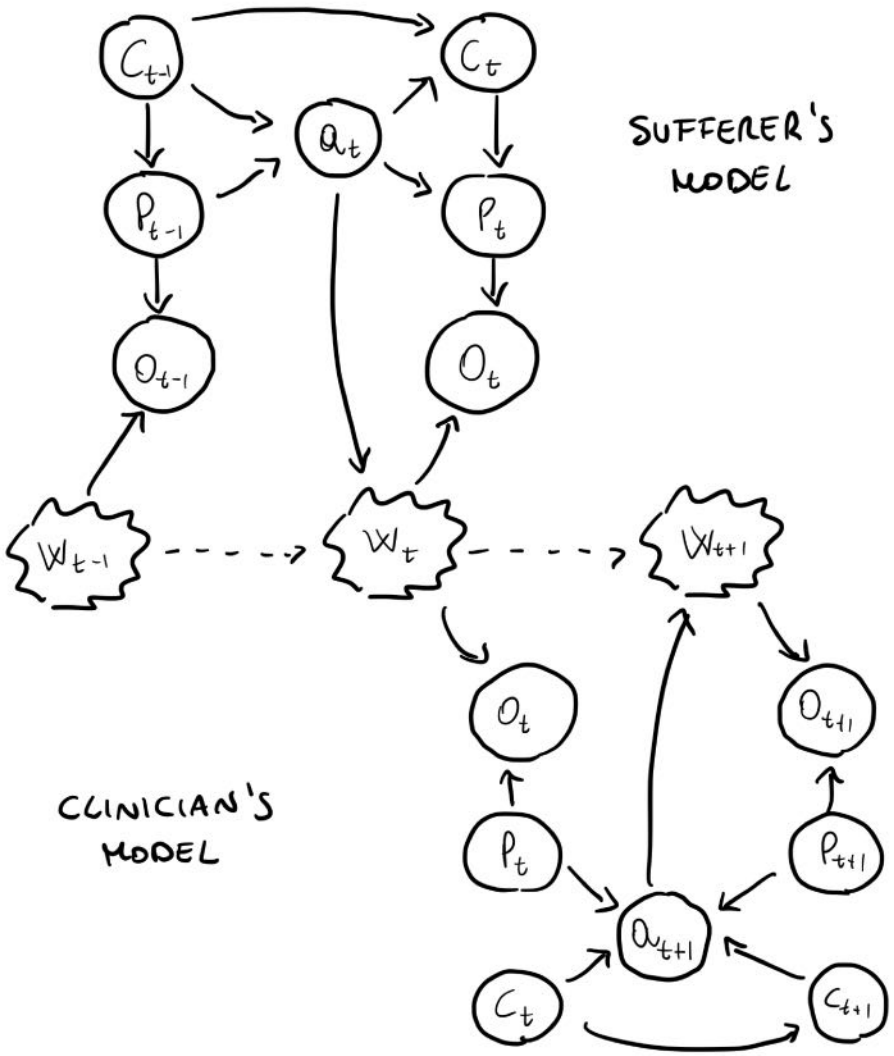


Figure 8.4: *A representation of the dynamic dyadic interaction in the form of the essential model presented in section 4. The point of contact between agents, namely communication, occurs through the impact in the world (W_t) that results from one agent's action, which is observed or experienced by the other agent.*

acteristics. Notably, biases in pain assessment are well-documented in the literature, particularly affecting female patients (Zhang et al., 2021), non-caucasian individuals (Hoffman et al., 2016), and those with a history of mental illness (Schäfer et al., 2016). Another element that we could regard as a common factor of the previous ones is the clinician's inherent estimation of the likelihood of deception, which is the focus of Kappesser et al. (2006).

The reference framework of that work is the Lens Model (Brunswik, 1952), a psychological framework that explains how individuals make decisions or judgments based

on limited and often ambiguous information. The model suggests that people use cues from the environment to make judgments. These cues are not always perfect indicators of the underlying reality but provide probabilistic hints that can guide decision-making (see Fig. 8.5).

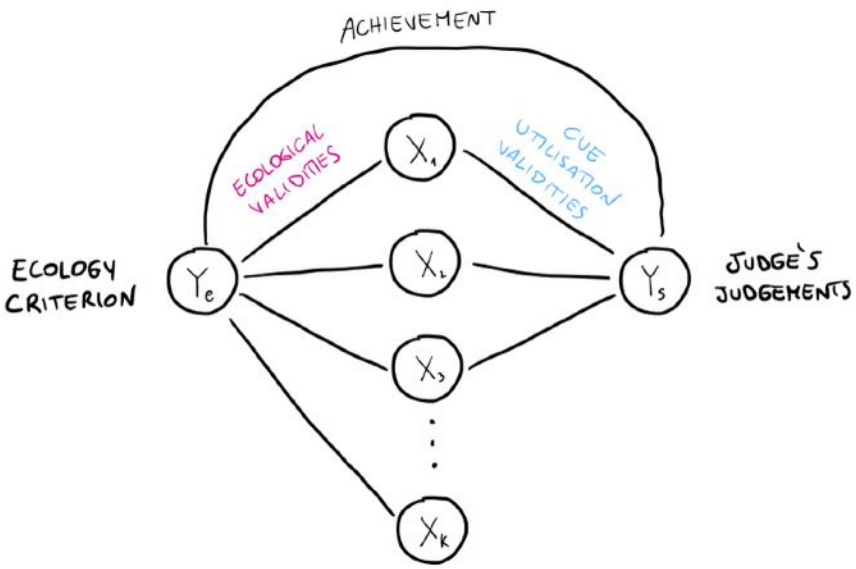


Figure 8.5: *A diagrammatic representation of the lens model. Ecological validity refers to how accurately each cue represents the value of the criterion. Utilisation denotes the relationship between each cue and the individual's judgement. Achievement is a metric that assesses how closely judgements align with the true values of the criterion.*

What is critical to note is that observers and pain sufferers use different information to assess pain, so their evaluations rarely align perfectly. Computationally, the statistical framework that best fits the problem at hand and that we have therefore chosen to use is the Partially Observable Markov Decision Process (POMDP), allowing us to model situations where the system's state is not fully observable. In a POMDP, an agent interacts with an environment in a series of time steps. At each time step, the agent:

- receives an observation: because the state of the environment is partially observable, the agent receives an observation that provides limited information about the true state;
- makes an inference on the hidden state using that observation, thus relying on the distribution $P(s_t|o_t)$;
- takes an action: based on the inferred state and policy, the agent chooses the action minimising the EFE. The action affects the environment;
- updates its beliefs: the agent updates its beliefs about the environment's state based on the action taken and the new observation received.

Variable type	Patient	Clinician
Factors (state)	pain = {low, high}	pain = {low, high}
Actions (policy)	expression = {relaxed, aroused} report = {low, high}	treatment = {psychological, pharmacological}
Outcomes (observation)	interoception = {relief, distress} exteroception = {safety, threat}	expression = {relaxed, aroused} report = {low, high}

Table 8.1: *List of model variables divided by agent's role: patient and clinician.*

It's important to note that in the case of dyadic interaction, each agent effectively serves as the environment for the other. Each agent possesses specific characteristics defined by the role they embody within the interaction, such as a patient or clinician in this particular case. Tab 8.1 shows the variables involved in both agents.

The patient agent is modelled with a single-factor internal state representing their perceived pain level, which can be classified into two discrete states: low and high. To infer their pain level, the patient relies on a combination of interoceptive and exteroceptive observations. The interoceptive observations include states of 'relief' and 'distress', while exteroceptive observations consist of 'safety' and 'threat'. These sensory inputs enable the patient to evaluate their pain condition effectively. Regarding actions, the patient can express their pain through two primary means: facial expressions and verbal reports. Both modalities of expression are binary, effectively communicating the pain level to external observers, specifically the clinician. See 8.6 for a detailed representation.

The clinician agent's objective is to accurately infer the pain level of the patient. The clinician's observations are primarily derived from the patient's actions: the verbal report and facial expressions, which are the outputs from the patient model. Additionally, the clinician considers a 'cheating cue', a critical element in the model that may indicate potential deception by the patient. This cue is taken into account for assessing whether the patient might be exaggerating their pain to obtain medications or, conversely, underreporting it due to fear of treatment or social conditioning influences. Both are well-known events in clinical contexts. Refer to Figure 8.7 for a detailed representation.

Through the integration of POMDP and active inference, the model facilitates a dynamic interaction between the patient and clinician, allowing for complex inferences about pain states that incorporate both observed actions and potential deceptive behaviours. Refer to Figure 8.8 for a schematic depiction of the interaction process between the two agents.

As revealed by the dynamics of the interaction, there is a substantial asymmetry between the two agents. While the action that allows the evolution of the internal state (the patient's pain) is essentially the treatment decided by the clinician, in the patient's generative model, the action that changes the state is their intentional reaction to the painful state. This is justified by the fact that in a transaction-based communication model, the agent holds the belief (thus a structured model) that they can influence the other's action through their communicative act.

In transitioning from the clinician's action, which is the treatment, to the patient's observations encompassing exteroceptive and interoceptive responses, we have intro-

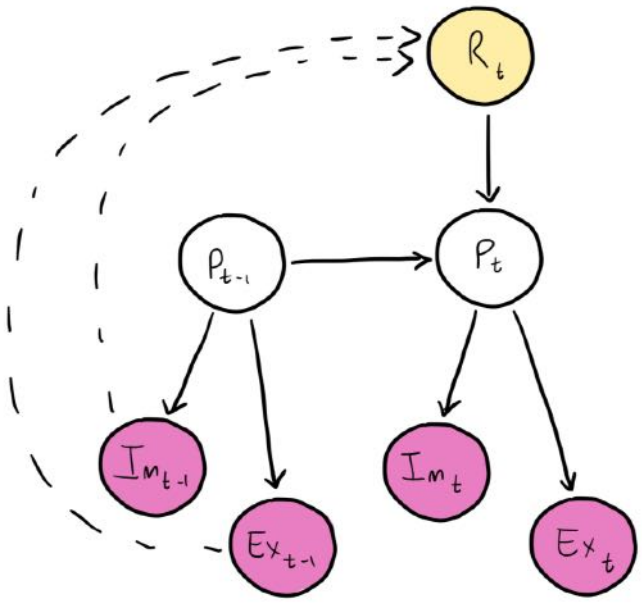


Figure 8.6: *POMDP patient's model. Highlighted in yellow is the action, specifically the ensemble of reports R_t, including verbal reports and facial expressions, which indirectly influence the hidden state of the pain level P_t. The observations, represented in pink, encompass the ensemble of interoceptive and exteroceptive reactions C_t occurring in the patient.*

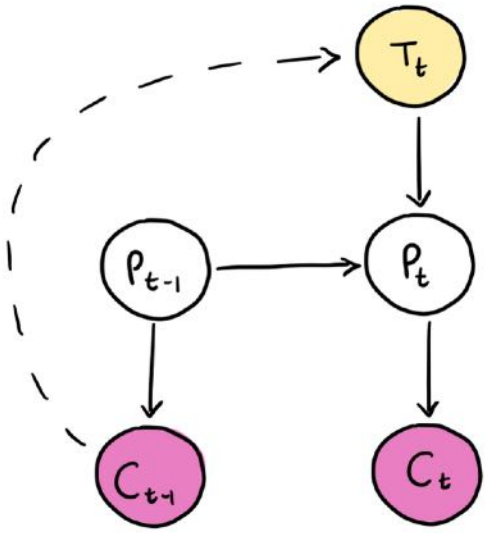


Figure 8.7: *POMDP clinician's model. Highlighted in yellow is the action, specifically the treatment T_t, which affects the hidden state of the patient's pain level P_t. The observations, represented in pink, comprise the ensemble of cues C_t visible to the clinician.*

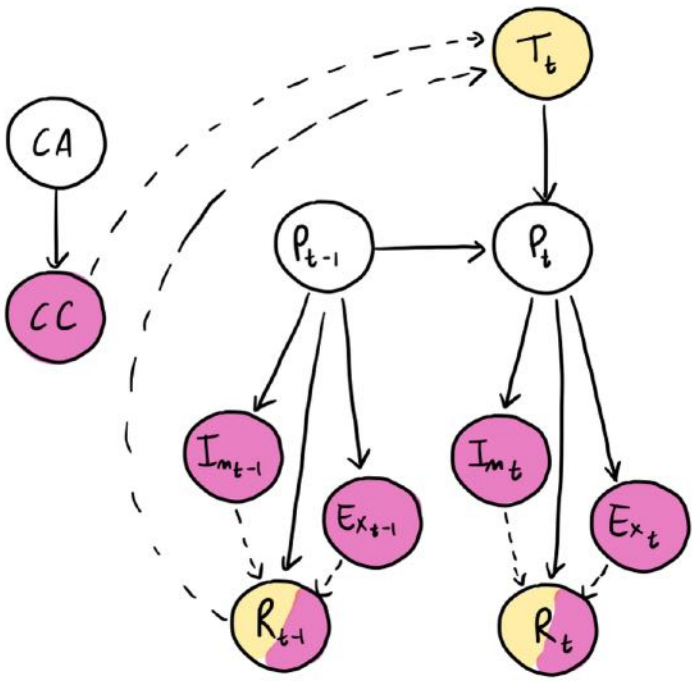


Figure 8.8: *Diagram of the overall generative process in the interaction. The patient's pain, which triggers interoceptive and exteroceptive reactions, evolves based on the treatment administered by the clinician. This treatment is influenced by the patient's reports and by the cheating cue CC, an observation arising from the patient's cheating cue CA. The variables depicted in pink are primarily observations, while those in yellow are actions. The report variable R_t serves a dual role, acting as an action from the patient's perspective and an observation from the clinician's perspective.*

duced a stochastic mapping function. This function essentially represents a probability distribution that guarantees the alignment of the clinician's interventions with the patient's consequent perceptions. The mapping can be defined as follows:

Let's consider the treatment given by the clinician as T and the observable responses of the patient as O, with O comprising both exteroceptive and interoceptive elements. The stochastic mapping function M then defines the likelihood of observing a particular set of responses O, given a specific treatment T. The mapping function M can be formally expressed as a conditional probability distribution:

$$P(O_t|T_t, O_{t-1}) = M(T \rightarrow O) \tag{8.1}$$

Here, M could take into account various factors such as the patient's individual pain threshold, the effectiveness of the treatment, and other environmental and psychological factors that could influence the patient's perception of pain. This stochastic nature of M acknowledges the inherent uncertainties and variabilities in how different patients might respond to the same treatment.

The relationship between involved variables and related distributions is encoded in

matrices A, B, C, and D. These matrices are tailored to each agent—clinician and patient—reflecting their unique perspectives and actions.

Matrix A, the observation matrix, governs the probability of observing a certain outcome given the true state of the system. For the patient, this could relate to the probability of perceiving relief or distress (interoception) and safety or threat (exteroception) as outcomes of treatment. It represents the observation likelihood $P(o_t \mid s_t)$. For the clinician, it might represent the likelihood of observing certain verbal reports or expressions of pain from the patient.

Matrix B, or the state transition matrix, defines how the internal state of an agent changes over time, representing the distribution $P(s_t \mid s_{t-1})$. In the patient's case, this matrix determines the transitions between different levels of pain states, for example, from low to high pain or vice versa. For the clinician, the state transition matrix might define changes in belief states regarding the patient's pain level or treatment strategies.

To guide behaviour towards specific objectives within active inference, it is crucial to incorporate a representation of a desired state or objective within the generative model. Unlike reinforcement learning, which utilises reward functions, active inference employs a prior distribution over observations, also termed as 'prior preferences' or 'goal distribution'. The decision-making process is influenced by this distribution of preferences, encoded in matrix C prompting agents to select actions that are anticipated to result in outcomes aligned with their preferred observations. Similarly, for the other two matrices, there is a distinction between the clinician and the patient.

Lastly, matrix D represents the agent's beliefs about the distribution over hidden states at the first timestep of the time horizon, so the prior $P(s_0)$.

8.4.2 Simulations

"Pseudo"-interaction experiments

Before commencing comprehensive simulations involving both the patient and clinician as agents, an initial set of simulations was conducted featuring only the clinician as the active agent, with the patient being treated merely as part of the simulation environment. This approach facilitated the replication of the experiment by Kappesser et al. (2006), which examined two significant factors influencing the underestimation of patients' pain by healthcare professionals: the absence of direct communication with patients and the anticipation of deceitful behaviour by patients.

In the referenced study, 120 healthcare professionals assessed the pain levels of patients by viewing videotaped facial expressions. The participants were divided into three groups: the first group observed only the facial expressions, the second group additionally received the patients' self-reported pain ratings, and the last group, besides viewing the facial expressions and self-reports, was primed with a context cue to anticipate deceitful behaviour.

The findings revealed that healthcare professionals generally underestimated the pain experienced by patients, but the degree of underestimation varied based on the provided cues. Observers who only saw the patients' facial expressions significantly underestimated pain compared to those who also had access to the patients' self-reports. Remarkably, healthcare professionals primed to expect deceitful behaviour underestimated pain as much as those who relied solely on visual cues, suggesting that both the

absence of verbal communication and an active suspicion of deceit could lead to the underestimation of pain.

Thus, the clinician in the preliminary simulations was modelled in three distinct conditions: observing only facial expressions (reflecting Group 1 from the study), observing facial expressions in conjunction with verbal reports (Group 2), and considering both previous cues plus a cheating cue (Group 3). These observations assist the clinician in inferring the patient's state (pain or no pain) and determining the likelihood of deceit regarding the reported state.

The preliminary simulations conducted yielded results that mirror the observable patterns from the reference study by Kappesser et al. (2006). As depicted in Fig. 8.9, in the absence of a cheating cue, we can discern two notable trends: firstly, the first group, which only observes facial expressions without additional information, exhibits less confidence in their pain assessments; this is understandable given their more limited informational context. The other two groups, privy to both facial and verbal cues, display greater confidence in their judgments, and with the cheating cue indicating that the patient is truthful, their beliefs are almost congruent.

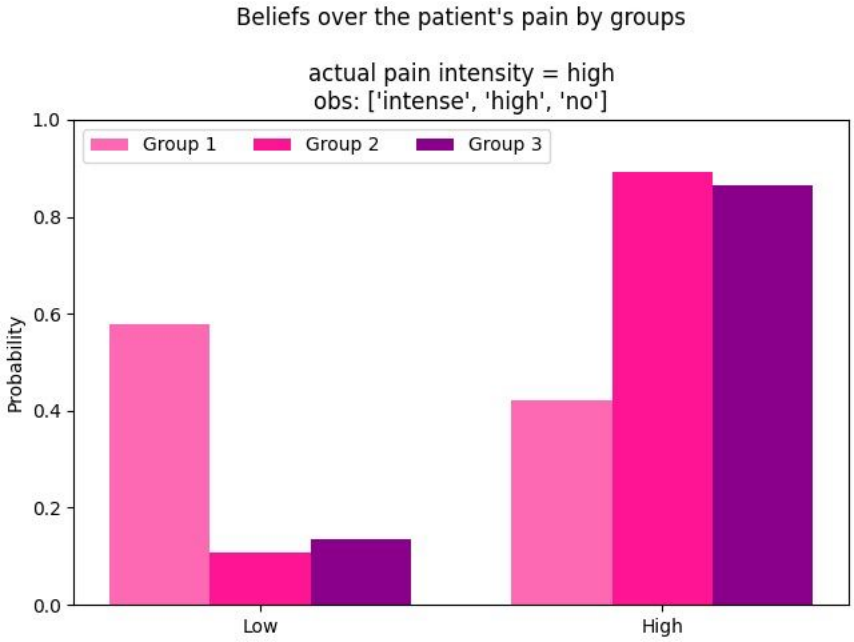


Figure 8.9: *The bar plot illustrates the clinician's beliefs regarding the patient's level of pain for each group, in the case of negative cheating cue, based on their access to information: the first group has access solely to facial expressions, the second to both facial expressions and verbal reports, and the third to all three cues—facial expressions, verbal reports, and the cheating cue.*

In contrast, Fig. 8.10 presents the scenario of an affirmative cheating cue. As the third group is the sole recipient of this cue, it can accurately identify deceit, leading to the inference of a lower level of pain when warranted. Furthermore, these simulations qualitatively replicate the findings of the referenced paper, particularly that the first and third groups are prone to underestimating pain when compared with the patient's self-reported pain, highlighting the influence of additional communicative cues and the

presence of a cheating cue on clinical pain assessment.

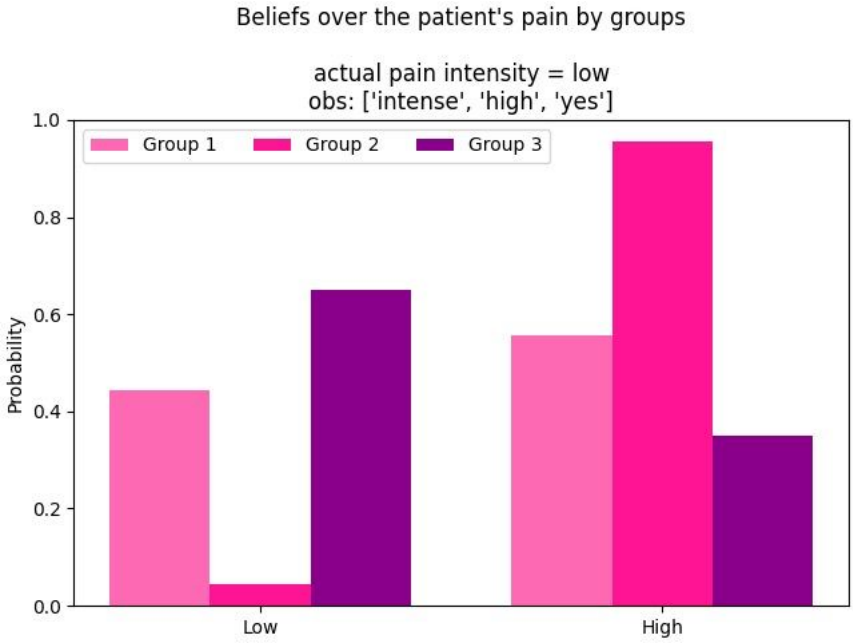


Figure 8.10: *The bar plot displays the clinician's beliefs about the patient's level of pain for each group, considering the presence of an affirmative cheating cue: the first group has access only to facial expressions, the second to facial expressions and verbal reports, and the third to all three cues - facial expressions, verbal reports, and the cheating cue, which affects their assessment of pain.*

These results uniquely display the clinician's inference side. Progressing further into the simulation, we generated data that accounts for the complete interaction, which includes not only the assessment but also the treatment selection and the patient's response. Moreover, the simulation extended to scenarios where the clinician could assess, diagnose, and treat multiple patients over time, specifically to examine how their prior probability distribution evolves. The intent was to investigate how repeated exposure to patients suffering from pain might alter a clinician's perception.

This phenomenon is often attributed to a habituation process, where clinicians become desensitised to pain due to repeated exposure, potentially shifting their internal scale of pain severity. Supporting this, some studies indicate that longer clinical experience correlates with greater underestimation of pain (Choinière et al., 1990). However, other research finds no such link (Dudley and Holm, 1984), suggesting the relationship between clinical experience and pain perception is complex and not universally applicable. This underscores the need for ongoing awareness and training in clinical settings to maintain sensitivity to patient pain and ensure appropriate pain management.

The initial results depicted in the plot of Fig. 8.11 illustrate the evolution of the clinician's prior probability that the patient's pain level is high, based on the clinician's prior belief before any patient interaction. As time progresses, it's notable that the clinician in the second group—who has access to both the patient's facial expression and verbal reports—initially sets a higher prior probability for severe pain. Interestingly, after 200 iterations, these beliefs stabilise without significant changes, suggesting a

consistent pattern where clinicians with access to more comprehensive patient information, and no cheating cue, tend to estimate higher levels of pain. This aligns with the findings mentioned earlier, where clinicians' prolonged exposure to patient pain doesn't necessarily alter their sensitivity or estimation accuracy over time. Instead, the type and amount of information accessible to clinicians play a crucial role in shaping their initial assessments of pain.

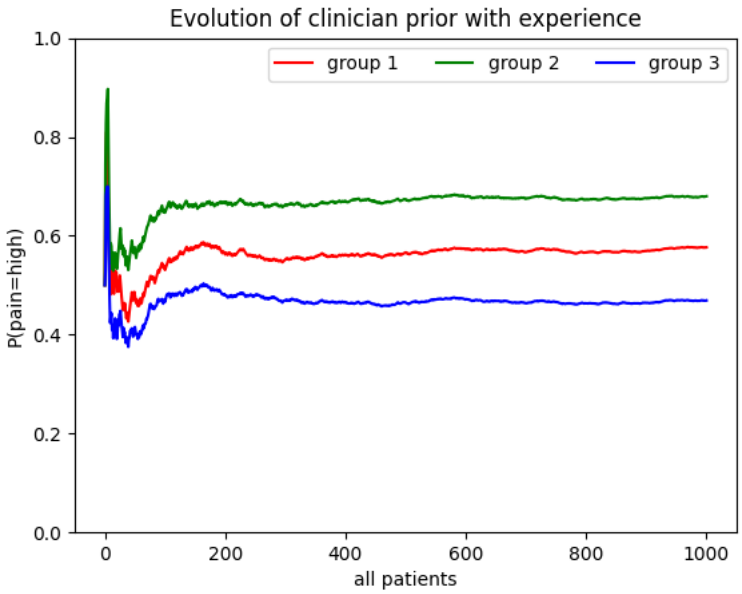


Figure 8.11: *The evolution of the clinician's initial belief regarding the patient's pain level over 1000 iterations, before observing any cues.*

In the Fig 8.12, we observe the evolution of the posterior probability that the patient's pain is high after the clinician's observations. The second group, which has access to both facial expressions and verbal reports, shows a slightly higher probability of high pain, illustrated in green. Intriguingly, the third group, depicted in blue and considering all cues including the cheating cue, exhibits three peaks instead of the usual two. This pattern suggests that the inclusion of the cheating cue may introduce additional uncertainty into the clinician's assessment, potentially leading to more fluctuating estimations of the pain level. This variability might indicate a struggle to ascertain the actual pain state due to the conflicting information the cheating cue provides.

However, in the simulations presented so far, the two groups without access to the cheating cue were modelled as being unaware of this variable's existence, lacking both the observation and state related to cheating, simulating a scenario of a "naive" clinician, unaware of this possibility. Subsequently, we redid the simulations, modelling all three groups identically concerning the cheating cue, setting it as negative in the first two groups and positive in the third. This adjustment allowed us to explore how the clinician's awareness or ignorance of potential deceit affects their diagnostic process and treatment decisions.

As shown in Fig. 8.13 and 8.14, the results are similar, but with awareness and uniform prior on cheating, both the prior and posterior probabilities of high pain slightly

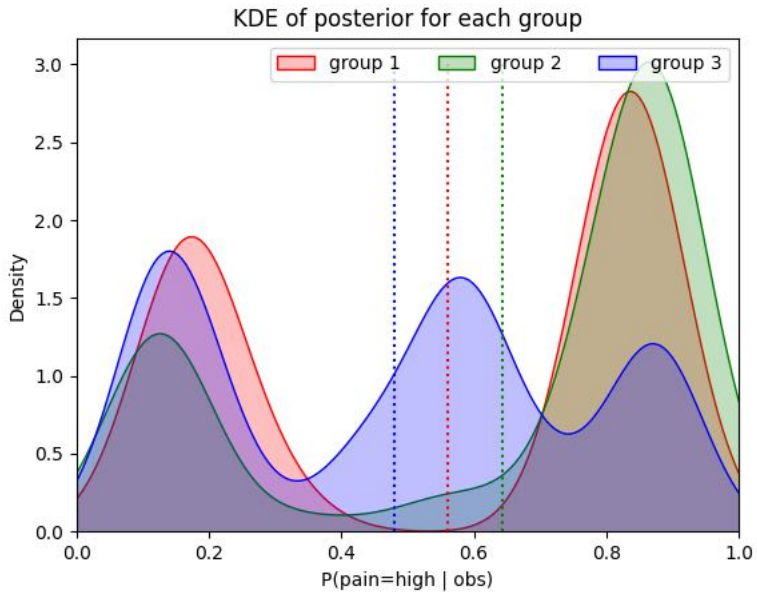


Figure 8.12: *KDE plot summarising the clinician's posterior belief on the patient's pain level after considering all available cues across multiple interactions.*

decrease for the groups involved in the modification (not over time, but in comparison to the unaware condition). This could suggest that these two groups are less inclined to assign a high level of pain, potentially due to increased suspicion or skepticism regarding the patient's reports.

Main interaction experiments

Proceeding with the main simulation, the interaction between the two agents—clinician and patient—was modelled as detailed in Algorithm 3. This simulation employs the framework previously established, engaging both agents in a structured sequence of actions and observations that reflect the dynamic and complex nature of clinical interactions.

The algorithm outlines the step-by-step process whereby the patient communicates their pain level through both facial expressions and verbal reports. Concurrently, the clinician observes these patient outputs, integrating them with the additional context provided by the 'cheating cue' to make an informed inference about the patient's actual pain level.

This interaction is iteratively simulated to reflect multiple exchanges within a typical clinical encounter, allowing for the analysis of the evolving understanding of the patient's state by the clinician, and adjustments in the patient's expressions based on the clinician's responses. The simulation serves not only to validate the model's conceptual accuracy but also to explore the implications of different interaction dynamics, such as trust and deception, within a clinical setting.

Figure 8.15 illustrates the results of the first simulation, in which the observation relating to the cheating cue is negative, meaning the clinician has no reason to believe

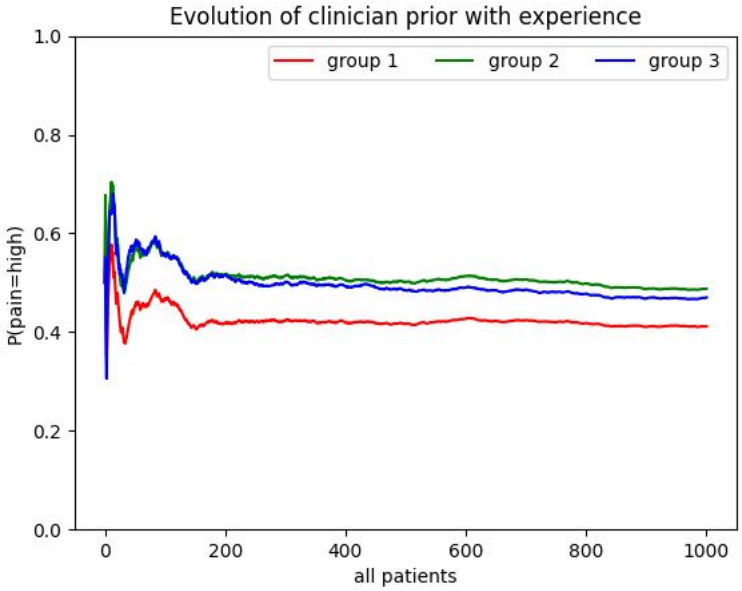


Figure 8.13: *The progression of the clinician's initial belief about the patient's pain level over 1000 iterations, before any cues are observed, taking into account their awareness of potential cheating.*

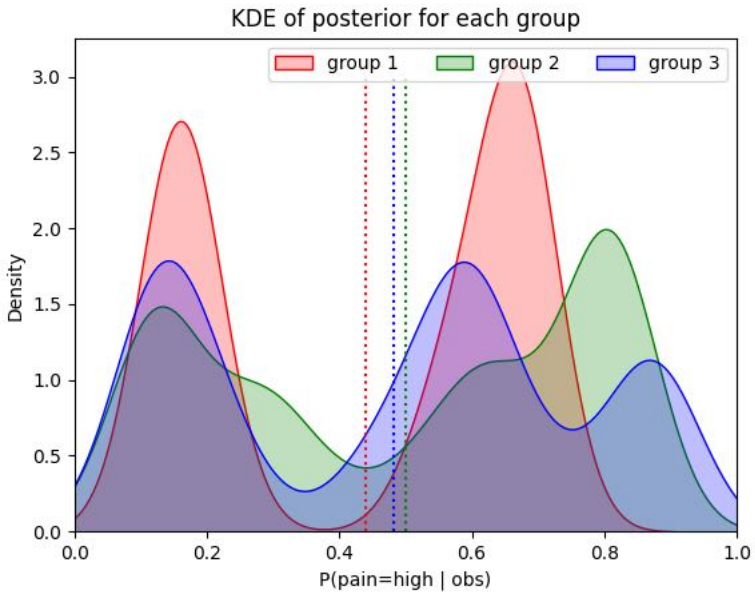


Figure 8.14: *KDE plot depicting the clinician's posterior belief on the patient's pain level after integrating all available cues across multiple interactions, considering their awareness of potential cheating.*

that the patient is lying about their pain level. Across 16 iterations, it is observable that, whether starting from a low or high pain level, the clinician generally makes a correct

Algorithm 3 Dyadic interaction

procedure $dyadic\_interaction$(patient, clinician)
  $n\_steps \leftarrow S$
  $patient\_action \sim patient\_q_\pi$ ▷ Pain reaction sampling from prior
  for `t in range(n_steps)` **do**
    $clinician\_obs \leftarrow patient\_action$ ▷ Report retrieval
    $clinician\_q_s \leftarrow clinician\_infer\_state(clinician\_obs)$ ▷ State inference
    $clinician\_q_\pi \leftarrow clinician\_infer\_policies(clinician\_q_s)$
    $clinician\_action \sim clinician\_q_\pi$ ▷ Treatment sampling
    $patient\_obs \leftarrow map(clinicia\_action)$ ▷ Stochastic mapping function
    $patient\_q_s \leftarrow patient\_infer\_state(patient\_obs)$
    $patient\_q_\pi \leftarrow patient\_infer\_policies(patient\_q_s)$ ▷ State inference
    $patient\_action \sim q_\pi$ ▷ Pain reaction sampling
  end for
end procedure

inference about the patient's state. In the latter scenario, the clinician also tends to rate the pain higher. Of course, since these are stochastic simulations, it is possible that the patient's pain level may increase even if the clinician administers the correct treatment.

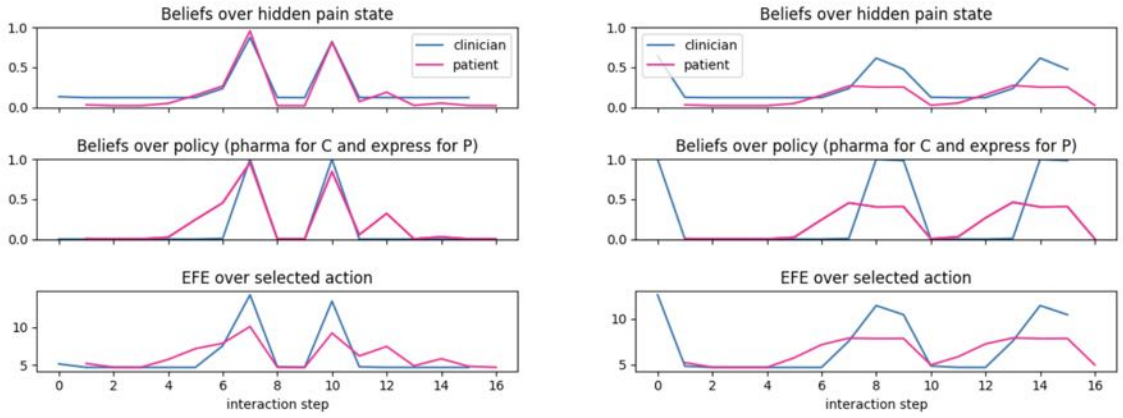


Figure 8.15: *Simulation Results with Negative Cheating Cue. The figure displays results from 16 simulation steps where the patient starts with either low pain (first column) or high pain (second column). The top row illustrates the probability that pain is present from both the clinician's and the patient's perspectives. The middle row depicts the probability of the chosen policy regarding pharmacological treatment from the clinician's viewpoint and the probability of reporting pain from the patient's side. The bottom row presents the Expected Free Energy associated with the selected action.*

Conversely, Figure 8.16 demonstrates how, in the presence of a positive cheating cue, the dynamics of inference change significantly. Particularly in the second plot, it appears that during actual pain events, the clinician's ratings are lower. However, in the absence of pain, the clinician tends to have a higher belief, possibly due to their observations being less robust compared to the patient's. This might reflect the clinician's

increased uncertainty in these situations.

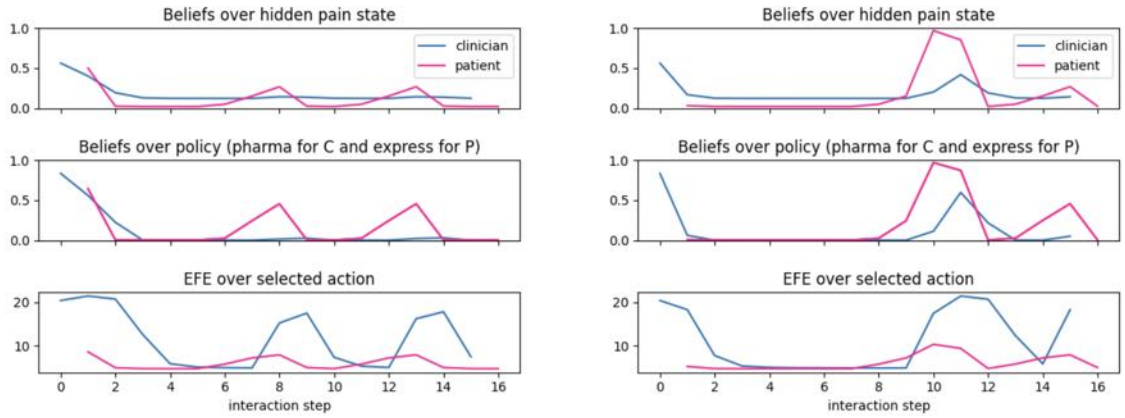


Figure 8.16: *Simulation Results with Affirmative Cheating Cue. This figure illustrates outcomes from 16 simulation steps, beginning with scenarios of both low and high pain levels in the patient (first and second columns, respectively). The top row indicates the probabilities assessed by both the clinician and the patient regarding the presence of pain. The middle row details the likelihood of the clinician opting for pharmacological intervention and the patient's likelihood of reporting pain. The bottom row demonstrates the Expected Free Energy for the chosen actions.*

8.5 Pain communication and understanding as a pragmatic problem

8.5.1 The RSA framework

The Rational Speech Act (RSA) model provides an agent-based methodology for formalising pragmatic reasoning, portraying listeners and speakers as engaged in mutual, recursive inference about each other's intentions. Initially crafted to reflect Gricean conversational principles, this framework also accommodates modern theoretical advancements by integrating a relevance distribution that weights the likelihood of various messages (but cfr. Appendix D for a brief introduction to pragmatics).

The basic architecture is the following (cfr. Figure 8.17).

The task of the listener L is to estimate the probability of a particular intended message m given the observed utterance u by the speaker, which we notate $P_L(m \mid u)$. Here, the m conveys information over the states of affairs (generically, the "world") as conceptualised by the speaker. By convention, the utterance u comprises linguistic as well as nonlinguistic components.

The listener is assumed to compute the posterior probability P_L via Bayesian inference through the integration of two components, the likelihood of the utterance given the message and the prior probability of the message:

$$P_L(m \mid u) \propto P_S(u \mid m)P(m).$$

The characteristic feature of RSA is the way that the likelihood term P_S (representing the speaker) is computed. The listener L is assumed to have an internal model of

Figure 8.17: *Schematic overview of the pain communication process between a patient (initially the speaker) and a caregiver (listener) during which different sources of information are integrated. Modified after Bohn and Frank (2019).*

the speaker S, who is modelled as choosing their utterance by maximising their own utility $U_S(u; m)$:

$$P_S(u \mid m) \propto \exp \alpha U_S(u; m).$$

The scalar value α can be interpreted as an indicator of how rational the speaker is in choosing utterances (i.e., how strongly they prefer the higher utility option). The speaker's utility is higher the more information they transmit through their utterance. Utility maximisation through cooperative communication reflects the central idea that humans communicate in a relevant (Sperber and Wilson, 1986) and cooperative (Clark and Brennan, 1991; Grice, 1989; Tomasello, 2010) way.

The utility of an utterance in turn depends on how much epistemic certainty it provides to the listener:

$$U_S(u; m) = \log P_{Lit}(m \mid u)$$

To avoid infinite recursion, the listener is taken to be a literal listener, say P_{Lit} who interprets utterances in accordance with their literal semantics:

$$P_{Lit}(m \mid u) \propto \delta_{[[u]](m)} P(m).$$

Here, $[[u]]$ is a semantic denotation for each sentence, concerning whether or not the utterance is true of a given message. $P(m)$ is the prior probability of the conveyed message. This prior term can be considered a distribution over relevant messages in context: it represents evidence for or against a particular message, independent of the utterance.

Through this recursive reference back to a listener, the model captures the interdependence of speaker and listener in communicative interactions (cfr., Figure 8.17). The combination of these two terms — speaker likelihood and prior — the listener's belief represents the outcome of a social-cognitive inference about the likely intended meaning of an utterance in context.

The RSA framework, which builds upon and synthesises a number of formal traditions in the study of human inference, from game theory to models of human reasoning it is well suited for our purposes. It is a description of the computational problem being solved by agents rather than being a model of a psychological process; thus, it provides a theoretical model.

Also, it is suitable to capture and formalise most relevant inferential theories of pragmatics that we have discussed in this section (Grice, 1989; Clark and Brennan, 1991; Sperber and Wilson, 1986; Tomasello, 2010). These theories have been immensely influential, but they are verbal descriptions of the psychological processes involved in communication, and the actual computations that lead to inference are not further specified.

RSA and its variants have now been used successfully to describe and predict a wide variety of phenomena, including implicature (Goodman and Stuhlmüller, 2013), hyperbole (Kao et al., 2014), vagueness (Lassiter and Goodman, 2017), generic language (Tessler and Goodman, 2019), and politeness (Yoon et al., 2020).

Importantly, RSA can offer a pragmatic perspective on language development. Bohn and Frank (2019) have argued for pragmatic reasoning supporting children's learning, comprehension, and use of language, providing evidence for developmental continuity between early nonverbal communication, language learning, and linguistic pragmatics.

8.5.2 RSA-based patient-clinician interaction

Within our model's context, the clinician (listener) and the patient (speaker) interact within a medical setting. The patient expresses their pain verbally and non-verbally, and the clinician, through careful observation, attempts to infer the severity of the pain being experienced.

Our modelling approach captures the immediate communication of the speaker's intent. At any given time t, we envisage distinct states that represent basic beliefs about pain, inferred from visible injuries such as cuts, burns, or rashes, or derived from medical records.

The model updates these beliefs based on observed actions A_{t-1} from the prior state (at time $t-1$), sampling from the following distribution: $B_t \sim P(B_t \mid B_{t-1}, A_{t-1})$ Here, B_t denotes a set of pain states $B_t = \{B_{s_1}, B_{s_2}, B_{s_3}\} = \{mild, moderate, severe\}$. While discussing the model, we generally omit the time index t for simplicity, except when specifically needed.

We define $P_S(s)$ as the prior probability of belief about these states, where $s \in \{s_1, s_2, s_3\}$:

$$P_S(s) = Cat(s \mid \pi_1, \pi_2, \pi_3) \tag{8.2}$$

In this scenario, the categorical distribution $Cat()$ and its parameters π_i determine the likelihood of each state s_i. For instance, a patient's medical records indicating no

previous pain might suggest a higher likelihood for mild pain.

If a patient reports mild pain, it reflects a direct communication of their current state. However, claims of moderate or severe pain might suggest that the patient is either genuinely concerned about their condition or possibly exaggerating.

The clinician, as the listener, is tasked with making the most accurate inference possible, considering all available information, which includes the affective state of the speaker. The model also allows for the possibility that the speaker may aim to express an affective state represented by $F = V \times A$, where V and A denote the valence and arousal components of their emotional state.

At the psychological level, the core affect state F offers a segmented representation of the complex, continuous affect space, showcasing the nuances of emotions felt by the individual. For simplicity, we assume a basic binary representation of F, focusing solely on arousal values within A:

$$a \in A = \{low, high\}. \tag{8.3}$$

This adds a layer of complexity to the interpretation of the patient's communications.

In summary, the model we introduce extends the RSA framework by incorporating a prior probability distribution $P(a \mid s)$ where $a \in A$ represents the influence of affect given a particular pain state. This enriches the model by considering both the literal and affective dimensions of pain communication, capturing the complex interplay between a patient's expressed and experienced states.

The priors relating to the binary random variable A, given the state s, are represented as:

$$P_A(a \mid s) = Bern(a \mid \pi^A(s_i)) \tag{8.4}$$

where $Bern()$ is the Bernoulli distribution with $\pi^A(s_i)$ parameter, identifying the probability of a related to the state s_i.

To engage in a communicative act, the speaker (the patient) must select a communicative objective, denoted as $g \in G \subset \mathcal{D}$, from their spectrum of potential desires contained within set D. Fundamentally, the goal maps the comprehensive semantic space $M = S \times A$ to a specific subset aligning with the speaker's intent M_X:

$$g_t : M \to M_X \tag{8.5}$$

Hence, the outcome manifests as a diverse subset M_X of states and affect, such as:

$$g(s, a) = g_s(s, a) = s \; g_a(s, a) = a \tag{8.6}$$

The prior distribution over goals $P_G(g)$ is chosen as a categorical distribution, with parameters π_i^G defining the probability of selecting the goal $g_i \sim P_G(g) = Cat(g_i \mid \pi_i^G, i = s, a)$.

The most straightforward scenario involves uniform sampling, where all π_i^G values are equal. In our scenario, the goal does not signify a true desire: articulating pain either objectively (g_s) or affectively (g_a) is not strictly intentional; thus, the goal here is more akin to an attitude.

For simplicity, we posit that utterances directly correspond to specific pain states, with each utterance u_i matching a state s_i, $u_i = s_i, i = 1, 2, 3$. These utterances are governed by a categorical prior distribution:

$$u_i \sim P_U(u) = Cat(u_i \mid \pi_i^U, i = 1, 2, 3), \tag{8.7}$$

where all π_i^U values are identical. It is assumed that the speaker articulates the exact utterance, while the listener perceives it clearly.

From the listener's (our clinician's) viewpoint, the communicative strategy $\mathcal{A}^{utt}$ implemented by the speaking patient can be outlined as follows. The speaker targets their message at an idealised and mentally constructed literal listener. This literal listener, termed L_0, serves as the baseline for recursive speaker-listener interactions. L_0 interprets the utterance u literally, ignoring the speaker's communicative goals and unable to detect any affective signals. Formally, the literal interpretation is a function $[[u]] : S \to Bool = \{0, 1\}$:

$$[[u]](s) = \delta_{u=s}. \tag{8.8}$$

The inferential process of the Literal listener L_0, considering the utterance u and the communication goal g, depends on the posterior computed as follow:

$$P_{L_0}(s, a \mid u, g) \propto \delta_{u=s} P_A(a \mid s) P_S(s). \tag{8.9}$$

The pragmatic speaker, labelled S_1, produces the utterance u, aligning itself with Gricean theories of communication (Grice, 1975), in order to be informative. This intent is tailored to a specific goal g, endorsing a principle of relevance that underscores the importance of pertinence in interpreting communications beyond their literal meanings.

The achievement of this goal is based on the literal listener's inference (Eq. 8.9), through which the speaker, via the posterior P_{S_1}, simulates the most suitable utterance u under the defined goal g, emulating the inferential process P_{L_0}:

$$P_{S_1}(u \mid s, a, g) = P_{Lit}(s, a \mid u, g) P_U(u). \tag{8.10}$$

Within the RSA framework, the speaker S_1 selects utterances based on a decision-making principle known as the softmax decision rule, described as follows:

$$P_{S_1}(u \mid s, a, g) \propto e^{\alpha U_1(u|s,a,g)} \tag{8.11}$$

Here, α is a parameter that modifies the speaker's rationality level. A higher α value leads the speaker's choices to align more closely with utility maximisation. At $\alpha = 1$ the equations simplify to:

$$U_1(u \mid s, a, g) = \log(P_{L_0}(s, a \mid u, g)^\alpha P_U(u)^\alpha). \tag{8.12}$$

The pragmatic listener L_1 updates their prior beliefs about the intended meaning $P_{L_1}(m)$ by interpreting the utterance u:

$$P_{L_1}(s, a, g \mid u) \propto P_{S_1}(u \mid s, a, g) P_G(g) P_A(a \mid s) P_S(s). \tag{8.13}$$

The clinician's initial model also considers non-verbal cues. The pragmatic listener assesses the speaker's affective state through their facial expressions, which adds depth to the interpretation of verbal clues.

It is assumed that the speaker and listener share a common context. Pain is subjective, making a shared context challenging to establish universally. However, medical records, test results, or visible injuries can serve as a common ground, influencing not only beliefs about pain states but also the affective interpretation.

The model analyses the speaker's facial expressions based on Action Unit (AU) expressions, linking them to affective states:

$$(v_F, a_F) \rightarrow AU_i \tag{8.14}$$

This informal theory of emotional behaviour helps the listener interpret the speaker's non-verbal cues. For simplicity, it is assumed that knowledge is acquired from datasets like those used in affect recognition tasks, enabling inverse inference:

$$(v, a) \leftarrow AU_i \tag{8.15}$$

For instance, logistic regression could be used based on probability distributions like:

$$P_{V_F}(v_F \mid AU) = Bern(v_F \mid \pi^V(AU)) \tag{8.16}$$

$$P_{A_F}(a_F \mid AU) = Bern(a_F \mid \pi^A(AU)) \tag{8.17}$$

These distributions provide expected values:

$$E[v_F \mid AU] = \pi_i^V(AU) \tag{8.18}$$

$$E[a_F \mid AU] = \pi_i^A(AU) \tag{8.19}$$

The listener evaluates emotional valence and arousal, considering both pre-established mental expectations and those inferred from facial expressions. Only arousal is considered here, with its posterior distribution approximated as follows:

$$\tilde{p}_A = \tilde{P}A(a \mid s, AU) \approx PA_F(a \mid AU)P_A(a \mid s) \tag{8.20}$$

After normalisation, this approximation resembles 'weak cue integration'. The resulting probability $\tilde{p}_A$ forms the basis for sampling posterior arousal:

$$a \sim \tilde{P}_A(a \mid s, AU) = Bern(a \mid \tilde{p}_A). \tag{8.21}$$

Various factors may influence the clinician's perceptions, including implicit biases and stereotypes that could affect their interpretation of the patient's pain, shaped by assumptions based on gender, race, age, or other personal characteristics. Past experiences and institutional pressures might also impact their judgement.

The model includes a gender-based stereotype, where gender observed by the clinician affects their assessment:

$$ge \in GE = male, female \tag{8.22}$$

It is often noted that women's pain might be underestimated, suggesting that clinicians might believe a female patient to be experiencing less severe pain (Zhang et al., 2021). This stereotype affects the clinician's prior distribution over states $P_S(s)$. Consequently, a secondary prior distribution is defined to account for gender-based expectations:

$$P_{S_{GE}}(s \mid ge) = Cat(s \mid \pi_1^S(ge), \pi_2^S(ge), \pi_3^S(ge)) \tag{8.23}$$

where $Cat()$ is a categorical distribution and parameter $\pi_i^S(ge)$ identify the probability of the state s_i given the gender ge. The posterior distribution of pain states, given the gender ge, is computed as follows:

$$\{\tilde{\pi}_1^S, \tilde{\pi}_2^S, \tilde{\pi}_3^S\} = \tilde{P}_S(s \mid ge) \approx P_{S_{GE}}(s \mid ge) P_S(s). \tag{8.24}$$

After normalisation, the obtained probabilities $\{\tilde{\pi}_1^S, \tilde{\pi}_2^S, \tilde{\pi}_3^S\}$ are used as parameters for the categorical distribution, to sample the posterior pain state defined as:

$$s \sim \tilde{P}_S(s \mid ge) = Cat(s \mid \tilde{\pi}_1^S, \tilde{\pi}_2^S, \tilde{\pi}_3^S). \tag{8.25}$$

Incorporating a broad range of factors, such as social, cultural, and communicative aspects, would complicate the clarity of the model. Therefore, a straightforward example was presented to demonstrate how these factors can influence a clinician's initial assumptions about pain states. In this instance, internal beliefs are shaped by observations. Still, it is also feasible to establish other prior distributions over states that are influenced by various factors or the clinician's implicit biases.

8.5.3 Implementation

We now provide an overview of the Clinician RSA model implementation outlined in the previous section. While we do not delve into specific code details, the pseudo-code provided below demonstrates the structural framework and operations of the model, as well as the probabilistic reasoning involved in the patient-clinician interaction.

Variables

First, let us delineate the variables of the model as outlined in Algorithm 4. As previously discussed in the theoretical framework, we categorise pain using three distinct classifications. Consequently, states and utterances within our model are characterised by these three categories of pain, ensuring a direct one-to-one correspondence between each state and its respective utterance. Following this, we specify the binary values for arousal, and subsequently define the parameters associated with goals.

The goals *state* and *affect* represent two ways of reporting pain. The first is the objective and direct reporting, while the second can interpret as an indirect speech act. The patient's perception of pain may be influenced in a non-conscious way by the affective state (e.g. during a treatment) or it is the patient who influences the reporting of pain. The reporting of pain differently from what it actually is, may be due either to asking for more help (over-reporting), or on the opposite, minimising due of fear or denial of the situation for example (under-reporting).

Algorithm 4 Model Data

$states \leftarrow \{mild, moderate, severe\}$
$utterances \leftarrow states$
$arausal \leftarrow \{low, high\}$
$goals \leftarrow \{state, affect\}$

Prior distributions and sampling

In Bayesian statistics and probabilistic modelling, a prior distribution represents the belief or uncertainty about the parameters of a model before observing any data. It encapsulates what is known or assumed, before considering any observed data.

Prior distributions are essential, because they introduce regularisation and encode existing knowledge or assumptions about the model. They are often subjective and may be influenced by domain knowledge, assumptions or available information.

In a real-world scenario, these priors could be modified based on empirical data, domain experience or learning from previous interactions. Priors play a crucial role in Bayesian inference, as they represent initial beliefs or assumptions about the likelihood of different outcomes. They are combined with observed data to derive posterior probabilities, which represent an updated belief after considering the evidence.

Adjusting priors based on domain knowledge or empirical data can improve the accuracy of the model in reflecting real-world scenarios.

The choice of a prior distribution influences the posterior distribution obtained through sampling. Different priors may lead to different posterior estimates. Generating samples from the prior distribution allows the parameter space to be explored, understanding the range of plausible values before considering the observed data.

Sampling refers to the method of generating observations from a specified probability distribution. This process is fundamental in probabilistic programming, where it involves drawing samples from probability distributions to estimate or simulate intricate models.

The utility of sampling extends to the simulation of data, estimation of parameters, and making predictions. Specifically, in the context of Bayesian inference, sampling is crucial for approximating the posterior distribution. This approximation is achieved by sequentially drawing samples from the prior distribution and evaluating these through the likelihood function, thereby facilitating a more comprehensive understanding of the model's behaviour under various conditions.

State

The prior distribution over states $P_S(s)$ (Eq. 8.2) represents the initial belief about possible states before any observation or information is taken into account. Specifically, it refers to the clinician's initial beliefs about the patient's true pain.

This prior is fundamental in that it encodes prior knowledge or assumptions about the probability of occurrence of various states. It shapes the agent's expectations of the world before any specific observation is made.

The algorithm 5 shows the state sampling function based on the prior distribution. In this case, the clinician is an objective clinician since the distribution over states is

uniform. Sampling is done by following a categorical distribution via Pyro's sample function, with the prior on the state as probabilities.

Algorithm 5 State Prior

procedure $state\_prior$
  $probs \leftarrow [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ ▷ Uniform prior probabilities
  $sampled \leftarrow sample(Cat(probs))$ ▷ Sample state
  return $sampled$
end procedure

If we want to include more prior distributions over states, for example the equation 8.25, the following procedure (algorithm 6) shows the implementation. This example considers three influencing factors, which are includes the context (injuries, medical records, etc.), an implicit clinician bias and the gender stereotype. The posterior distributions is calculated as normalised join distribution.

Algorithm 6 State Prior with multiple prior distributions

$state\_prior \leftarrow [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$
$context\_prior \leftarrow [\frac{1}{5}, \frac{1}{2}, \frac{1}{3}]$
$gender\_prior \leftarrow \{male : [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}], female : [\frac{1}{2}, \frac{1}{3}, \frac{1}{5}]\}$
procedure $state\_prior$(gender)
  $probs \leftarrow state\_prior * context\_prior * gender\_prior[gender]$ ▷ Integration cues
  $probs \leftarrow normalise(probs)$
  $sampled \leftarrow sample(Cat(probs))$ ▷ Sample state
  return $sampled$
end procedure

Arousal

The arousal prior probabilities signify the likelihood of experiencing high arousal associated with pain: more intense pains are more likely than milder ones. It is also essential to consider the arousal evident in the observed facial expressions.

To sample the arousal state, it is necessary first to differentiate between the sampling processes of the literal listener and the pragmatic listener. Only the pragmatic listener takes into account the facial expression, whereas the literal listener exclusively relies on interpreting based on prior probabilities.

The literal listener calculates P_A concentrating on the prior as depicted in Eq. 8.4. Conversely, the pragmatic listener calculates a posterior distribution by integrating the two prior distributions, $P_A(a \mid s)$ and $P_{A_F}(a \mid AU)$, which is subsequently normalised as shown in Eq. 8.21.

Algorithm 7 Arousal Prior

$arousal\_prior \leftarrow \{mild : 0.3, moderate : 0.7, severe : 0.9\}$
procedure $arousal\_sample$(state, $fe\_arousal$)
  $prior\_prob \leftarrow arousal\_prior[state]$ ▷ High arousal prob. of state
  if $fe\_arousal == None$ **then**
    $sampled \leftarrow sample(Bern(prior\_prob))$ ▷ Sample arousal
    return $sampled$
  else
    $arousal\_prior\_probs \leftarrow [prior\_prob, 1 - prior\_prob]$
    $arousal\_other\_probs \leftarrow [fe\_arousal, 1 - fe\_arousal]$
    $probs \leftarrow arousal\_prior\_probs * arousal\_other\_probs$ ▷ Integration cues
    $probs \leftarrow normalise(probs)$
    $prob \leftarrow probs[0]$ ▷ Select probability of high arousal
    $sampled \leftarrow sample(Bern(prob))$ ▷ Sample arousal
    return $sampled$
  end if
end procedure

Utterance

The prior distribution over utterances $P_U(u)$ reflects the listener's expectations or beliefs about which utterances are most likely to be produced by the speaker in a given context or situation. In our case, how the patient is more likely to produce a pain descriptor than another depending of the meaning they want to express.

The following pseudo-code shows a simple utterance sample function, where the probability prior distribution over utterances is uniform.

Algorithm 8 Utterance Prior

procedure $utterance\_prior$
  $probs \leftarrow [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ ▷ Uniform prior probabilities
  $sampled \leftarrow sample(Cat(probs))$ ▷ Sample utterance
  return $sampled$
end procedure

Goals

The prior distribution over goals $P_G(g)$ establishes the initial belief or expectation of each possible goal.

In the sampling procedure outlined in Algorithm 9, we have employed a uniform distribution for the sake of simplicity. However, similar to other prior distributions we have examined, this can be adapted to specific contexts to enhance accuracy. For instance, the goal prior might be updated based on the characteristics of a particular patient, who could exhibit a more objective and direct approach, or be more influenced by their affective state.

Algorithm 9 Goal Prior

procedure $goal\_prior$
  $probs \leftarrow [\frac{1}{5}, \frac{1}{5}]$ ▷ Uniform prior probabilities
  $sampled \leftarrow sample(Cat(probs))$ ▷ Sample goal
  return $sampled$
end procedure

In addition, we have to implement the function $g(s, a)$ (Eq. 8.6) that gives us the pain state or the affective state depending on the goal. The importance of this function lies in its ability to potentially accommodate both perspectives when interpreting or conveying information. It represents the speaker's intention to convey different types of information based on the goal, and the listener's interpretation of language strictly by its explicit or indirect meaning.

Algorithm 10 Goal State

procedure $goal\_state$(goal, state, arousal)
  if $goal == state$ **then**
    return $state$
  else if $goal == affect$ **then**
    return $arousal$
  end if
end procedure

Meaning function

The meaning function $\delta_{u=s}$ represents the core of the literal listener because it is used to interpret utterances strictly according to their literal semantic meaning. In our case, we have a direct one-to-one mapping between utterance and state, so the function simply compares the utterance and state, returning a Boolean value if they coincide or not.

Algorithm 11 Meaning Function

procedure $meaning\_function$(utterance, state)
  return $utterance == state$
end procedure

Literal listener

The literal listener L_0 computes the posterior distribution based on the literal interpretation of utterance and communication goal (see Eq. 8.9). L_0 computes first the prior distribution over states $P_S(s)$ and over arousal $P_A(a \mid s)$ Then, through the meaning function $\delta_{u=s}$ it interprets the utterance u.

If the literal meaning of utterance u matches the literal meaning of sampled state s, a zero log-factor is assigned; otherwise, a very negative log-factor (e.g. -9999.0) is assigned. This log-factor is used to update the probabilistic model: using the $factor$ function of Pyro, the calculated log-factor is incorporated into the probabilistic model,

thus influencing the probability distribution based on the literal interpretation of the utterance.

Finally, the value given by $g(s, a)$ associated with the goal g is returned.

Algorithm 12 Literal Listener

procedure $literal\_listener$(utterance, goal)
  $sampled\_state \leftarrow state\_prior()$ ▷ Sample state
  $sampled\_arousal \leftarrow arousal\_sample(sampled\_state)$ ▷ Sample arousal
  $goal\_value \leftarrow goal\_state(goal, sampled\_state, sampled\_arousal)$ ▷ Goal state
  if $literal\_meaning(utterance, state)$ **then** ▷ Literal interpretation
    $log\_factor \leftarrow 0.0$
  else
    $log\_factor \leftarrow -999999.0$
  end if
  $factor(log\_factor)$ ▷ Incorporate literal meaning
  return $goal\_value$
end procedure

Pragmatic speaker

The implemented pragmatic speaker (Algorithm 13) computes the posterior distribution over utterances generating, as described in Equation 8.11.

The scaling factor has been defined with its default value, indicating the minimum of deterministic in choosing utterance. The speaker uses probabilistic reasoning to select utterances that maximise the likelihood of successful communication within a given goal g, in the following way: the speaker chooses an utterance u through the prior distribution $P_U(u)$, then it computes the literal listener's distribution and conditions it with the intended meaning (given by $g(s, a)$) based on goal g. Specifically, the listener's distribution is scaled using the scaling factor and then conditioned making an observation respect the conveyed meaning. The model incorporates this conditioned distribution to update the probability distribution in the utterance choice. Finally, the sampled utterance is returned.

Algorithm 13 Pragmatic Speaker

procedure $pragmatic\_speaker$(state, arousal, goal)
  $alpha \leftarrow 1.0$ ▷ Scaling factor
  $goal\_value \leftarrow goal\_state(goal, state, arousal)$
  $sampled\_utterance \leftarrow utterance\_prior()$ ▷ Sample utterance
  $l\_0 \leftarrow literal\_listener(sampled\_utterance, goal)$ ▷ Reasoning on listener
  $sample(alpha * l\_0, obs = goal\_value)$ ▷ Incorporate listener interpretation
  return $sampled\_utterance$
end procedure

Pragmatic listener

The pragmatic listener calculates the probability distribution shown in Equation 8.13.

First, it computes the prior distributions over states $P_S(s)$, over arousal with facial expression $\tilde{P}_A(a \mid s, AU)$ and over goals $P_G(g)$. On the next step, it performs the social recursive reasoning through the pragmatic speaker. The obtained speaker's distribution is conditioned by observing the utterance.

Algorithm 14 Pragmatic Listener

procedure $pragmatic\_listener$(utterance, arousal)
  $sampled\_state \leftarrow state\_prior()$ ▷ Sample state
  $sampled\_arousal \leftarrow arousal\_sample(sampled\_state, arousal)$ ▷ Sample arousal
  $sampled\_goal \leftarrow goal\_prior()$ ▷ Sample goal
  $s\_1 \leftarrow pragmatic\_speaker(sampled\_state, sampled\_arousal, sampled\_goal)$
  $sample(s\_1, obs = utterance)$ ▷ Condition speaker's distr. with utterance
  return $sampled\_state, sampled\_arousal, sampled\_goal$
end procedure

8.5.4 Simulations

In this study, we did not utilise data directly from actual patient-clinician interactions due to the inability to collect such data and the lack of suitable existing datasets. This absence of real data necessitated the adoption of alternative methods to analyse patient-clinician interactions. Consequently, we created hypothetical data based on controlled linguistic models, clinical scenarios, or theoretical assumptions. This synthetic data aims to emulate the dynamics of communication between clinicians and patients, incorporating pragmatic features and capturing various facets of healthcare communication. The hypothetical data primarily consists of prior probability distributions (over states, utterances, and goals) that frame the interaction between doctors and patients, emphasising the importance of focusing on specific clinical scenarios rather than relying on overly simplistic general models.

In the absence of empirical data, hypothetical case studies and scenarios were constructed. Despite the inherent limitations, meticulous considerations were given to how broadly the findings from simulated and hypothetical scenarios could be applied to actual patient-clinician interactions. We now present results from simulations with the following parameters:

- Prior distribution over utterances $P_U(u)$ is uniform;
- Prior distribution over goals $P_G(g)$ is uniform;
- The probabilities of having an high arousal are:
  $\pi^A(mild) = 0.3$, $\pi^A(moderate) = 0.7$, $\pi^A(severe) = 0.9$ (see Eq. 8.4);
- We suppose the probability of high arousal related to facial expressions are:
  $\pi^A(relaxed) = 0.01$, $\pi^A(intense) = 0.99$ (see Eq. 8.17);

The results were derived by altering the distributions over states, specifically by varying the context where medical records contribute to a shared understanding of pain states between the doctor and the patient. Essentially, the scenario involves an objective

clinician who relies solely on medical records or the results of medical examinations to assess patients' reported pain.

For each context, we detail scenarios where the patient accurately reports their pain, as well as cases where patients either underreport or overreport their pain.

Mild pain

We initiate our simulation in a context where clinical data suggests a high likelihood of mild pain for the patient. The prior distribution over pain states $P_S(s)$ is defined as $[0.7, 0.2, 0.1]$ corresponding to the states $[\text{mild}, \text{moderate}, \text{severe}]$.

If the patient reports experiencing mild pain, this aligns with the clinical indications, reinforcing the physician's confidence that the patient is genuinely experiencing mild pain as depicted in Figures 8.18 and 8.19.

Both figures predominantly suggest mild pain, though they differ in the patient's level of arousal, influenced mainly by facial expressions. In Figure 8.18, the patient appears relaxed, while in Figure 8.19, a more intense expression is evident, possibly reflecting a heightened pain sensitivity.

Conversely, Figures 8.20, 8.21, and 8.22 illustrate scenarios where the patient reports more severe pain than indicated by clinical records. Given the reliance on these records, the clinician might infer dishonesty, suggested by the continued high likelihood of mild pain, indicative of potential deceit.

The model inherently incorporates the possibility of cheating, not to ascribe malevolent intent but to capture the involuntary aspects of a patient misreporting pain. Such discrepancies might not stem from a deliberate desire to deceive but rather from the patient's affective state regarding pain, influenced here by their arousal state.

Observing Figures 8.20, 8.21, and 8.22, it is clear that the clinician may question the patient's actual pain state due to similar probabilities across mild and moderate pain states in the first two graphs, and mild to severe in the last.

In Figure 8.23, the assignment of a high probability to the mild state under intense arousal might suggest to the clinician that the patient has over-reported their pain, possibly out of a desire to hasten medical attention or ensure treatment.

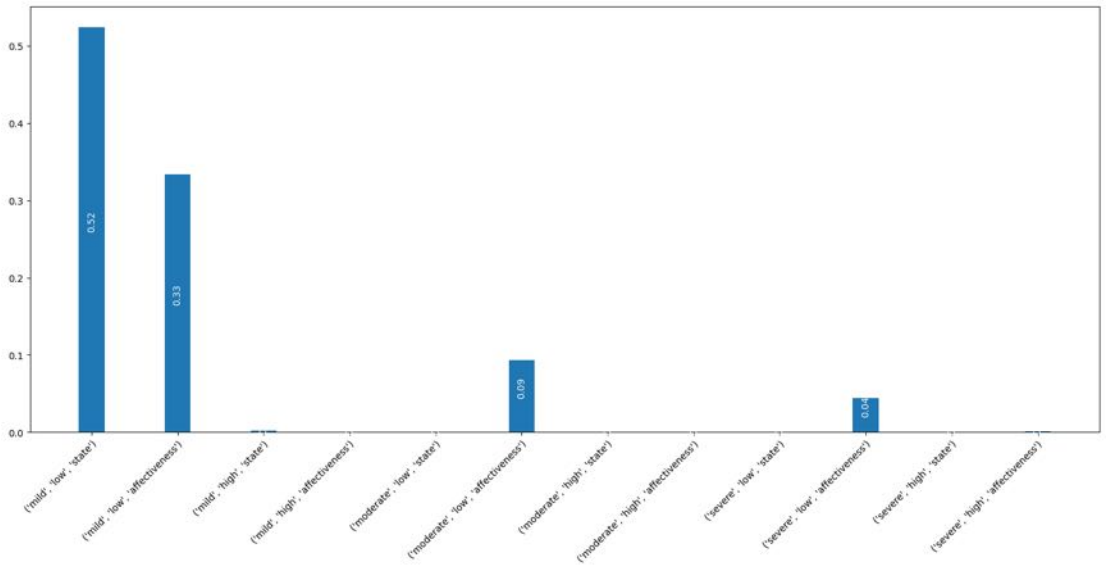


Figure 8.18: *Posterior probabilities across pain states in a mild pain context, where the patient reports mild pain with a relaxed facial expression. The alignment between the clinical context and the patient's report suggests a high confidence in the mild pain state.*

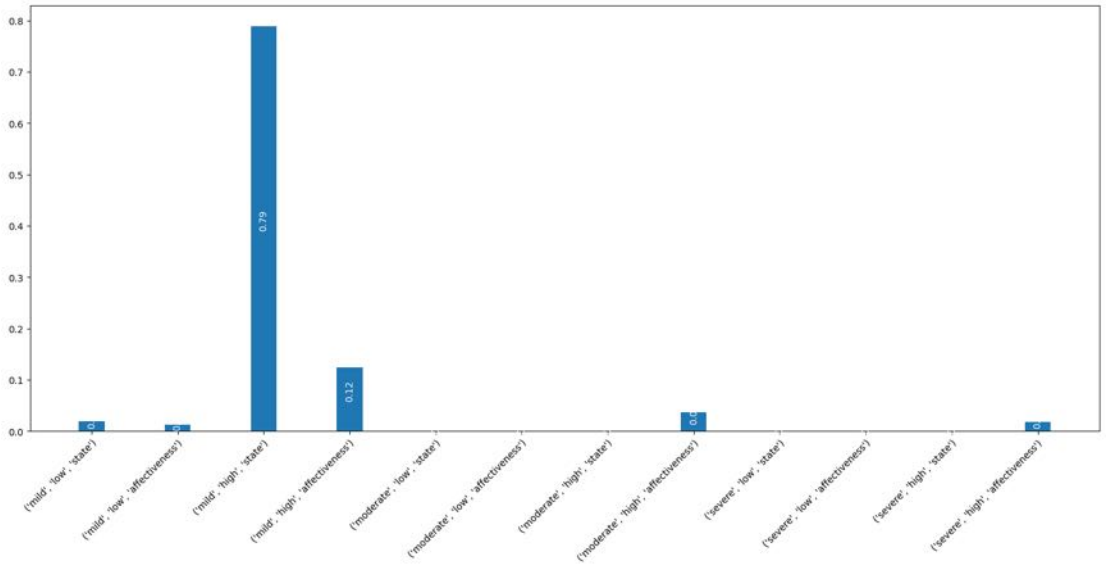


Figure 8.19: *Posterior probabilities in a mild pain context, with the patient reporting mild pain but displaying an intense facial expression. The contrast between the relaxed expression and the reported pain level indicates a potential heightened pain sensitivity or emotional response.*

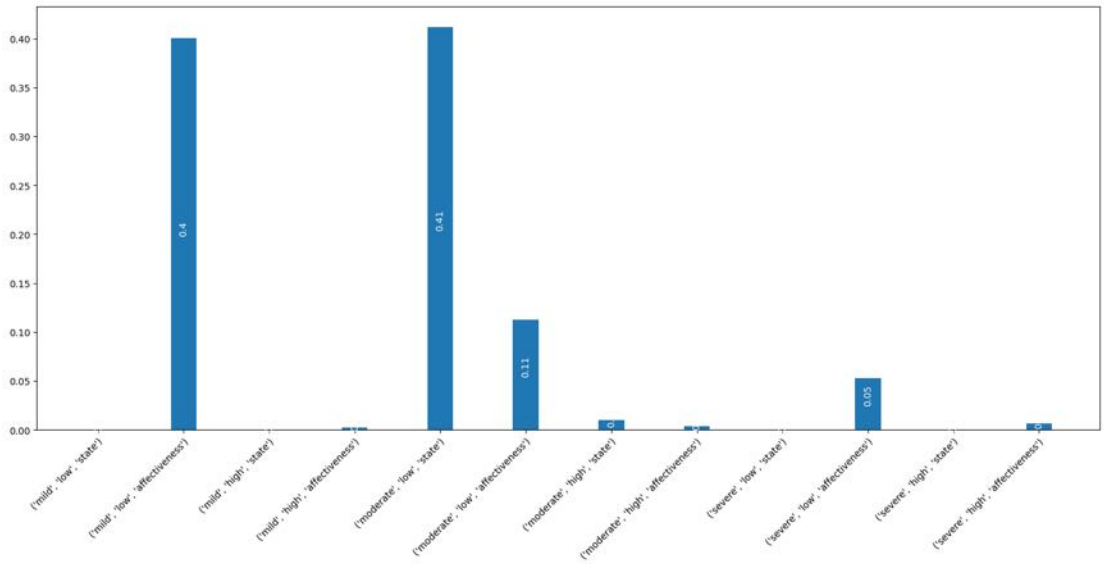


Figure 8.20: *Posterior probabilities when a patient in a mild pain context reports moderate pain with a relaxed expression. Despite the patient reporting a higher level of pain, the relaxed expression may lead the clinician to question the accuracy of the report, with a continued emphasis on the mild pain state.*

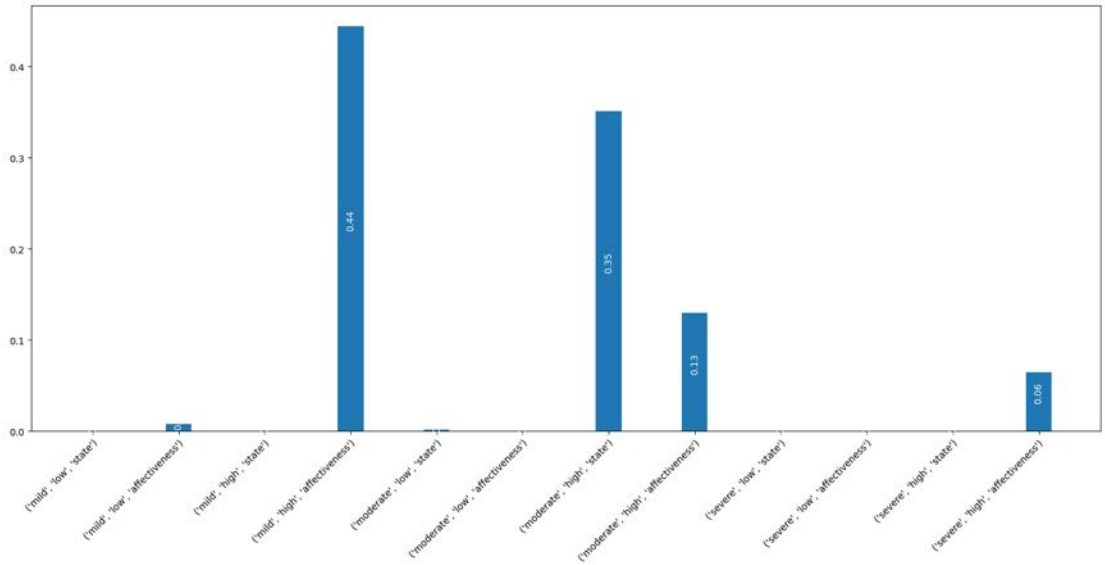


Figure 8.21: *Posterior probabilities for a patient reporting moderate pain in a mild pain context, with an intense facial expression. The combination of a more intense expression with a higher reported pain level could suggest a discrepancy between the clinical indications and the patient's experience and, as a consequence, more uncertainty for the clinician.*

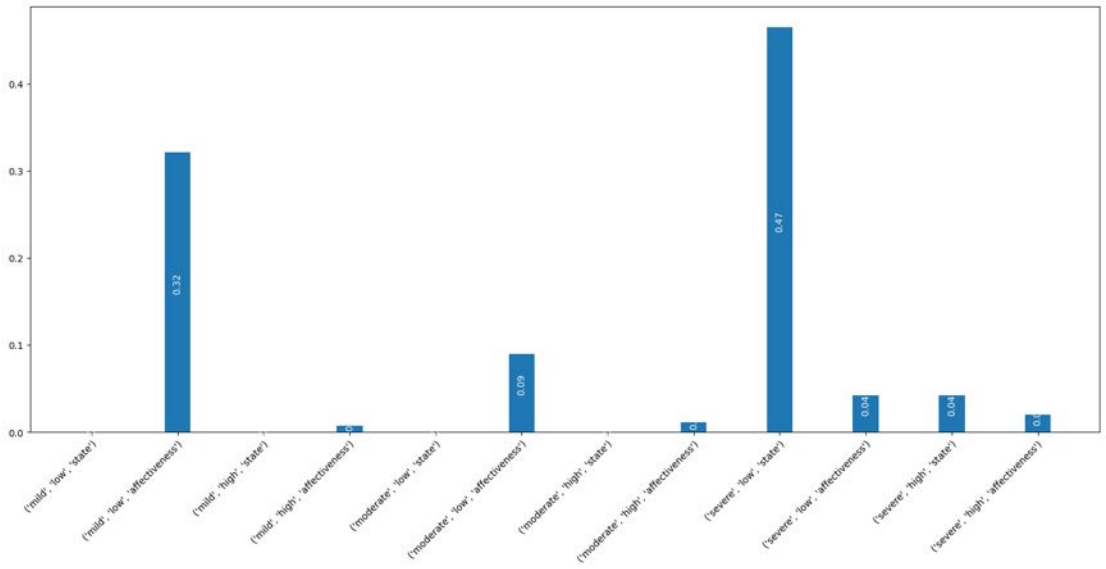


Figure 8.22: *Posterior probabilities when a patient reports severe pain in a mild pain context, yet displays a relaxed expression. The significant mismatch between the reported pain and the clinical context might raise concerns about the accuracy of the patient's report, raising uncertainty.*

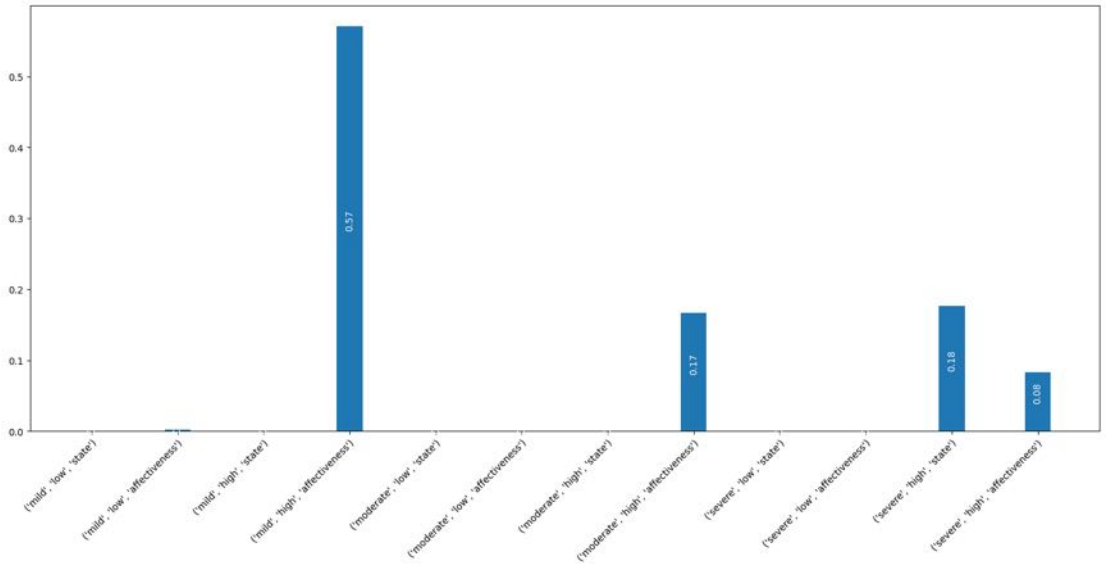


Figure 8.23: *Posterior probabilities for a patient reporting severe pain with an intense expression in a mild pain context. The clinician may interpret this as an overreporting of pain, possibly due to anxiety or a desire to prompt quicker medical attention, thus "mild" remains the most plausible state for the clinician.*

Moderate pain

In this section, we explore a scenario where the clinician anticipates moderate pain, characterised by the prior probability distribution over states $[0.1, 0.7, 0.2]$. The straightforward instances depicted in Figures 8.26 and 8.27 display scenarios where the patient's reported pain aligns precisely with the clinician's expectations of moderate pain.

Considering a relaxed expression while reporting mild pain, as seen in Figure 8.24, the clinician, adhering to medical records, might still attribute moderate pain to the patient. This attribution could stem from the patient's relaxed demeanour, which might not be typical but can occur, for instance, during certain medical or physiotherapy treatments where the patient experiences less pain than the underlying condition might suggest.

Conversely, when the patient reports mild pain with an intense expression, as shown in Figure 8.25, the probabilities of mild and moderate pain states increase. This scenario could indicate a fear of pain or the situation itself if the pain is actually moderate, or it could reflect a higher pain tolerance if the pain is genuinely mild.

On the other hand, Figures 8.28 and 8.29 illustrate scenarios where the patient reports more severe pain than anticipated. These cases could arise from the patient's anxiety or agitation, potentially leading them to experience and report greater pain than what might be clinically expected, irrespective of their actual pain tolerance.

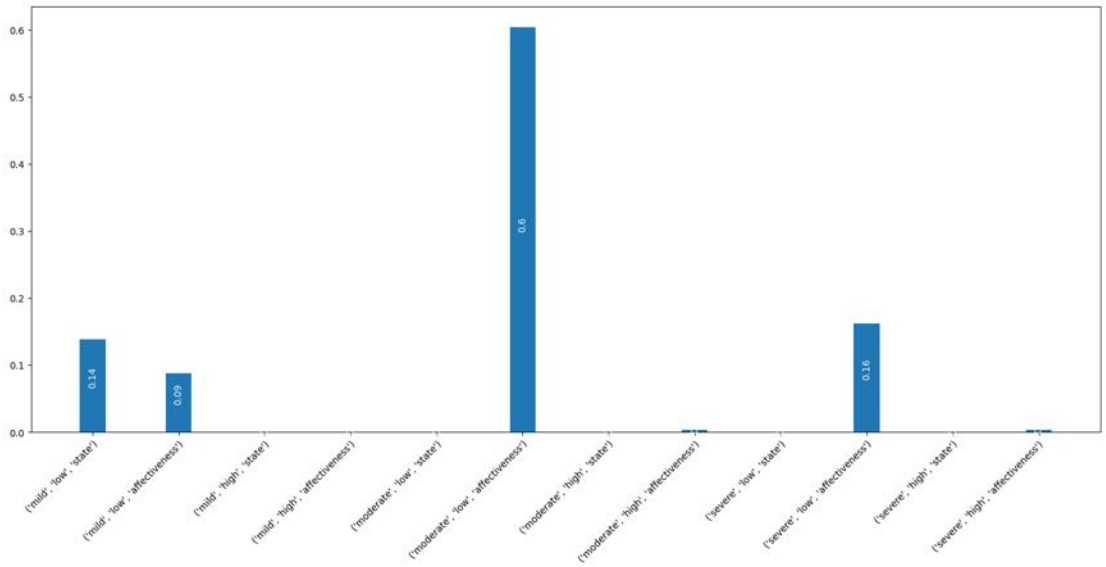


Figure 8.24: *Posterior probabilities across pain states in a moderate pain context, with the patient reporting mild pain and displaying a relaxed expression. The clinician might still attribute moderate pain despite the lower reported pain level, based on the clinical context.*

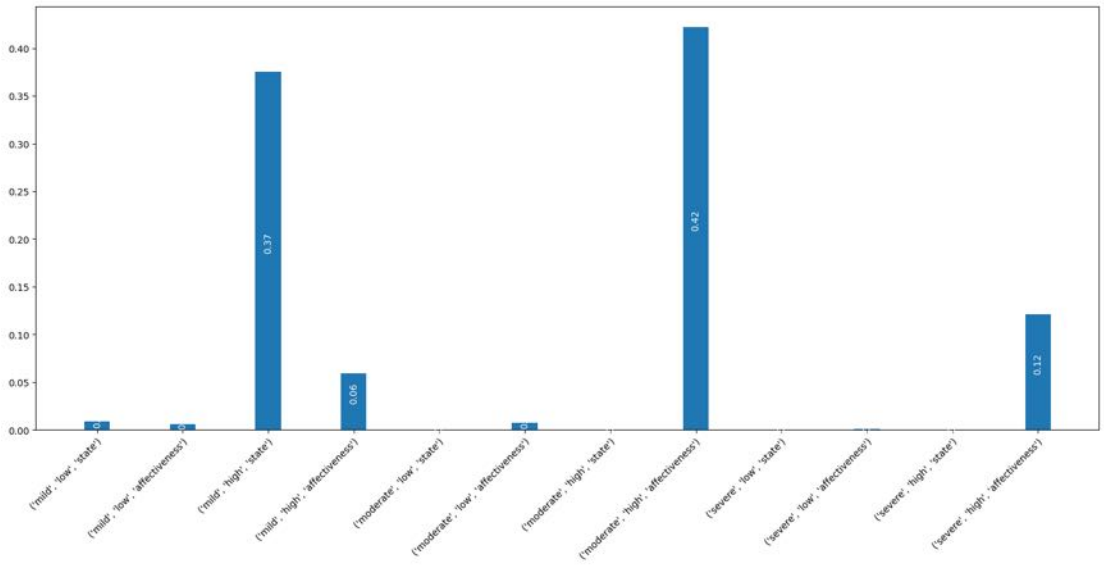


Figure 8.25: *Posterior probabilities in a moderate pain context, where the patient reports mild pain but has an intense facial expression. Here we see an increased likelihood of both mild and moderate pain states.*

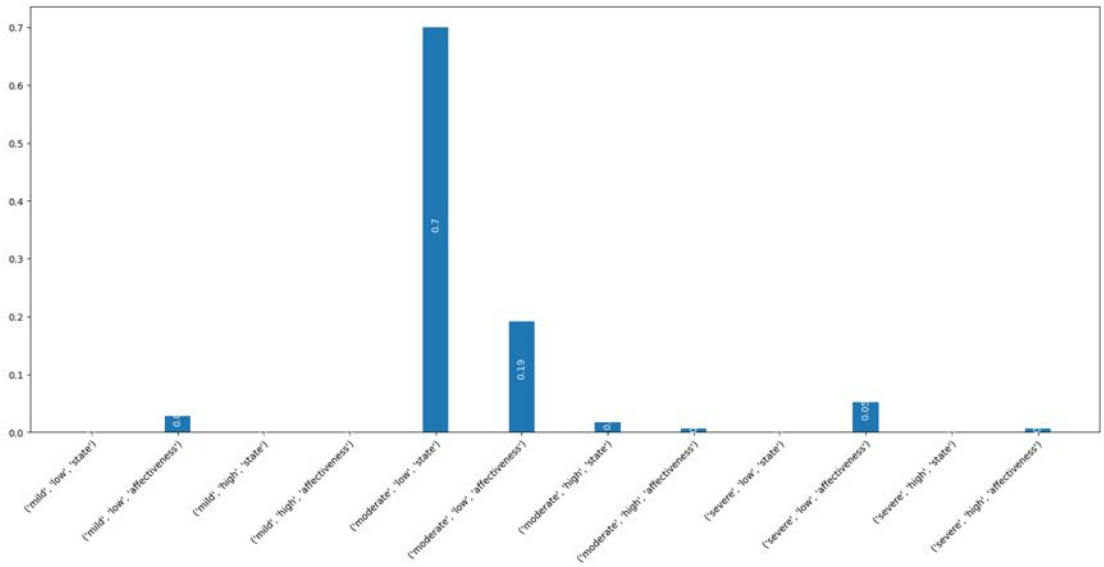


Figure 8.26: *Posterior probabilities for a patient reporting moderate pain in a moderate pain context, with a relaxed expression. The alignment between reported and expected pain levels suggests accurate self-reporting by the patient.*

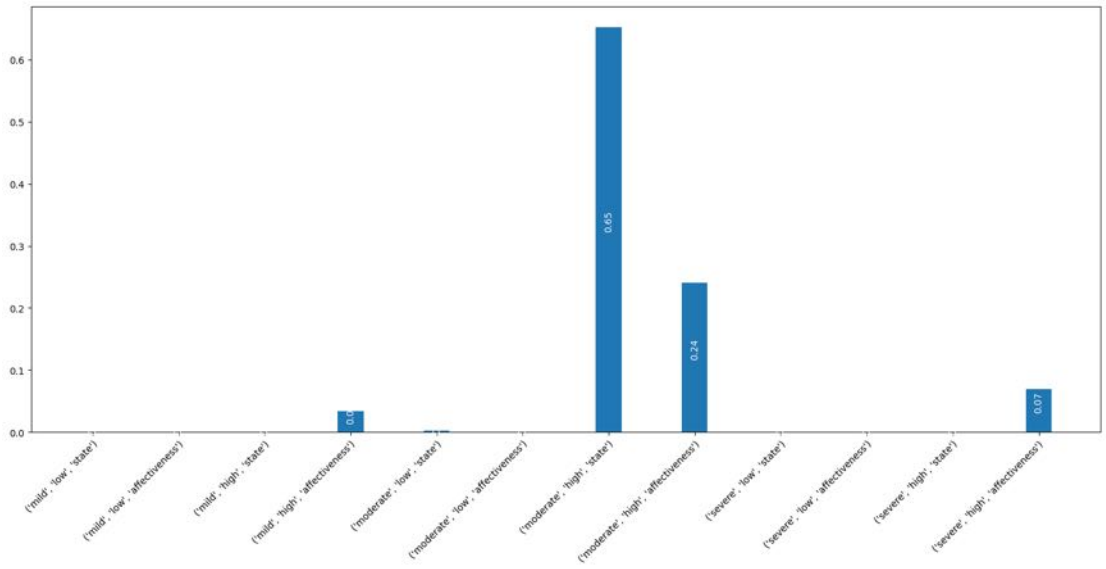


Figure 8.27: *Posterior probabilities when a patient reports moderate pain with an intense expression in a moderate pain context. The clinician might consider this as a possible indication of higher pain sensitivity or anxiety.*

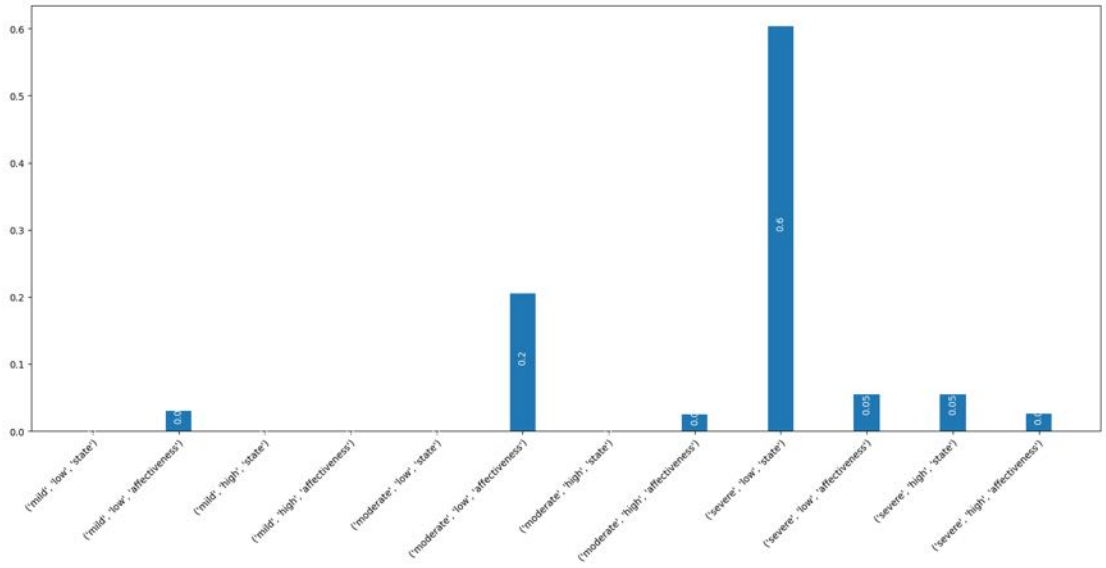


Figure 8.28: *Posterior probabilities across pain states in a moderate pain context, with the patient reporting severe pain but showing a relaxed expression. This combination may lead the clinician to slightly question the reported pain level, even if severe pain is considered the most plausible state.*

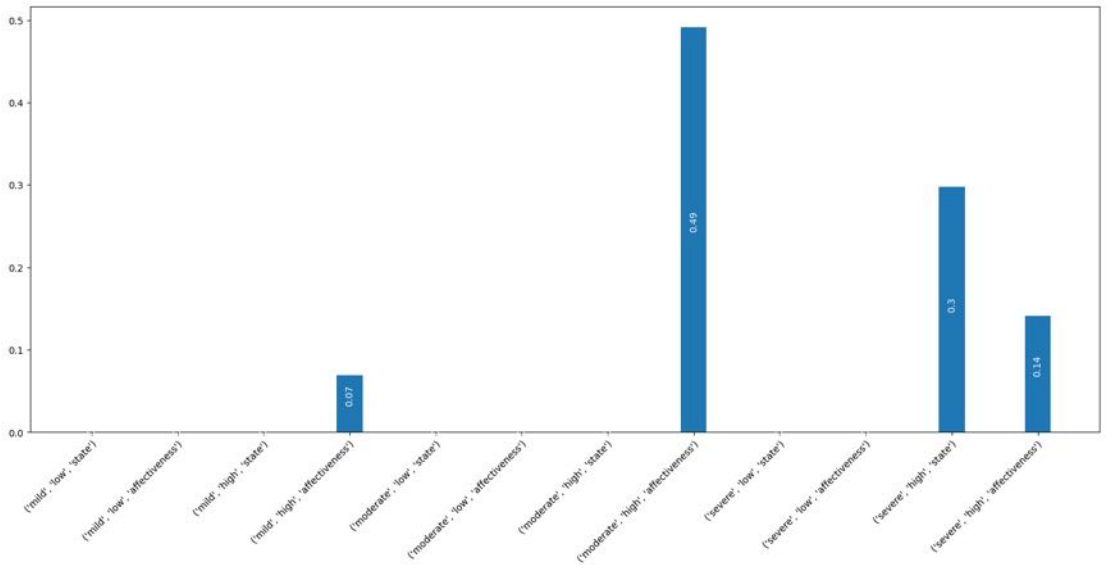


Figure 8.29: *Posterior probabilities for a patient reporting severe pain in a moderate pain context, combined with an intense expression. This could be interpreted as the patient experiencing severe pain or possibly exaggerating due to distress.*

Severe pain

In scenarios where clinicians anticipate severe pain based on medical records, the prior distribution over pain states is set at $[0.1, 0.2, 0.7]$.

When a patient reports severe pain, as expected, the posterior distributions are illustrated in Figures 8.34 and 8.35. These figures reflect a cooperative patient who accurately communicates their pain level. The primary distinction between the two graphs lies in the level of arousal; the second graph 8.35 additionally considers the patient's affective state through the *affect* goal, emphasising an emotional dimension to the pain reporting.

Figures 8.30 and 8.31 depict scenarios where the patient reports mild pain. If the patient appears relaxed, as shown in Figure 8.30, there remains a high probability that the pain is actually severe, possibly under-reported due to the patient's affective state influenced by treatment. This mirrors the less plausible scenario in Figure 8.24 from the previous section, where the clinician is unlikely to believe mild pain reports in the face of severe pain indications.

In Figure 8.31, the patient may under-report pain not out of confidence but due to fear, evidenced by high arousal levels. This dynamic also applies to Figures 8.32 and 8.33, where the patient reports moderate pain. In the relaxed scenario depicted in Figure 8.32, the patient might genuinely experience moderate pain due to personal pain tolerance or a calm affective state.

Conversely, in Figure 8.33, the intense expression suggests the clinician might interpret this as the patient experiencing actual moderate pain due to tolerance, or possibly, the patient is downplaying severe pain out of fear or anxiety. This analysis underscores how affective states and expressions significantly influence clinical interpretations of reported pain levels.

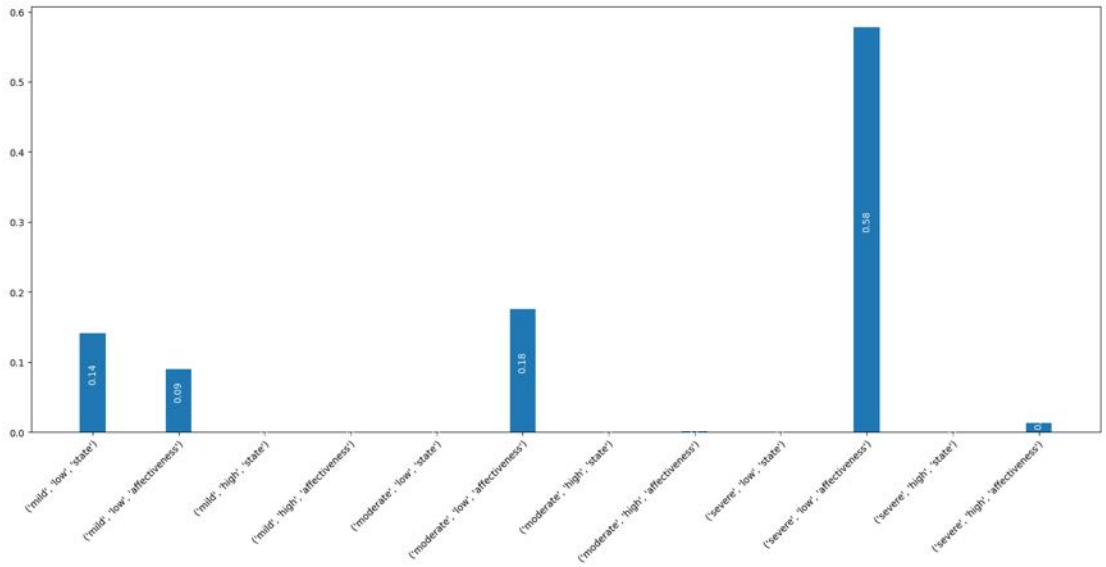


Figure 8.30: *Posterior probabilities across pain states in a severe pain context, with the patient reporting mild pain and appearing relaxed. The clinician may suspect that the patient is underreporting their pain, potentially due to their affective state.*

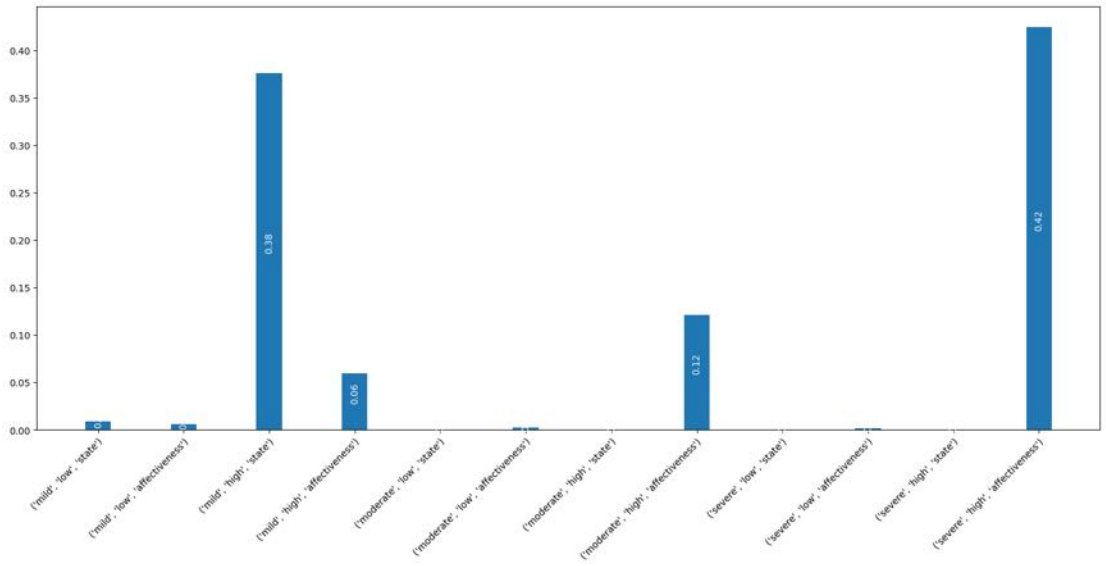


Figure 8.31: *Posterior probabilities in a severe pain context, with the patient reporting mild pain but displaying an intense expression. This combination might suggest the patient is downplaying their pain due to fear or anxiety. Nonetheless, the clinician remains uncertain about the actual level of pain experienced by the patient.*

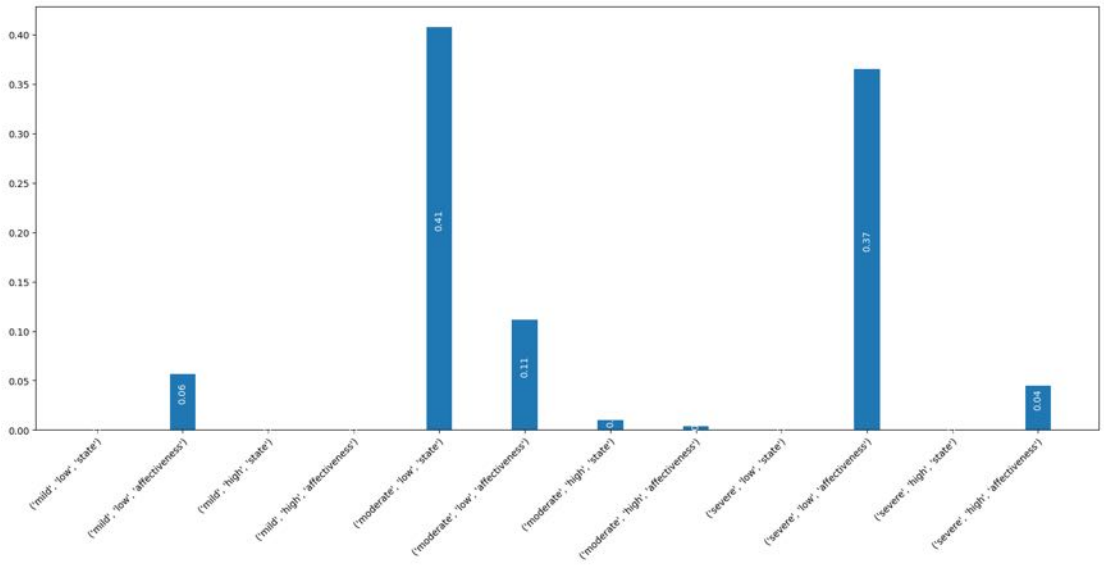


Figure 8.32: *Posterior probabilities for a patient reporting moderate pain in a severe pain context, with a relaxed expression. The relaxed demeanor may lead the clinician to question whether the patient is accurately reporting their pain level, remaining with high uncertainty.*

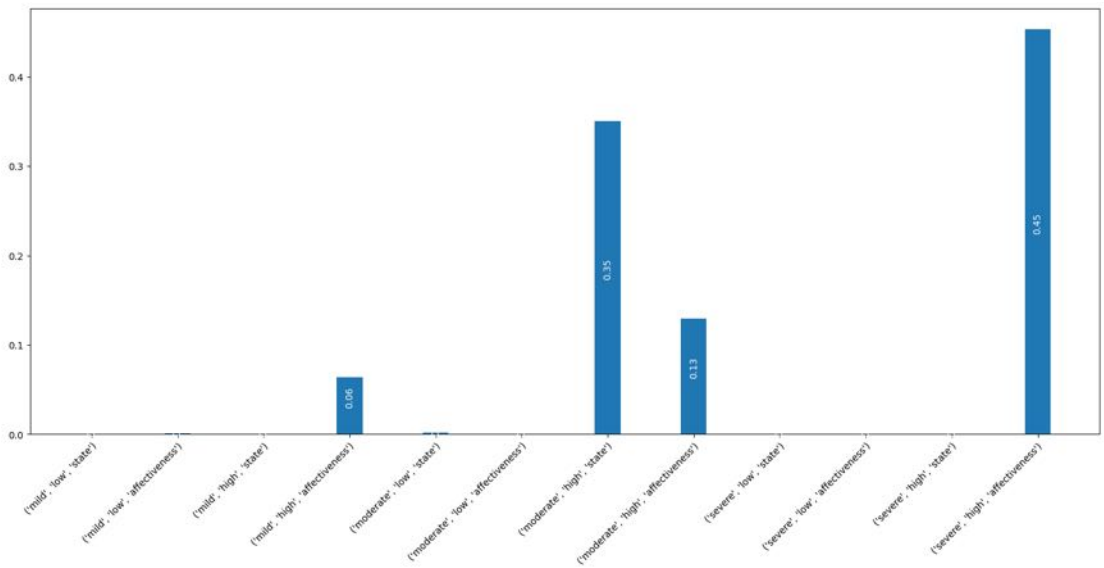


Figure 8.33: *Posterior probabilities in a severe pain context, where the patient reports moderate pain but has an intense expression. This might indicate that the patient is experiencing severe pain but is downplaying it, potentially due to fear.*

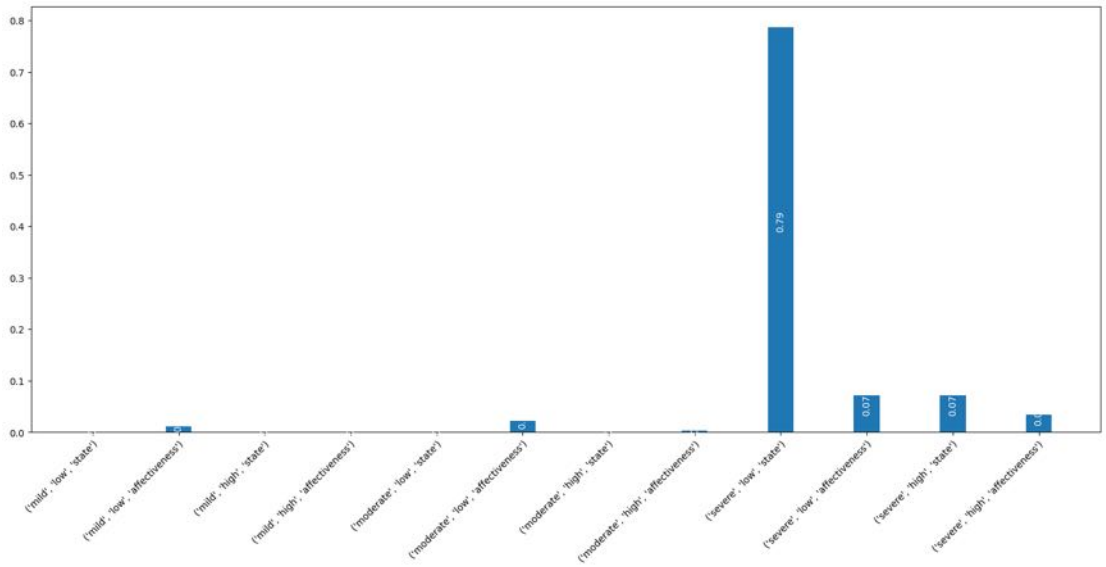


Figure 8.34: *Posterior probabilities across pain states in a severe pain context, with the patient reporting severe pain and displaying a relaxed expression. This discrepancy could suggest the patient has a high tolerance for pain, leading to minimal uncertainty in the clinician's overall assessment.*

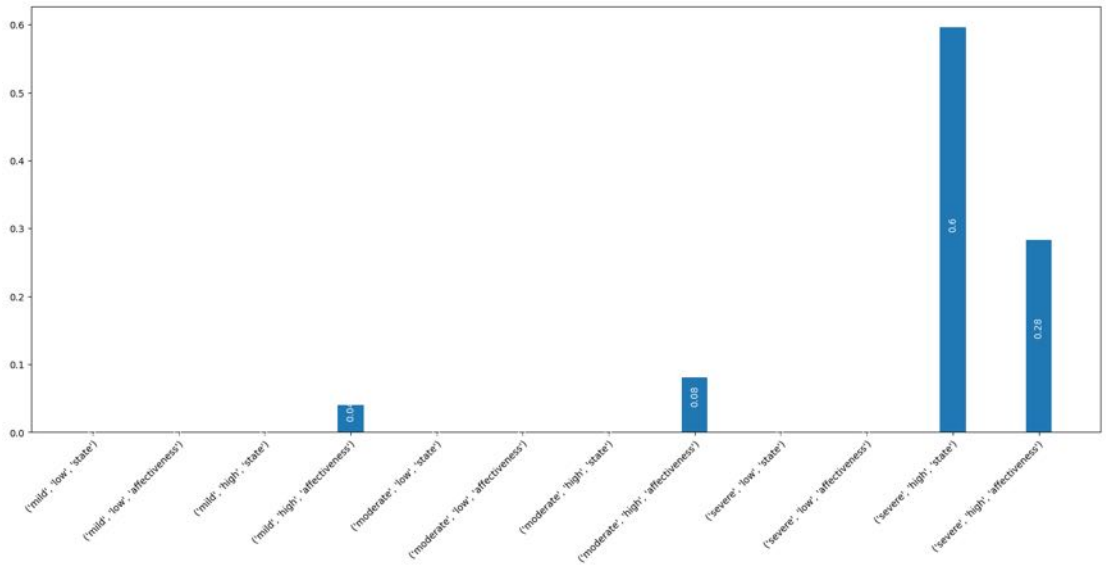


Figure 8.35: *Posterior probabilities for a patient reporting severe pain with an intense expression in a severe pain context. The intense expression may strengthen the clinician's belief in the accuracy of the reported severe pain.*

8.6 Discussion

The chapter proposes pain models rooted in the biopsychosocial approach, emphasising the importance of considering pain not only as a biological phenomenon but also as a psychological and social experience influenced by social determinants, cultural factors, and interpersonal relationships. Both presented models acknowledge that pain develops and manifests within social contexts and that communication between the sufferer and the observer is crucial for a comprehensive understanding of the pain experience.

As proposed by our taxonomy, class 3 models specifically include social aspects, modelling the interaction between the sufferer and the observer, such as a clinician or an individual capable of providing assistance. These models emphasise the social and communicative aspects of pain, portraying it as a social experience that emerges and evolves through interactions between the sufferer and the observer. Two main approaches are discussed: one based on Partially Observable Markov Decision Processes (POMDP) and active inference, and the other based on the Rational Speech Act (RSA) framework.

The POMDP and active inference approach involves a dynamic interaction between the patient and the clinician, allowing for complex inferences about pain states that incorporate both observed actions and potential deceptive behaviours. The clinician uses observations from the patient's verbal and non-verbal cues, along with a "cheating cue" to assess the likelihood of deception. This approach highlights the clinician's inference process about the patient's pain level and decision making on appropriate treatments, and, foremost, the role of the patient as a communicative agent in search of the best communicative strategy to achieve pain cessation.

The RSA-based approach focuses on the pragmatic reasoning behind pain communication, portraying both the listener (clinician) and the speaker (patient) as engaged in mutual, recursive inference about each other's intentions. This approach explicitly highlights the importance of communication, considering both literal and affective dimensions of pain expression. The clinician interprets the patient's verbal and non-verbal cues within the context of their shared understanding, influenced by cultural and social norms.

Both modelling approaches recognise that communication does not always have an exclusively informative value but can also include the intention to generate the desired action in the interacting agent, thereby influencing their behaviour to favour the communicative agent's goal. The receiving agent, the clinician in our modelling, is aware of this possible communicative distortion and thus searches for signals that can ensure the truthfulness of the communication or indicate deception. This is the role of the cheating cue in our models, which is modelled to represent the agents' awareness of this possibility.

However, this ability poses several challenges because it can easily turn into a disadvantage when a natural inclination towards non-naivety transforms into an unjustified bias that could worsen clinical practice and make clinicians less inclined to adequately treat certain categories. Scientific evidence demonstrates this in the case of women, where gender bias in pain assessment is a well-documented issue (Zhang et al., 2021).

In general, a model that aims to describe and explain pain at a high level cannot ignore social factors and therefore requires modelling based on an active conceptuali-

sation that is shared to some extent among the involved agents. These aspects can be considered in Class 3 modelling. Understanding and addressing biases, such as those related to gender, is crucial for accurate pain assessment and effective treatment, highlighting the importance of a comprehensive, biopsychosocial approach to pain management.

CHAPTER 9

Conclusions

In this thesis, we have addressed the tangled topic of pain modelling from a Bayesian perspective. As extensively discussed in the initial chapters, pain is a multifaceted phenomenon that requires an integrated approach, incorporating insights from various disciplines such as neurobiology, psychology, philosophy, sociology, and computer science to analyse computational approaches to the subject. The study and analysis of these elements were crucial for constructing a taxonomy that organises existing works in the state of the art according to principled criteria. This stratification aims to do justice to the complexity of conceptualising the phenomenon of pain.

We observed how Class 0 models allow for a description of pain as a static and pointwise inference of the cause of incoming nociceptive information. With Class 1 models, this notion is extended to consider temporal dynamics, essential for a subject attempting to estimate their condition in a changing environment. We proposed two distinct models within this category: a discriminative approximation based on Graph Neural Networks (GNN) and a generative one eventually leading to a sort of Input-Output Hidden Markov Model implementation model. Both have been validated through experimental datasets available in the literature.

Adding another layer of analysis, we considered the motivational function of pain, introducing Class 2 models, where the experience of pain is examined within the action-perception loop, taking into account an agent's interactions with their environment. The two approaches proposed in this thesis—one based on simulations and active inference, and the other data-based—were developed from experimental data in the context of fear generalisation. This approach allowed us to integrate the discussion of pain with that of affect, which plays a crucial role in the pain experience.

Finally, we explored the highest level of the hierarchy, the social level, through models that simulate the interaction between two agents—a sufferer and an observer—within

two different frameworks: active inference and the Rational Speech Act (RSA) framework. Incorporating the social level enabled the exploration of high-level components, particularly voluntary (verbal and facial expressions) and involuntary (facial micro-expressions) communication of pain. At this level, we encounter specific issues, typically sociological, such as interactions that extend beyond the informational level (communication as transaction) and resulting biases and prejudices commonly encountered in clinical pain management practice. For example, gender biases in pain assessment, as highlighted in various studies, demonstrate how clinicians may misinterpret pain expressions based on preconceived notions.

In essence, the framework we have provided aims to systematise the phenomenon from a biopsychosocial perspective and proposes models capable of representing the various facets of pain with progressively higher levels of abstraction and/or completeness. These models suggest principles that can be used by future computational models to contextualise and adapt their characteristics to the application context: not all contexts will require social modelling (e.g., pain in neonates), while in certain contexts it could be fundamental (e.g., clinical pain management that commonly involves a "transaction" between doctor and patient).

In short, the main contributions of this thesis include:

- Model pluralism: given the complexity of pain, implementing a single model that accounts for all facets is impractical. Although this thesis outlines a high-level comprehensive model, it proposes a pluralistic approach, allowing for the implementation of particular levels of the hierarchy and specific factors of the heterarchy based on context. Comparing different approaches for the same context, we also show how a model-based approach, which is more interpretable and computationally lighter, can achieve comparable performance to data-hungry and heavier models (e.g. GNNs), depending on the context and data.

- Active Inference modelling of pain: this work advances pain modelling through active inference across multiple levels: affective-motivational (in the context of fear generalisation) and social.

- Process/model distinction: this work highlights and addresses the distinction between the process related to the phenomenon itself and the modelling done by the experiencer or the observer, thereby clarifying the different layers of analysis required.

- Model emulator: in addition to the concept of a generative model of the process held by the observer, we introduce another level of simplification of a model emulator in the observer, providing another level of abstraction.

- Dyadic interaction modelling: by modelling dyadic interactions in the context of pain, the thesis brings to light the social aspects of pain overlooked in computational literature, particularly simulating clinical settings.

- RSA framework for pain modelling: incorporating the Rational Speech Act framework, we explore the aspects of communication and social interaction in pain expression and assessment.

To the best of our knowledge, all these contributions are novel to the field of computational modelling of pain.

This comprehensive approach underscores the need to understand pain as a dynamic, transactional process influenced by a multitude of factors. By integrating multidisciplinary insights and addressing various levels of pain experience, from individual perception to social interaction, this thesis contributes to a more nuanced and holistic understanding of pain, providing a robust framework for future research and practical applications in pain management and assessment. The proposed models and frameworks are adaptable and can be tailored to specific contexts, thus offering a versatile toolkit for researchers in the field of pain study and management.

It is important to acknowledge that the theoretical model presented in this thesis serves as a blueprint rather than a fully implemented system. Due to the intrinsic complexity of the topic and the current lack of comprehensive datasets, the theoretical model proposed has not been implemented in its entirety. Indeed, a limitation we have had to address in this work is the current lack of suitable data that comprehensively encompass the various hierarchical levels and the complexity of the pain phenomenon. Although pain data sets do exist, they often do not cover the full spectrum required for an integrated model. This limitation is not only costly to overcome, but in some cases is also impractical. Therefore, it is essential to develop models that can learn effectively from small sample datasets. As demonstrated in this work, another viable approach is to rely on simulations and compare their results with the behavioural evidence found in the clinical and psychological literature. Efforts should be invested in creating machine learning techniques capable of learning and generalising from modest-sized datasets while using simulation-based approaches to validate and enhance these models against established behavioural data.

Overall, we hope that the initial step proposed in this dissertation will pave the way for novel insights and discoveries in the complex and multifaceted domain of pain research.

APPENDIX $\mathcal{A}$

Probabilistic Graphical Models

PROBABILISTIC graphical models (PGM) allow to enormously simplify complex joint distributions using conditional independence property in order to achieve factorisations 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 factorisation 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 hold in P, but not the *vice versa*, that is P could have independencies not reflected in $\mathcal{G}$.

These key concepts allow the compact factorised 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 factorises 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 factorise 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 factorises according to $\mathcal{G}$.*

Theorem A.1 proves the factorisation of P according to $\mathcal{G}$, but also the converse holds: factorisation 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 factorises 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 $X_1 \leftrightharpoons \cdots \leftrightharpoons X_n$ a trail in $\mathcal{G}$. Let $\mathbf{Z}$ be a subset of *observed variables*. The trail $X_1 \leftrightharpoons \cdots \leftrightharpoons X_n$ 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 factorises 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 factorise over $\mathcal{G}$, that is, for all distributions except for a set of measure zero in the space of CPD parameterizations, 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 factorise over $\mathcal{G}$.

These results state that for almost all parameterizations 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 modeling 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 reparameterize 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 $T > 0$).

A.1 Forney-style Factor Graph

A Forney-style factor graph (FFG) (**?**) offers a graphical representation of a factorised probabilistic model. In an FFG, edges represent variables and nodes specify relations between variables. As a simple example, consider a generative model (joint probability distribution) over variables $X_1, \ldots, X_5$ that factors as

$$f(X_1, \ldots, X_5) = f_a(X_1)\, f_b(X_1, X_2)\, f_c(X_2, X_3, X_4)\, f_d(X_4, X_5)\,, \tag{A.3}$$

where $f_\bullet(\cdot)$ denotes a probability density function. This factorised model can be represented graphically as an FFG, as shown in Fig. A.1. Note that although an FFG is

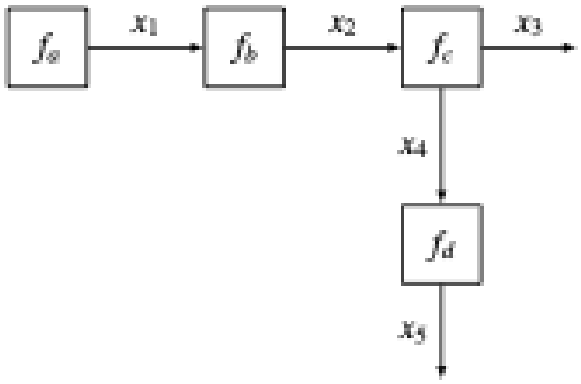


Figure A.1: *Forney-style factor graph (FFG) representation of Eq. A.3. In an FFG, edges correspond to variables and nodes represent factors that encode constraints among variables. A node connects to all edges that correspond to variables that occur in its factor function. For example, node f_b connects to edges X_1 and X_2 since those variables occur in $f_b(X_1, X_2)$. Variables that occur in just one factor (X_3 and X_5 in this case) are represented by half-edges. While an FFG is principally an undirected graph, we usually specify a direction for the (half-)edges to indicate the generative direction of the model and to anchor the direction of messages flowing on the graph.*

principally an undirected graph, in the case of generative models we specify a direction for the edges to indicate the "generative direction". The edge direction simply anchors the direction of messages flowing on the graph (we speak of forward and backward messages that flow with or against the edge direction, respectively). In other words, the edge directionality is purely a notational issue and has no computational consequences.

The FFG representation of a probabilistic model helps to automate probabilistic inference tasks. As an example, consider we observe $X_5 = \hat{X}_5$ and are interested in calculating the marginal posterior probability distribution of X_2 given this observation.

In the FFG context, observing the realization of a variable leads to the introduction of an extra factor in the model which "clamps" the variable to its observed value. In our example where X_5 is observed at value $\hat{X}_5$, we extend the generative model to $f(X_1, \ldots, X_5) \cdot \delta(X_5 - \hat{x}_5)$. Following the notation introduced in **?**, we denote such "clamping" factors in the FFG by solid black nodes. The FFG of the extended model is illustrated in Fig. A.2.

Computing the marginal posterior distribution of X_2 under the observation $X_5 = \hat{X}_5$ involves integrating the extended model over all variables except X_2, and renormalizing:

$$f(X_2 \mid X_5 = \hat{X}_5) \propto \int \cdots \int f(X_1, \ldots, X_5) \cdot \delta(X_5 - \hat{X}_5) dX_1 dX_3 dX_4 dX_5 \tag{A.4}$$

$$= \overbrace{\int \underbrace{f_a(X_1)}_{1} \, f_b(X_1, x_2) dX_1}^{2} \overbrace{\iint f_c(X_2, X_3, X_4) \underbrace{\left(\int f_d(X_4, X_5) \cdot \delta(X_5 - \hat{X}_5) dX_5 \right)}_{3} dX_3 dX_4}^{4} . \tag{A.5}$$

The nested integrals in Eq. A.5 result from substituting the factorisation of Eq. A.3 and rearranging the integrals according to the distributive law. Rearranging large integrals of this type as a product of nested sub-integrals can be automated by exploiting the FFG representation of the corresponding model. The sub-integrals indicated by circled numbers correspond to integrals over parts of the model (indicated by dashed boxes in Fig. A.2), and their solutions can be interpreted as messages flowing on the FFG. Therefore, this procedure is known as *message passing* (or summary propagation). The messages are ordered ("scheduled") in such a way that there are only backward dependencies, i.e., each message can be calculated from preceding messages in the schedule. Crucially, these schedules can be generated automatically, for example by performing a depth-first search on the FFG.

Message passing is generally efficient because the computation of every message is node-local in the FFG. More specifically, the message flowing out of a factor node f_a can be calculated from the analytic form of factor f_a and all messages inbound to node f_a. If the analytic forms of the incoming messages are known (which is often the case), a pre-derived *message computation rule* can be used to compute the outgoing message. These rules can be stored in a lookup table for reuse in any model that involves that specific factor-message combination. This important locality property thus enables efficient and automated probabilistic inference.

In the case of marginalization, the messages are derived according to the so-called *sum-product rule*[3], which leads to the sum-product (belief propagation) algorithm. As an example derivation we consider the outgoing message of an "equality constraint node" (see Fig. A.3, left; see also **?**), which constrains three variables x, y, z to equal values through the factor $f_=(X, Y, Z) = \delta(Z - X)\,\delta(Z - Y)$.

For given incoming messages $\mu_1(X)$ and $\mu_2(Y)$ on edges X and Y (depicted by 1 and 2 in the Fig. A.3, right), the outgoing sum-product message on the z-edge is given

[3]The name sum-product rule derives from the observation that each sub-integral in Eq. A.5 comprises a sum (integral) of a product of factors.

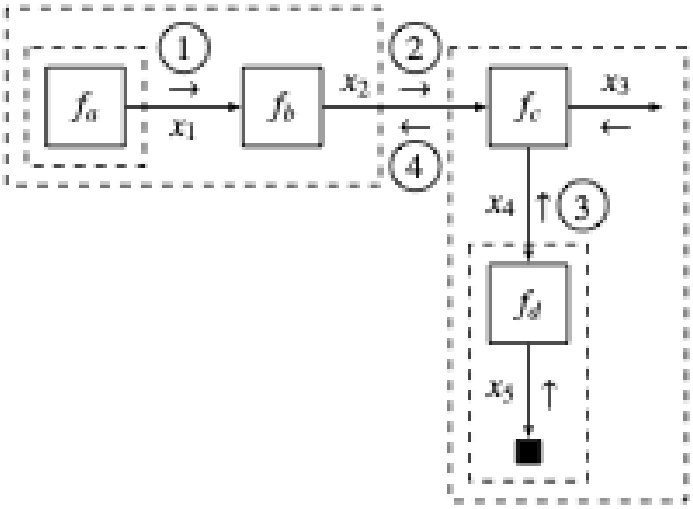


Figure A.2: *Visualization of the message passing schedule corresponding to Eq. A.5 with observed variable $X_5 = \hat{X}_5$. The observation is indicated by terminating edge X_5 by a small solid node that technically represents the factor $\delta(X_5 - \hat{X}_5)$. Messages are represented by numbered arrows, and the message sequence is chosen such that there are only backward dependencies. Dashed boxes mark the parts of the graph that are covered by the respective messages coming out of those boxes. The marginal posterior distribution $f(X_2 \mid X_5 = \hat{X}_5)$ is obtained by taking the product of the messages that flow on edge X_2 and normalizing.*

by

$$\mu_3(Z) = \iint \mu_1(X)\,\mu_2(Y)\,f_=(X,Y,Z)dX\,dY \tag{A.6}$$

$$= \mu_1(Z)\,\mu_2(Z)\,. \tag{A.7}$$

Note that the outgoing message is only a function of Z and that it is calculated from only node-local information, namely the incoming messages at the corresponding node and the definition of the node factor itself. Equality constraint nodes are quite prevalent in FFGs because they constitute a branching mechanism that distributes variables over multiple (more than two) factors in the graph. If we interpret message $\mu_1(\cdot)$ as a prior and message $\mu_2(\cdot)$ as a likelihood function, then message $\mu_3(\cdot)$ becomes proportional to the posterior distribution over z. Therefore, message passing through the equality node effectively fuses information from two sources by executing Bayes rule (up to a normalizing constant).

Figure A.3: *FFG representation (left) and message passing schedule (right) for an equality constraint node.*

APPENDIX $\mathcal{B}$

Active Inference: the bare essentials

Active inference provides a unifying theory for perception, action, decision-making, and learning in biological or artificial agents (Da Costa et al., 2020). Our active inference agent rests on the tuple $(\mathcal{O}, \mathcal{S}, \mathcal{A}, P, Q)$. This is composed of: a finite set of observations $\mathcal{O}$, a finite set of states $\mathcal{S}$, a finite set of actions $\mathcal{A}$, a generative model P and an approximate posterior Q.

Active inference proposes a solution for action and perception by assuming that actions will fulfill predictions that are based on inferred states of the world, given some observations. The generative model contains beliefs about future states and action plans, where plans that lead to preferred observations are more likely. Perception and action are achieved through the optimisation of two complementary objective functions, the variational free-energy F, and the expected free-energy G. These quantities to optimise are derived based on the generative model and the approximate posterior, as detailed later.

Variational free-energy measures the fit between the generative model and past and current sensory observations, while expected free-energy scores future possible courses of action according to prior preferences and predicted observations. Fig. B.1 depicts the general high level idea.

For a first time reader of active inference, we advise consulting Friston et al. (2017) for a more extensive introduction.

In the following, we explain the form of the generative model and how the model parameters relate to states, actions, and observations. Based on this model, we present the expressions for the free-energy and expected free-energy that are used to derive the equations for perception and decision making.

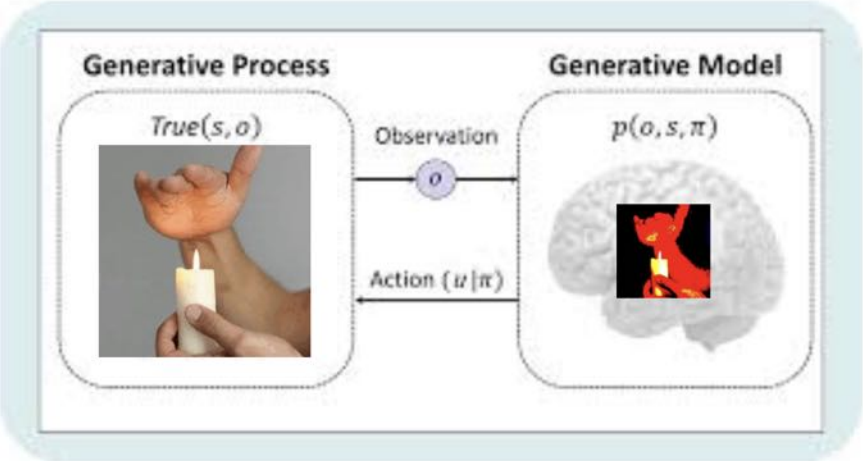


Figure B.1: *High-level visualization of an active inference agent. The generative process describes the true causes of the agent's observations $O = o$ that might be visual and/or nociceptive. An agent can apply actions $A = u$ (move the hand) under a policy π (avoid skin damage) to change the state of the world and to get observations that are aligned with its internal preferences/beliefs that are part of its generative model.*

B.1 Generative models

In active inference (Friston et al., 2017), the generative model P is chosen to be a Markov process that allows to infer the states of the environment and to predict the effects of actions as well as future observations. This is expressed as a joint probability distribution $P(\bar{o}, \bar{s}, \eta, \pi)$, where $\bar{o}$ is a sequence of observations, $\bar{s}$ is a sequence of states, η represents model parameters, and π is a plan.

By using the chain rule, we can write:

$$P(\bar{o}, \bar{s}, \eta, \pi) = P(\bar{o}|\bar{s}, \eta, \pi)P(\bar{s}|\eta, \pi)P(\eta|\pi)P(\pi) \tag{B.1}$$

Note that $\bar{o}$ is conditionally independent from the model parameters η and π given $\bar{s}$. In addition, under the Markov property, the next state and current observations depend only on the current state:

$$P(\bar{o}|\bar{s}, \eta, \pi) = \prod_{\tau=1}^{T} P(o_\tau|s_\tau) \tag{B.2}$$

The model is further simplified considering that $\bar{s}$ and η are conditionally independent given π:

$$P(\bar{s}|\eta, \pi) = \prod_{\tau=1}^{T} P(s_\tau|s_{\tau-1}, \pi) \tag{B.3}$$

Finally, consider the model parameters explicitly:

$$P(\bar{o}, \bar{s}, \eta, \pi) = P(\bar{o}, \bar{s}, A, B, D, \pi) =$$
$$P(\pi)P(A)P(B)P(D)\prod_{\tau=1}^{T} P(s_\tau|s_{\tau-1}, \pi)P(o_\tau|s_\tau) \tag{B.4}$$

$P(A), P(B), P(D)$ are Dirichlet distributions over the model parameters (Friston et al., 2017). In case the model parameters are fixed by the user it holds:

$$P(\bar{o}, \bar{s}, \pi) = P(\pi) \prod_{\tau=1}^{T} P(s_\tau | s_{\tau-1}, \pi) P(o_\tau | s_\tau) \tag{B.5}$$

The probability distributions in eq. (**??**) are represented internally by an active inference agent through the following parameters:

- $A \in [0,1]^{r,m}$ is a matrix representation of the conditional probability $P(o_\tau|s_\tau)$, where r is the number of possible observations and m the number of possible states. A is also called the likelihood matrix, and it indicates the probability of observations given a specific state. Each column of A is a categorical distribution. It holds that $P(o_\tau|s_\tau, A) = Cat(As_\tau)$. For a generic entry $A_{ij} = P(o_\tau = i|s_\tau = j)$.
- $B \in [0,1]^{m,m}$ represents a transition matrix. In particular $P(s_{\tau+1}|s_\tau,\ a_\tau) = Cat(B_{a_\tau} s_\tau)$. For a symbolic action a_τ in a plan π, B_{a_τ} represents the probability of state $s_{\tau+1}$ while applying action a_τ from state s_τ. The columns of B_{a_τ} are categorical distributions.
- π is a sequence of actions over a time horizon T. π is the posterior distribution, a vector holding the probability of different plans. These probabilities depend on the expected free-energy in future time steps under plans given the current belief: $P(\pi) = \sigma(-G(\pi))$. Here, σ indicates the softmax function used to normalize probabilities.

An active inference agents contains also a model $D \in [0,1]^m$ that represents the belief about the initial state at $\tau = 1$. So $P(s_0) = Cat(D)$. Additionally, an agent also represents prior preferences about desired observations for goal-directed behavior in $C \in \mathbb{R}^r$, such that $P(o_\tau) = C$.

Additional generative model parameters could be considered such as the E vector to encode priors over plans used to represent habits (Smith et al., 2022; Hesp et al., 2021). The parameter E could be used to include common sense knowledge in the decision making process.

The above model represents the minimal "kernel" of AIF and it is summarised in Figure B.2, in the form of a Forney-like factor graph (see. Appendix A)

Table B.1 summarises the notation adopted. The top part contains quantities computed by active inference, while the bottom part the domain parameters required. These internal models will be used by an active inference agent to compute the free-energy and the expected free-energy.

Given the generative model above, we are interested in finding the posterior hidden causes of sensory data. For the sake of these derivations, we consider that the parameters associated with the task are known and do not introduce uncertainty. Using Bayes rule:

$$P(\bar{s}, \pi|\bar{o}) = \frac{P(\bar{o}|\bar{s}, \pi) P(\bar{s}, \pi)}{P(\bar{o})} \tag{B.6}$$

Computing the model evidence $P(\bar{o})$ exactly is a well-known and often intractable problem in Bayesian statistics. The exact posterior is then computed minimizing the

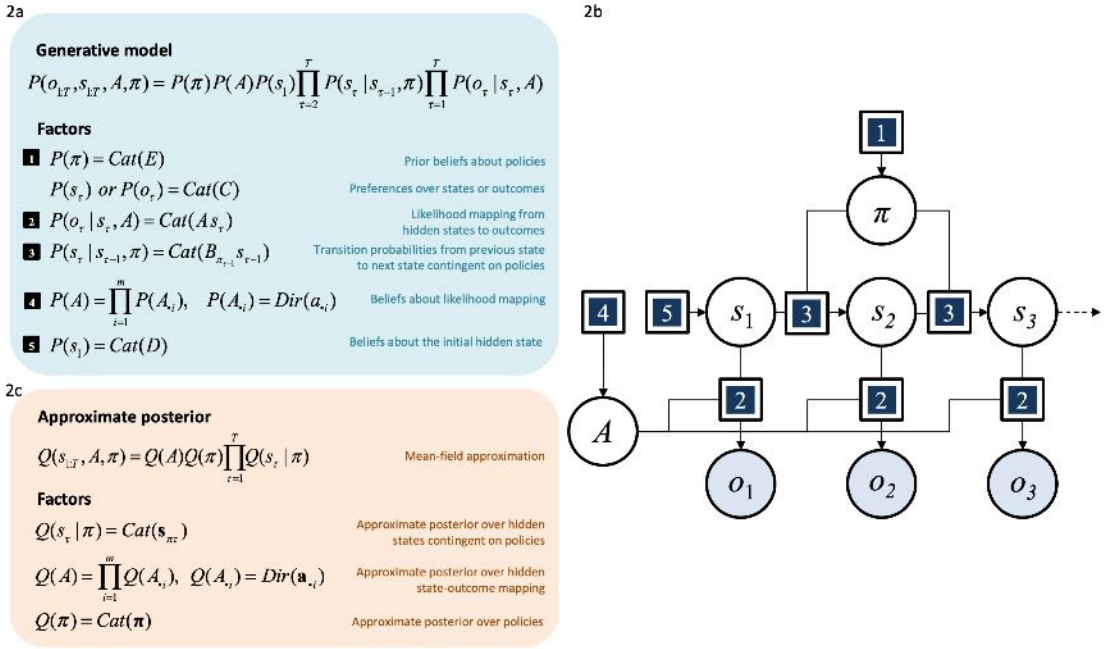


Figure B.2: ***Example of a discrete state-space generative model.*** *Panel 2a, specifies the form of the generative model, which is how the agent represents the world. The generative model is a joint probability distribution over (hidden) states, outcomes and other variables that cause outcomes. In this representation, states unfold in time causing an observation at each time-step. The likelihood matrix A encodes the probabilities of state-outcome pairs. The policy π specifies which action to perform at each time-step. Note that the agent's preferences may be specified either in terms of states or outcomes. It is important to distinguish between states (resp. outcomes) that are random variables, and the possible values that they can take in S (resp. in O), which we refer to as possible states (resp. possible outcomes). Note that this type of representation comprises a finite number of timesteps, actions, policies, states, outcomes, possible states and possible outcomes. In Panel 2b, the generative model is displayed as a probabilistic graphical model (cfr Appendix A)expressed in factor graph form (***?***). The variables in circles are random variables, while squares represent factors, whose specific form are given in Panel 2a. The arrows represent causal relationships (i.e., conditional probability distributions). The variables highlighted in grey can be observed by the agent, while the remaining variables are inferred through approximate Bayesian inference and called hidden or latent variables. Active inference agents perform inference by optimising the parameters of an approximate posterior distribution. Panel 2c specifies how this approximate posterior factorises under a particular mean-field approximation. A glossary of terms used in this figure is available in Table B.1. The mathematics of generative models is heavily dependent on Markov blankets. The Markov blanket of a random variable in a probabilistic graphical model are those variables that share a common factor. Crucially, a variable conditioned upon its Markov blanket is conditionally independent of all other variables. We will use this property extensively (and implicitly) in the text.*

Kullback-Leibler divergence (D_{KL}, or KL-Divergence) with respect to an approximate posterior distribution $Q(\bar{s}, \pi)$. Doing so, we can define the free-energy as a functional of approximate posterior beliefs which result in an upper bound on surprise. By definition D_{KL} is a non-negative quantity given by the expectation of the logarithmic difference between $Q(\bar{s}, \pi)$ and $P(\bar{s}, \pi|\bar{o})$. Applying the KL-Divergence:

$$\begin{aligned} &D_{KL}\left[Q(\bar{s}, \pi)||P(\bar{s}, \pi|\bar{o})\right] = \\ &\mathbb{E}_{Q(\bar{s},\pi)}\left[\ln Q(\bar{s}, \pi) - \ln P(\bar{s}, \pi|\bar{o})\right] \geq 0 \end{aligned} \tag{B.7}$$

D_{KL} is the information loss when Q is used instead of P. Considering equation (B.6) and the chain rule, equation (B.7) can be rewritten as:

$$\begin{aligned} D_{KL}\left[\cdot\right] &= \mathbb{E}_{Q(\bar{s},\pi)}\left[\ln Q(\bar{s}, \pi) - \ln \frac{P(\bar{o}, \bar{s}, \pi)}{P(\bar{o})}\right] \\ &= \underbrace{\mathbb{E}_{Q(\bar{s},\pi)}\left[\ln Q(\bar{s}, \pi) - \ln P(\bar{o}, \bar{s}, \pi)\right]}_{F[Q(\bar{s},\pi)]} + \ln P(\bar{o}) \end{aligned} \tag{B.8}$$

We have just defined the free-energy as the upper bound of surprise:

$$F\left[Q(\bar{s}, \pi)\right] \geq -\ln P(\bar{o}) \tag{B.9}$$

B.2 Variational Free-energy

To fully characterize the free-energy in equation (B.8), we need to specify a form for the approximate posterior $Q(\bar{s}, \pi)$. There are different ways to choose a family of probability distributions (Schwöbel et al., 2018), compromising between complexity and accuracy of the approximation. In this work, we choose the mean-field approximation. It holds:

$$Q(\bar{s}, \pi) = Q(\bar{s}|\pi)Q(\pi) = Q(\pi)\prod_{\tau=1}^{T} Q(s_\tau|\pi) \tag{B.10}$$

Under mean-field approximation, the plan-dependent states at each time step are approximately independent of the states at any other time step. We can now find an expression for the variational free-energy. Considering the mean-field approximation and the generative model in eq. (B.5) we can write:

$$\begin{aligned} F\left[Q(\bar{s}, \pi)\right] = \mathbb{E}_{Q(\bar{s},\pi)}\Bigg[&\ln Q(\pi) + \sum_{\tau=1}^{T} \ln Q(s_\tau|\pi) \\ &- \ln P(\pi) - \sum_{\tau=1}^{T} \ln P(s_\tau|s_{\tau-1}, \pi) - \sum_{\tau=1}^{T} \ln P(o_\tau|s_\tau)\Bigg] \end{aligned} \tag{B.11}$$

Since $Q(\bar{s}, \pi) = Q(\bar{s}|\pi)Q(\pi)$, and since the expectation of a sum is the sum of the expectation, we can write:

$$F\left[\cdot\right] = D_{KL}\left[Q(\pi)||P(\pi)\right] + \mathbb{E}_{Q(\pi)}\left[F(\pi)\left[Q(\bar{s}|\pi)\right]\right] \tag{B.12}$$

where

$$F(\pi)\left[Q(\bar{s}|\pi)\right] = \mathbb{E}_{Q(\bar{s}|\pi)}\Bigg[\sum_{\tau=1}^{T}\ln Q(s_\tau|\pi) - \sum_{\tau=1}^{T}\ln P(s_\tau|s_{t-\tau},\pi) - \sum_{\tau=1}^{T}\ln P(o_\tau|s_\tau)\Bigg] \tag{B.13}$$

One can notice that $F(\pi)$ is accumulated over time, or in other words, it is the sum of free energies over time and plans:

$$F(\pi) = \sum_{\tau=1}^{T} F(\pi,\tau) \tag{B.14}$$

Substituting the agent's belief about the current state at time τ given π with s_τ^π, we obtain a matrix form for $F(\pi,\tau)$ that we can compute given the generative model:

$$F(\pi) = \sum_{\tau=1}^{T} s_\tau^{\pi\top}\Big[\ln s_\tau^\pi - \ln\left(B_{a_{\tau-1}} s_{\tau-1}^\pi\right) - \ln\left(A^\top o_\tau\right)\Big] \tag{B.15}$$

Given a plan π, the probability of state transition $P(s_\tau|s_{\tau-1},\pi)$ is given by the transition matrix under plan π at time τ, multiplied by the probability of the state at the previous time step. In the special case of $\tau = 1$, we can write:

$$F(\pi,1) = s_1^{\pi\top}\big[\ln s_1^\pi - \ln D - \ln\left(A^\top o_1\right)\big] \tag{B.16}$$

Finally, we can compute the expectation of the plan dependant variational free-energy $F(\pi)$ as $\mathbb{E}_{Q(\pi)}\big[F(\pi)\big] = \pi^\top F_\pi$. We indicate $F_\pi = (F(\pi_1), F(\pi_2)...)^\top$ for every allowable plan. To derive state and plan updates that minimize free-energy, F in equation (B.12) is partially differentiated and set to zero, as we will see in the next appendixes.

B.3 Perception and State estimation

According to active inference, both perception and decision making are based on the minimization of free-energy. In particular, for state estimation, we take partial derivatives of F with respect to the states and set the gradient to zero.

We differentiate F with respect to the sufficient statistics of the probability distribution of the states. Note that the only part of F dependent on the states is $F(\pi)$. Then:

$$\frac{\partial F}{\partial s_\tau^\pi} = \frac{\partial F}{\partial F(\pi)}\frac{\partial F(\pi)}{\partial s_\tau^\pi} = \pi^\top\big[1 + \ln s_\tau^\pi - \ln\left(B_{a_{\tau-1}} s_{\tau-1}^\pi\right) - \ln\left(B_{a_\tau}^\top s_{\tau+1}^\pi\right) - \ln\left(A^\top o_\tau\right)\big] \tag{B.17}$$

Setting the gradient to zero and using the softmax function for normalization:

$$s_\tau^\pi = \sigma(\ln\left(B_{a_{\tau-1}} s_{\tau-1}^\pi\right) + \ln\left(B_{a_\tau}^\top s_{\tau+1}^\pi\right) + \ln\left(A^\top o_\tau\right)) \tag{B.18}$$

Note that the softmax function is insensitive to the constant 1. Also, for $\tau = 1$ the term $\ln(B_{a_{\tau-1}} s^{\pi}_{\tau-1})$ is replaced by D. Finally, $\ln(A^{\top} o_{\tau})$ contributes only to past and present time steps, so for this term is null for $t < \tau \leq T$ since those observations are still to be received.

Eventually, the posterior distribution of the state, conditioned by a plan, is given by:

$$s^{\pi}_{\tau=1} = \sigma(\ln D + \ln(B^{\top}_{a_\tau} s^{\pi}_{\tau+1}) + \ln(A^{\top} o_{\tau})) \tag{B.19a}$$

$$s^{\pi}_{1<\tau<T} = \sigma(\ln(B_{a_{\tau-1}} s^{\pi}_{\tau-1}) + \ln(B^{\top}_{a_\tau} s^{\pi}_{\tau+1}) + \ln(A^{\top} o_{\tau})) \tag{B.19b}$$

$$s^{\pi}_{\tau=T} = \sigma(\ln(B_{a_{\tau-1}} s^{\pi}_{\tau-1}) + \ln(A^{\top} o_{\tau})) \tag{B.19c}$$

where σ is the softmax function. The column of $B^{\top}_{a_\tau}$ are normalized.

B.4 Expected Free-energy

Active inference unifies action selection and perception by assuming that actions fulfill predictions based on inferred states. Since the internal model can be biased toward preferred states or observations (*prior desires*), active inference induces actions that will bring the current beliefs towards the preferred states. An agent builds beliefs about future states which are then used to compute the expected free-energy. The latter is necessary to evaluate alternative plans. Plans that lead to preferred observations are more likely. Preferred observations are specified in the model parameter C. This enables action to realize the next (proximal) *observation* predicted by the plan that leads to (distal) goals.

We indicate with $G(\pi)$ the expected free-energy obtained over future time steps until the time horizon T while following a plan π. Basically, this is the variational free-energy of future trajectories which measures the plausibility of plans according to future predicted observations (Sajid et al., 2021). To compute it we take the expectation of variational free-energy under the posterior predictive distribution $P(o_\tau|s_\tau)$. Following Sajid et al. (2021) we can write:

$$G(\pi) = \sum_{\tau=t+1}^{T} G(\pi, \tau) \tag{B.20}$$

then:

$$\begin{aligned} G(\pi,\tau) &= \mathbb{E}_{\tilde{Q}}\big[\ln Q(s_\tau|\pi) - \ln P(o_\tau, s_\tau|s_{\tau-1})\big] \\ &= \mathbb{E}_{\tilde{Q}}\big[\ln Q(s_\tau|\pi) - \ln P(s_\tau|o_\tau, s_{\tau-1}) - \ln P(o_\tau)\big] \end{aligned} \tag{B.21}$$

where $\tilde{Q} = P(o_\tau|s_\tau)Q(s_\tau|\pi)$. The expected free-energy is:

$$G(\pi,\tau) \geq \mathbb{E}_{\tilde{Q}}\big[\ln Q(s_\tau|\pi) - \ln Q(s_\tau|o_\tau, s_{\tau-1}, \pi) - \ln P(o_\tau)\big] \tag{B.22}$$

Equivalently, we can express the expected free-energy in terms of preferred observations (Da Costa et al., 2020):

$$G(\pi,\tau) = \mathbb{E}_{\tilde{Q}}\big[\ln Q(o_\tau|\pi) - \ln Q(o_\tau|s_\tau, s_{\tau-1}, \pi) - \ln P(o_\tau)\big] \tag{B.23}$$

Making use of $Q(o_\tau|s_\tau,\pi) = P(o_\tau|s_\tau)$ since the predicted observations in the future are only based on A which is plan independent given s_τ, we have:

$$G(\pi,\tau) = \underbrace{D_{KL}\left[Q(o_\tau|\pi)||P(o_\tau)\right]}_{Expected\ cost} + \underbrace{\mathbb{E}_{Q(s_\tau|\pi)}\left[H(P(o_\tau|s_\tau))\right]}_{Entropy} \tag{B.24}$$

were $H[P(o_\tau|s_\tau)] = \mathbb{E}_{P(o_\tau|s_\tau)}\left[-\ln P(o_\tau|s_\tau)\right]$ is the entropy. We are now ready to express the expected free-energy in matrix form, such that we can compute it. From the previous equation, one can notice that plan selection aims at minimizing the expected cost and ambiguity. The latter relates to the uncertainty about future observations given hidden states. In a sense, plans tend to bring the agent to future states that generate unambiguous information over states. On the other hand, the cost is the difference between predicted and prior beliefs about final states. Plans are more likely if they minimize cost, and lead to observations that match prior desires. Minimizing G leads to both exploitative (cost minimizing) and explorative (ambiguity minimizing) behavior. This results in a balance between goal-oriented and novelty-seeking behaviors

Substituting the sufficient statistics in equation (B.24), and recalling that the generative model specifies $P(o_\tau) = C$, one obtains (Smith et al., 2022):

$$G(\pi,\tau) = \underbrace{o_\tau^{\pi\top}\left[\ln o_\tau^\pi - \ln C\right]}_{Reward\ seeking} \underbrace{-diag(A^\top \ln A)^\top s_\tau^\pi}_{Information\ seeking} \tag{B.25}$$

Note that prior preferences are passed through the softmax function before computing the logarithm.

B.5 Planning and Decision Making: Updating plan distribution

The update rule for the distribution over possible plans follows directly from the variational free-energy:

$$F\left[\cdot\right] = D_{KL}\left[Q(\pi)||P(\pi)\right] + \pi^\top F_\pi \tag{B.26}$$

The first term of the equation above can be written as:

$$D_{KL}\left[Q(\pi)||P(\pi)\right] = \mathbb{E}_{Q(\pi)}\left[\ln Q(\pi) - \ln P(\pi)\right] \tag{B.27}$$

Recalling that the approximate posterior over policies is a softmax function of the expected free-energy $Q(\pi) = \sigma(-G(\pi))$ (Da Costa et al., 2020; Friston et al., 2017), and taking the gradient with respect to π it results:

$$\frac{\partial F}{\partial \pi} = \ln \pi + G_\pi + F_\pi + 1 \tag{B.28}$$

where $G_\pi = (G(\pi_1), G(\pi_2), ...)^\top$ Finally, setting the gradient to zero and normalizing through softmax, the posterior distribution over plans is obtained:

$$\pi = \sigma(-G_\pi - F_\pi) \tag{B.29}$$

where the vector π encodes the posterior distribution over plans reflecting the predicted value of each plan. $F_\pi = (F(\pi_1), F(\pi_2), ...)^\top$ and $G_\pi = (G(\pi_1), G(\pi_2), ...)^\top$.

The plan that an agent should pursue is the most likely one.

Plan independent state-estimation Given the probability over p possible plans, and the plan dependent states s_τ^π, we can compute the overall probability distribution for the states over time through Bayesian Model Average:

$$s_\tau = \sum_i s_\tau^{\pi_i} \pi_i, \text{ where } i \in \{1, ..., p\} \tag{B.30}$$

were $s_\tau^{\pi_i}$ is the probability of a state at time τ under plan i and π_i is the probability of plan i. This is the average prediction for the state at a certain time, so s_τ, according to the probability of each plan. In other words, this is a weighted average over different models. Models with high probability receive more weight, while models with lower probabilities are discounted.

Action selection The action for the agent to be executed is the first action of the most likely plan:

$$\lambda = \max(\underbrace{[\pi_1, \pi_2, ..., \pi_p]}_{\pi^\top}), \; a_\tau = \pi_\lambda(\tau = 1) \tag{B.31}$$

where λ is the index of the most likely plan.

The active inference algorithm is summarised in pseudo-code in Algorithm 15.

Algorithm 15 Action selection with active inference

Set C ▷ prior preferences
for $\tau = 1 : T$ **do**
    If not specified, get state from D if $\tau == 1$
    If not specified, get observation from A
    Compute F for each plan ▷ eq. (B.15)
    Update posterior state s_τ^π ▷ eq. (B.19)
    Compute G for each plan ▷ eq. (B.25)
    Bayesian model averaging ▷ eq. (B.30)
    Action selection ▷ eq. (B.31)
end for
Return a ▷ Preferred action

Appendix B. Active Inference: the bare essentials

Table B.1: *Notation for Active Inference*

Symbol Description
$s_\tau \in \{0,1\}^m$ One-hot encoding of the hidden state at time τ, with m mutually exclusive states in the discrete state space.
$s_\tau^\pi \in [0,1]^m$ Posterior distribution over the state under a plan π, where the elements sum up to one.
$o_\tau \in \{0,1\}^r$ Observation at time τ that can have r mutually exclusive possible values.
$o_\tau^\pi \in [0,1]^r$ Posterior distribution of observations under a plan.
π Plan specifying a sequence of symbolic actions $\pi = [a_\tau,\ a_{\tau+1}, ..., a_T]^\top$, where T is the time horizon.
$\pi \in [0,1]^p$ Posterior distribution over plans, where p is the number of different plans.
$F(\pi) \in \mathbb{R}$ Plan specific variational free-energy.
$F_\pi \in \mathbb{R}^p$ $F_\pi = (F(\pi_1), F(\pi_2), ...)^\top$ is a column vector containing the free-energy for every plan.
$G(\pi, \tau) \in \mathbb{R}$ Expected free-energy for a plan at time τ.
$G_\pi \in \mathbb{R}^p$ $G_\pi = (G(\pi_1), G(\pi_2), ...)^\top$ is a column vector containing the expected free-energy for every plan.
$A \in [0,1]^{r \times m}$ Likelihood matrix, mapping from hidden states to observations $P(o_\tau \vert s_\tau, A) = Cat(As_\tau)$.
$B_{a_\tau} \in [0,1]^{m \times m}$ Transition matrix, $P(s_{\tau+1} \vert s_\tau,\ a_\tau) = Cat(B_{a_\tau} s_\tau)$.
$C \in \mathbb{R}^r$ Prior preferences over observations $P(o_\tau) = C$.
$D \in [0,1]^m$ Prior over initial states $P(s_0) = Cat(D)$.
σ Softmax function.

APPENDIX C

Neurobiology of Fear Generalisation

In 2014 Lissek proposed a neural model illustrating the areas of the brain involved in the fear generalisation tasks (Lissek et al., 2014). This model suggests that the generalisation of conditioned fear involves a network of brain regions, including the hippocampus, sensory cortex, amygdala, insula, and vmPFC. These regions are associated with both fear excitation and inhibition. The model is supported by research in classical conditioning conducted in both animals and humans.

As Figure C.1 shows, after acquiring fear to CS+, when encountering a stimulus similar to CS+ (i.e., GS3), the thalamus plays a role in transmitting sensory information about GS3 to the amygdala-based fear circuits through a rapid and less detailed pathway. This leads to an initial fear response to GS3.

The thalamus performs a dual role by transmitting sensory information about GS3 to the visual cortices, allowing for more advanced sensory processing. This route is slower, but it activates neural representations of GS3 in the visual cortex.

The process involves the assessment of the similarity between brain activity patterns representing GS3 and those previously associated with CS+. When there's significant overlap, CA3 neurons in the hippocampus initiate "pattern completion," activating the stored pattern linked to the past experience (CS+). This pattern completion leads to the activation of brain structures associated with fear excitation (highlighted in yellow, including anterior insula, dACC, amygdala), ultimately resulting in the physiological and behavioural fear response.

In 2021, Lissek proposed an updated and more detailed model of conditioned fear generalisation describing also neural structures activated for positive and negative generalisation (Webler et al., 2021) (see Figure C.2).

As already described in the first model, the sensory thalamic nuclei play a role by sending visual information about the GS to the amygdala-based fear circuits via a

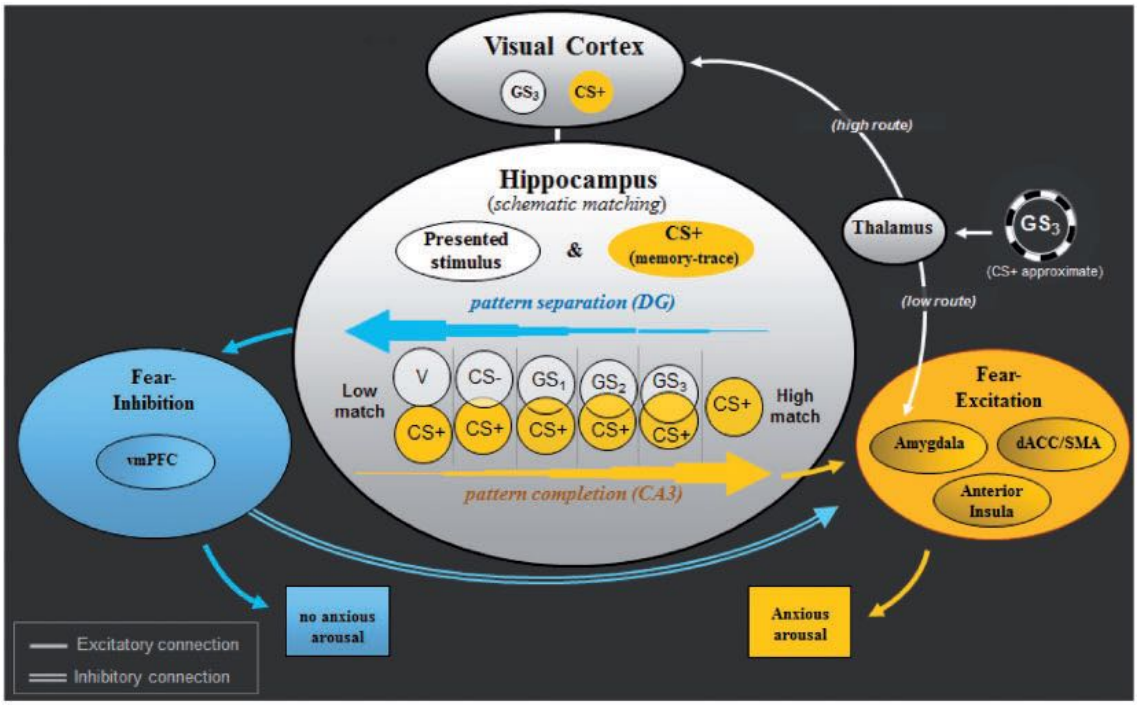


Figure C.1: *Neural model of conditioned fear generalisation. (Figure from Lissek et al. (2014))*

rapid "low road" and to visual cortices through a "high road." Projections from the basolateral amygdala (BLA) and the locus coeruleus (LC), activated by the low road, target specific hippocampal subfields (CA1 and CA3), setting the stage for "pattern completion" in the hippocampus.

Meanwhile, the high road conveys detailed visual representations of the GS to the hippocampus. Here, the degree of overlap between the cortical representation of the GS and the previously encoded CS+ is evaluated. If there's sufficient alignment between the GS/CS+ representations and GS-induced activity in the BLA-CA1 and LC-CA3 pathways, the hippocampus initiates pattern completion. This triggers the activation of brain structures associated with threat processing, including the amygdala, anterior insula (AI), dorsomedial prefrontal cortex (dmPFC), periaqueductal gray (PAG), and LC. This activation results in the autonomic, neuroendocrine, and behavioural components of the generalised threat response.

These threat-related activations then engage executive control areas of the brain, such as the inferior parietal lobule (IPL), dorsolateral prefrontal cortex (dlPFC), and ventrolateral prefrontal cortex (vlPFC). These areas facilitate attentional and emotion-regulation processes to optimize responses.

In cases where there's an insufficient overlap between GS/CS+ representations and LC signaling, dentate gyrus neurons in the hippocampus implement "pattern separation," which activates default mode structures associated with fear inhibition and a return to a resting state. These structures include the ventromedial prefrontal cortex (vmPFC), middle temporal gyrus (MTG), and angular gyrus (AG). This default mode activation reduces ongoing activity in amygdala-based fear networks initiated earlier by the low road, thereby alleviating generalised anxiety.

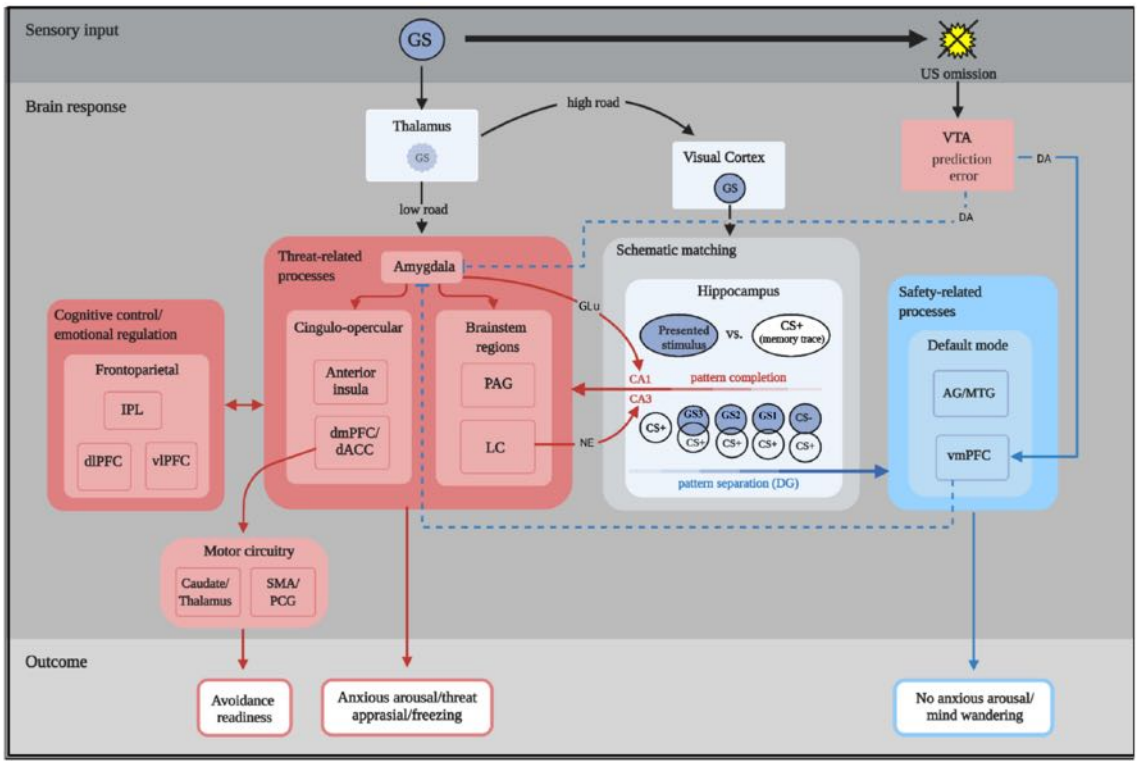


Figure C.2: *Updated neural working model of conditioned fear generalisation incorporating neural structures with activations shown or proposed to fall along positive (red structures) and negative gradients of generalisation (blue structures). (Figure from Webler et al. (2021))*

Lastly, when GS presentations are unexpectedly followed by the absence of an aversive stimulus (US), a positive prediction error signal is triggered in the ventral tegmental area (VTA) through dopaminergic pathways. This supports safety learning, strengthening the association between the GS and safety cues in the amygdala and vmPFC. This prediction error process contributes to the attenuation of generalised fear with repeated exposure to unreinforced GSs.

The work by Webler et al. (2021) delineates brain substrates of conditioned fear-generalisation and formulates a working neural model. It reported whole-brain fMRI results and applied generalisation-gradient methodology to identify brain activations that gradually strengthen (positive generalisation) or weaken (negative generalisation) as presented stimuli increase in CS+ resemblance. Positive generalisation was instantiated in cingulo-opercular, frontoparietal, striatalthalamic, and midbrain regions (locus coeruleus, periaqueductal grey, ventral tegmental area), while negative generalisation was implemented in default-mode network nodes (ventromedial prefrontal cortex, hippocampus, middle temporal gyrus, angular gyrus) and amygdala.

In the following discussion, based on the fMRI study by Lissek (Webler et al., 2021), we will elaborate on the potential psychological contributions of these key regions involved in positive and negative generalisation.

C.0.1 Neural substrates of positive generalisation

Cingulo-opercular loci

The cingulo-opercular network is known for detecting important environmental events and facilitating cognitive processes to respond effectively. Lissek et al. identified two key nodes within this network, the anterior insula (AI) and the dorsal medial prefrontal cortex/anterior cingulate cortex (dmPFC/dACC), as being involved in positive generalisation. This suggests that they play a significant role in detecting threats and that this sensitivity gradually diminishes as stimuli become less similar to the threat cue (CS+).

Furthermore, these nodes may have distinct functions related to generalisation. The AI is associated with being aware of bodily sensations related to fear, so positive generalisation effects in AI might indicate an increasing awareness of anxious bodily states as presented stimuli resemble the CS+. On the other hand, the dmPFC/dACC has been linked to fear responses in both rodents and humans, particularly in evaluating risk in the fear response. Positive generalisation effects in dmPFC/dACC may signify an increasing perception of risk as stimuli become more similar to the CS+.

Frontoparietal regions

The frontoparietal network plays a crucial role in various advanced cognitive functions such as attention, cognitive control, and emotional regulation, as supported by previous research (Marek and Dosenbach, 2018; Rees et al., 2002)). In the study by Webler et al. (2021), two specific regions within the frontoparietal network, namely the left prefrontal cortex (lPFC) encompassing the dorsolateral prefrontal cortex (dlPFC) and ventrolateral prefrontal cortex (vlPFC), as well as the inferior parietal lobule (IPL), displayed positive generalisation effects.

Lateral prefrontal cortex (lPFC)

According to attentional control theory, anxiety can impair goal-directed attention by increasing cognitive load due to heightened stimulus-driven attention. Following this theory, anxiety-induced increases in cognitive load might have required greater engagement of the lPFC to perform these task-related activities. Therefore, positive generalisation effects in the lPFC may reflect the increased cognitive load driven by threat-related stimuli, which scales with the similarity of the stimuli to the CS+.

The second interpretation is supported by the well-established roles of the dorsolateral prefrontal cortex (dlPFC) and ventrolateral prefrontal cortex (vlPFC) in emotion regulation. Studies involving neuromodulation techniques have suggested that the lPFC can down-regulate negative emotions by inhibiting subcortical structures like the amygdala. For example, both repetitive transcranial magnetic stimulation (rTMS) and transcranial direct current stimulation (tDCS) have been shown to reduce amygdala activation in response to negatively valent stimuli in individuals with high trait anxiety. This interpretation suggests that the positive generalisation effects observed in the lPFC may reflect increased efforts to regulate fear by inhibiting the amygdala-based fear network, with the level of effort scaling with the degree of similarity between presented stimuli and the CS+.

Inferior parietal lobule

While the inferior parietal lobule (IPL) has been associated with various cognitive functions such as attentional reorienting, working memory, and memory retrieval, a recent theoretical perspective suggests that its primary role is in directing attention towards salient external events or attention-grabbing episodic memories. According to this theory, positive generalisation in the IPL may signify a shift of attention towards the external cue or relevant internal representations that are most closely associated with the highly threatening CS+. As the perceptual similarity to the CS+ decreases, this attentional shift towards the cue or internal representations gradually diminishes. In essence, positive generalisation in the IPL reflects the attentional response to external cues or memories that are most closely linked to the most threatening stimulus (CS+), with this response decreasing as stimuli become less similar to the CS+.

Brainstem nuclei

In line with the brainstem's pivotal function in generating autonomic and behavioural reactions to emotionally significant stimuli, three specific brainstem nuclei—namely, the locus coeruleus (LC), periaqueductal gray (PAG), and ventral tegmental area (VTA)—exhibited positive generalisation effects.

Striatal-thalamic areas

The striatum and thalamus play crucial roles in an "action-selection" circuit that helps us choose and carry out motivated behaviours. The striatal nuclei, including the caudate, serve as the input for this circuit, indicating whether a specific action should be performed or stopped. After further processing in other basal ganglia nuclei, selected actions are executed by releasing motoric thalamic nuclei (ventral lateral and ventral anterior nuclei; VLN, VAN) from inhibition. Recent findings suggesting positive generalisation effects in key regions of this circuit (caudate, VLN, VAN) may indicate an increased readiness for defensive responses to cues associated with higher threat levels.

Apart from its motoric functions, the thalamus also includes sensory processing regions. The pulvinar, the largest thalamic nucleus, is involved in processing important visual information. Additionally, there is evidence of heightened connectivity between the pulvinar and the amygdala when masked conditioned cues are presented, indicating the existence of a rapid pulvinar-amygdala visual pathway. In the study by Webler et al. (2021), the positive generalisation effects observed in the pulvinar may signify enhanced visual processing of biologically relevant cues and their close perceptual counterparts, suggesting the involvement of a rapid thalamic-amygdala circuit for threat processing.

C.0.2 Neural substrates of negative generalisation

Regions implicated in the default mode network

The default mode network is linked to various mental processes such as self-referential, spontaneous thinking (as mentioned by Andrews-Hanna et al. (2014)), remembering events from the past or planning for the future (as highlighted by Buckner et al. (2008)),

and, more recently, responding to safety cues in situations that pose a threat (as indicated by Marstaller et al. (2017)). Notably, certain regions of the default mode network, including both sides of the ventromedial prefrontal cortex (vmPFC) and specific areas like the left lateralized middle temporal gyrus (MTG), angular gyrus (AG), and anterior hippocampus, displayed negative generalisation effects (as shown in Webler et al. (2021)).

Amygdala

In line with numerous prior fMRI studies on fear-conditioning in humans, the analysis by Webler et al. (2021) did not reveal an increase in amygdala activation in response to the CS+ (conditioned stimulus associated with a threat). Instead, a decrease was observed in the amygdala reactivity to the CS+ as it became more distinct from other presented stimuli.

C.0.3 Animal evidence

In experiments on animals, studies have shown that lesions in either the hippocampus itself (as demonstrated by Wild and Blampied (1972) and Solomon and Moore (1975)) or the cortical inputs to the hippocampus (specifically, the postrhinal and perirhinal cortex, as observed by Bucci et al. (2002)) lead to an increased tendency to generalise fear responses from CS+ to CS-. These findings indicate that the activation of the hippocampus plays a crucial role in the successful discrimination between CS+ and CS-. This role may be attributed to the hippocampus's function in "pattern separation," as proposed by O'Reilly and Rudy in 2001. In this process, the hippocampus helps distinguish between neural representations of similar yet distinct sensory experiences.

Additional findings in animal studies have shown that lesions in the auditory cortex (as demonstrated by Jarrell et al. (1987), Teich et al. (1988), and also Armony et al. (1997)) and the medial geniculate nucleus of the thalamus (as observed by Antunes and Moita (2010)) result in an increased tendency to generalise conditioned fear responses to auditory CSs. These findings suggest that sensory regions of the brain, where the stimulus characteristics of CS+ and GSs are represented and potentially distinguished by the hippocampus, contribute to this generalisation.

Furthermore, another brain area implicated in the generalisation of classical conditioning in animals is the ventromedial orbitofrontal cortex (OPFC), as discussed by Zelinski et al. (2010). Specifically, when rats have lesions in the OPFC, they generalise freezing behaviour from a context associated with a shock to an unrelated context. In contrast, intact animals only exhibit freezing behaviour in the context paired with the shock. These results imply that activations in the OPFC are necessary to inhibit fear responses to stimulus events resembling CS+, supporting the idea of an inverse relationship between OPFC activations and generalised conditioned fear responses to GSs.

APPENDIX D

A Brevia on Pragmatics: How Use Contributes to Meaning

Much work in semantics follows the tradition of positing systematic but inflexible theories of meaning. In practice, however, the meanings listeners derive from language are heavily dependent on nearly all aspects of context, both linguistic and situational.

In daily conversations, spoken sentences do not always mean what they literally mean. In many cases, language use cannot be handled without pragmatics.

The term pragmatics concerns the flexible use of language in context, deriving from the Greek noun *pragma*, which refers to an act or deed. Literally, pragmatics refers to aspects of linguistic meaning that derive from the act of speaking in a particular situated context.

An "in principle" distinction between semantics and pragmatics in addressing meaning is outlined in Figure D.1 (though, in many practical cases such distinction tends to blur)

There are three representative theories currently supported in pragmatics: (1) speech act theory by Austin (1962), (2) theory of implicature by Grice (1989), and (3) relevance theory by Sperber and Wilson (1986). These theories have provided many reasonable explanations, analyses, suggestions, and implications regarding language use, and have had a great influence on several academic disciplines.

By and large, these approaches point out that communication cannot be reduced to a code model: a communicator encodes her intended message into a signal, which is decoded by the audience using an identical copy of the code. They laid the foundations for an inferential model of communication (Grice, markedly), as an alternative to the classical code model.

According to the inferential model, a communicator provides evidence of her intention to convey a certain meaning, which is inferred by the audience on the basis of the

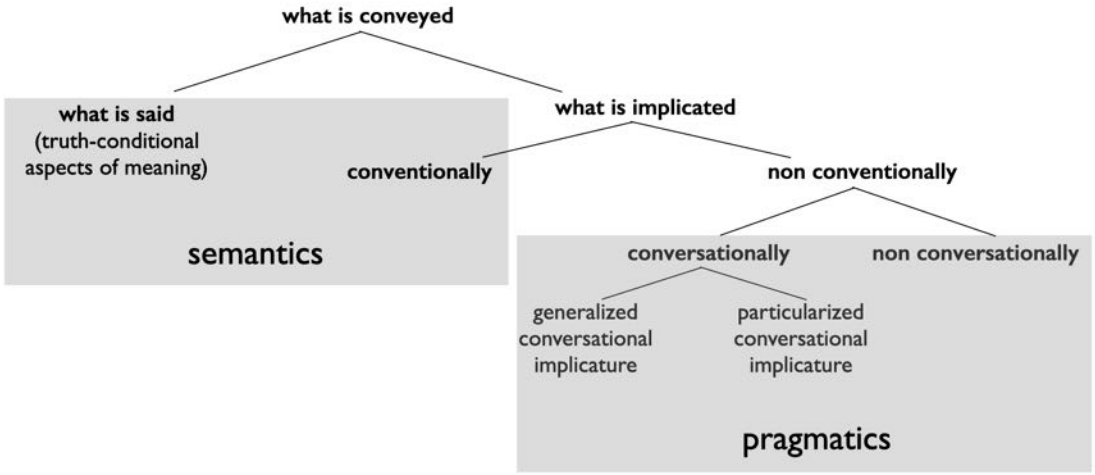


Figure D.1: *Kinds of meaning that are in principle addressed by semantics and pragmatics. These include both conventional and conversational (to be inferred) aspects of meaning. Adapted from Terkourafi (2021).*

evidence provided. An utterance is, obviously, a linguistically coded piece of evidence, so that verbal comprehension involves an element of decoding. However, the linguistic meaning recovered by decoding is just one of the inputs to a non-demonstrative inference process which yields an interpretation of the speaker's meaning (Sperber and Wilson, 1986).

D.0.1 Austin: speech acts

Half a century ago, Austin presented a new picture of analysing meaning; meaning is described in a relation among linguistic conventions correlated with words/sentences, the situation where the speaker actually says something to the listener, and associated intentions of the speaker. The idea that meaning exists among these relations is depicted successfully by the concept of *acts*: in uttering a sentence, that is, in utilising linguistic conventions, the speaker with an associated intention performs a linguistic or speech act to the listener. Austin's analysis of meaning is unique in the sense that meaning is not explained through some forms of reduction. In reductive theories of meaning, complexities of meaning expressed by a sentence are reduced by a single criterion to something else, and this is claimed to be the process of explaining the meaning of the sentence.

As we have seen modern truth-conditional semanticists adopt the Russellian idea of explaining the meaning of a sentence and the Russellian/Tarskian idea of correlating a sentence, as its meaning, with a fact or state of affairs: to explain the meaning of a sentence is to specify its truth conditions, i.e., to give necessary and sufficient conditions for the truth of that sentence.

Austin warned against oversimplifying complexities of meaning and emphasised

the importance of describing the total speech act in the total speech situation in which the language users employ the language: the speaker utters a sentence and performs a speech act to the hearer.

The preliminary distinction in Austin's approach is between assertions or statements – to which Austin refers with the term *constatives* ("My daughter's name is Frauke") and utterances with which something is done; Austin refers to these utterances with the term *performatives*("I bet you six pence Fury will win the race"). The latter make the action performed by the speaker explicit: these sentences perform an act (betting) and they are neither true nor false. Clearly, performatives can go wrong (e.g., the bet after the race is over): in this situation, the performative utterance is in general "unhappy". Thus performatives have to meet the so-called *felicity conditions*

Austin's felicity conditions define the elements which structure the speech situation, in terms of which a purported act succeeds/fails (Austin, 1962):

- (*Conventionality*) (i) There must exist a conventional procedure having a certain conventional effect (uttering of certain words by a speaker in certain circumstances) (ii) The circumstances and persons must be appropriate, as specified in the procedure.

- (*Actuality*) The procedure must be executed by all participants both (i) correctly and (ii) completely.

- (*Intentionality*) Often (i) the persons must have the requisite thoughts, feelings and intentions, as specified in the procedure, and (ii) if consequent conduct is specified, then the relevant parties must so do.

Through a description of the success/failure of the speech act purported, which is explained as a violation/observation of the felicity conditions, Austin formulated a method to describe a sentence in terms of the speech situation where it is uttered: by means of associated linguistic conventions, the speaker, with an associated intention, actually performs an act to the hearer, which induces a certain response from the listener.

The next turn in Austin's approach relies on showing how the difference between constative and performative utterances is somehow artificial: the apparently constative "France is hexagonal" is neither true nor false, it is simply a rough geometrical description, which might hold in certain circumstances. Thus, he proposes a framework in terms of which all speech acts, i.e. constatives as well as performatives, can be described: more fundamentally they are both embedded into the act of "saying something", the locutionary act. The study of utterances is the study of locutions, or of full units of speech. Further, locutionary acts are also and at the same time illocutionary acts, i.e. acts of doing something in saying something (e.g., accusing, asking and answering questions, apologizing, blaming, informing, ordering, assuring, warning, announcing an intention, making an appointment). Illocutionary acts conform to conventions and have a certain conventional force. Eventually, Austin finally contrasts locutionary ("He said to me: kiss her!") and illocutionary acts ("He urged me to kiss her"). with 'perlocutionary' acts ("He got me to kiss her"), i.e. acts of doing something by saying something like persuading, alerting, convincing, deterring, surprising and getting somebody to do something. Perlocutionary acts produce effects upon the

feelings, thoughts or actions of the addressee(s) and thus have psychological and/or behavioural consequences.

Also, perlocutionary acts are causal. Interestingly, the response or sequel of perlocutionary acts can also be achieved by non-verbal means (intimidation may be achieved by waving a stick or pointing a gun). Contrary to illocutionary acts, perlocutionary acts are not conventional: effects of the speaker's perlocutionary acts may be intended by the speaker, but they may also be unintended. A perlocutionary act is performed whenever the speaker is (at least partially) responsible for some act or state of the listener.

It is worth recalling, at this point, some further analysis due to Austin's student John R. Searle who systematised and somewhat formalised Austin's ideas. For Searle, speaking is performing illocutionary acts in a rule-governed form of behaviour: acts have an effect on the hearer; the hearer understands the speaker's utterance.

According to this insight, a sentence has two parts: a proposition-indicating element and the function-indicating device which reveals what illocutionary force the utterance is to have and thus what illocutionary act the speaker is performing in the utterance of the sentence. These devices include – at least for English - word order, stress, intonation contour, punctuation, the mood of the verb, and finally a set of so-called performative verbs. Meaning is more than a matter of intention, it is also a matter of convention. Both the intentional and conventional aspects of illocutionary acts must be captured and especially the relationship between them. In the performance of an illocutionary act the speaker intends to produce a certain effect by means of getting the hearer to recognise his intention to produce this effect, and furthermore, if he is using words literally, he intends this recognition to be achieved in virtue of the fact that the rules for using the expressions he utters associate the expressions with the production of that effect. Indeed, according to Searle there are three principal dimensions of differences between speech acts: the illocutionary point, the direction of fit, and the expressed psychological states. For example, utterances "I suggest we go to the movies" and "I insist that we go to the movies" both have the same illocutionary point but are presented with different strengths.

Another important distinction Searle makes between direct and indirect speech acts. The utterance "Can you pass the salt?" has a specific meaning but that also means something else: is also a request addressed to the hearer that should make him pass the salt to the speaker. The sentence has an ulterior illocutionary point beyond the illocutionary point contained in the meaning *per se* of the sentence. Namely, in indirect speech acts we observe a difference between what is said and what is actually meant by the speaker. The listener must then follow some sort of cooperative principle in conversation that operates on both the speaker and the listener and makes the inference that the speaker wants him to pass the salt.

Indirect speech acts, including perlocutionary acts, are often subject to social and/ or linguistic convention, which has to be learned in order to participate adequately in a society. It's communicative function is to be derived by means of sensible social reasoning, as when the speaker utters "It's cold in here" hoping that the listener will take the hint and turn the heating up or shut the window.

The cooperative principle of conversation that was set by H. Paul Grice's theory of implicature and conversational maxims.

D.0.2 Grice: the inferential stance

Grice presented an initial framework theory for pragmatic reasoning, positing that speakers are taken to be cooperative, choosing their utterances to convey particular meanings. Gricean listeners then attempt to infer the speaker's intended communicative goal, working backward from the form of the utterance. This goal inference framework for communication has been immensely influential.

The central point, again, is that there is a difference between what is said and what is actually meant by the speaker: the listener has to make certain inferences to recognise and understand this actual meaning which is implicated by the speaker in what he or she said. said. The notion of a conversational implicature is that of a default inference, one that captures our intuitions about a preferred or normal interpretation' of a sentence, an utterance, a conversation or a text. In the dialog

A Will you go to Mark's PhD party?

B I have to prepare my inaugural lecture

speaker A will understand that speaker B implies with his or her answer (an indirect speech act) that he or she will not or cannot go to this party.

The core of Grice's proposal was a set of conversational maxims (informativeness, truthfulness, relevance and clarity). These are a set of principles/categories that ground Grice's Cooperative Principle:

Quantity Avoid obscurity: (i) Make your contribution as informative as is required (for the current purposes of the exchange); (ii) Do not make your contribution more informative than is required

Quality Try to make your contribution one that is true: (i) Do not say what you believe to be false; (ii) Do not say that for which you lack adequate evidence.

Relation Be relevant.

Manner Be perspicuous

For instance, in the conversation

A Marco doesn't seem to have a girlfriend these days.

B He has been paying a lot of visits to Rome lately

speaker B implicates that Marco has, or may have, a girlfriend in Rome.

Indeed, interesting cases are those where the maxims are violated: ironic statements, metaphors, and understatements (e.g., speaking about a drunken man who has broken all his furniture as if "he was a little intoxicated") all break the maxim of Quality. Figures of speech like irony, metaphor and understatement are paradigmatic examples requesting implicatures.

Implicatures are not fully determinable, that is to say there is no one-to-one linkage between the form of an implicature and its intended meaning. A sentence like "Marco is a machine" might mean that Marco is unemotional, a hard worker, or efficient, depending on the circumstances of the conversation and the common ground of speaker and listener. Further, implicatures might involve complex inferences from non strictly linguistic behavior, such as prosody or silence (Senft, 2014):

A Mrs. X is an old bag.

(silence)

B The weather has been quite delightful this summer, hasn't it?

Violation here concerns the maxim of Relation. The utterance of speaker A there is acknowledged via a moment of embarrassed silence. Then, speaker B utters about the weather, blatantly refusing to make what he or she says relevant to A's preceding remark. Here, speaker B implicates not only that A's remark should be ignored, but also that A has committed a social *faux pas* (Senft, 2014).

Indeed, implicatures, though being non-conventional and distinguishable from other deductive processes,according to Grice can be calculated. However, attempts to build on these ideas by providing a specific set of formal principles that allow the derivation of pragmatic inferences have met with difficulty. The Rational Speech Act (RSA) theory, which we shall address later on, is one example where this goal has been successfully addressed.

The relevance-theoretic account is based on another of Grice's central claims: that utterances automatically create expectations which guide the hearer towards the speaker's meaning.

D.0.3 Sperber and Wilson: Relevance Theory

The relevance-theoretic account is based on Grice's central claim: utterances automatically create expectations which guide the listener towards the speaker's meaning. Here, the aim is to explain in cognitively realistic terms what these expectations of relevance amount to, and how they might contribute to an empirically plausible account of comprehension (Sperber and Wilson, 1986). In relevance-theoretic terms, any external stimulus or internal representation which provides an input to cognitive processes may be relevant to an individual at some time.

Indeed, the search for relevance is a basic feature of human cognition. As a result of constant selection pressure towards increasing efficiency, the human cognitive system has developed in such a way that our perceptual mechanisms tend automatically to pick out potentially relevant stimuli, our memory retrieval mechanisms tend automatically to activate potentially relevant assumptions, and our inferential mechanisms tend spontaneously to process them in the most productive way.

A sight, a sound, an utterance, a memory is relevant to an individual when it connects with background information he has available to yield conclusions that matter to him. Importantly, an input is relevant to an individual when its processing in a context of available assumptions yields a *positive cognitive effect* (Sperber and Wilson, 1986).

In such endeavour, the most important type of cognitive effect achieved by processing an input in a context is a contextual implication, a conclusion deducible from the input and the context together, but from neither input nor context alone.

An important role is played by *ostensive-inferential* communication, which involves an extra layer of intention:

1. The informative intention: the intention to inform an audience of something.
2. The communicative intention: the intention to inform the audience of one's informative intention

A clear example is provided by Sperber and Wilson (1986). A person might leave her empty glass in the partner line of vision, intending him to notice that she might like another drink. This is not yet a case of inferential communication because, although she did intend to affect her partner thoughts in a certain way, she gave no evidence of her intention. However, instead of covertly leaving her glass in his line of sight, she might touch his arm and point to her empty glass, wave it at him, ostentatiously put it down in front of the partner, stare at it meaningfully, or just say "My glass is empty". An ostensive stimulus is designed to attract the audience's attention. Given the universal tendency to maximise relevance, an audience will only pay attention to a stimulus that seems relevant enough.

The Communicative Principle of Relevance and the notion of optimal relevance are the key to relevance-theoretic pragmatics.

It is in the communicator interest to make ostensive stimulus as easy as possible for the audience to understand, and to provide evidence not just for the cognitive effects the communicator aims to achieve in the audience but also for further cognitive effects which, by holding attention, will help the communicator to achieve the goal.

Implicatures to identify, illocutionary indeterminacies to resolve, metaphors and ironies to interpret require an appropriate set of contextual assumptions, which the listener must also supply. The Communicative Principle of Relevance and the definition of optimal relevance suggest a practical procedure for performing these subtasks and constructing a hypothesis about the speaker's meaning. The hearer should take the linguistically encoded sentence meaning; following a path of least effort, he should enrich it at the explicit level and complement it at the implicit level until the resulting interpretation meets his expectation of relevance.

In many non-verbal cases (e.g. pointing to one's empty glass, failing to respond to a question), use of an ostensive stimulus merely adds an extra layer of intention recognition to a basic layer of information that the audience might have picked up anyway.

Clearly, the range of meanings that can be non-verbally conveyed is necessarily limited by the range of concepts the communicator can evoke in the audience by drawing attention to observable features of the environment. Verbal communication can achieve a degree of explicitness not available in non-verbal communication. Yet, the relevance-theoretic comprehension procedure applies in the same way to the resolution of linguistic underdeterminacies at both explicit and implicit levels.

Most notable, for Relevance Theory comprehension is an on-line process, and hypotheses about explicatures, implicated premises and implicated conclusions are developed in parallel against a background of expectations (or anticipatory hypotheses) which may be revised or elaborated as the utterance unfolds.

Some utterances (technical instructions, for instance) achieve relevance by conveying a few strong implicatures. Other utterances achieve relevance by weakly suggesting a wide array of possible implications, each of which is a weak implicature of the utterance. This is typical of poetic uses of language, and has been discussed in relevance theory under the heading of poetic effect.

Meaning is thus recovered by a mixture of decoding and inference based on a variety of linguistic and non-linguistic clues: for example word order, mood indicators, tone of voice, facial expression (Sperber and Wilson, 1986)

More generally, on both Gricean and relevance-theoretic accounts, the interpretation of every utterance involves a high degree of metarepresentational capacity, since comprehension rests on the ability to attribute both informative and communicative intentions. For instance, there is evidence that irony involves a higher order of metarepresentational ability than metaphor. Higher order representational performance involves the ability to recognise that the speaker is thinking, not directly about a state of affairs in the world, but about another thought or utterance that she attributes to someone else. Experimental evidence from the literature on autism, child development and right hemisphere damage, has shown that the comprehension of irony correlates with second-order metarepresentational abilities, while the comprehension of metaphor requires only first-order abilities.

From a psychological perspective, this raises the question of how pragmatic abilities are acquired, and how they fit into the overall architecture of the mind. Relevance theory addresses the issue and in this sense it qualifies as a cognitive psychological theory.

Grice's analysis treats comprehension as a variety of the Theory of Mind (ToM) or mind-reading. However, there are different interpretations in the literature as to the mind-reading problem (see Goldman and Sripada, 2005 for a discussion).

Mind-reading is the capacity to identify the mental states of others, for example, their beliefs, desires, intentions, goals, experiences, sensations and also emotion states. One approach to mind-reading holds that mental-state attributors deploy a naive psychological theory to infer mental states in others from their behavior, the environment, and/or their other mental states. According to different versions of this "theory-theory" (TT), the naive psychological theory is either a component of an innate, dedicated module or is acquired. by domain-general learning. A second approach holds that people typically execute mind-reading by a different sort of process, a simulation process. Roughly, according to simulation theory (ST), an attributor arrives at a mental attribution by simulating, replicating, or reproducing in his own mind the same state as the target's, or by attempting to do so. For example, the attributor would pretend to be in initial states thought to correspond to those of the target, feeds these states into parts of his own cognitive equipment (e.g. a decision-making mechanism), which would operate on them to produce an output state that is imputed to the target. TT vs. ST is a longstanding controversy (Goldman and Sripada, 2005), though much recent neuroscientific work is quite receptive to simulationist ideas (Gallese, 2007; Rizzolatti and Sinigaglia, 2016). In recent years a number of researchers have moved away from pure forms of TT or ST in the direction of some sort of TT/ST hybrid (e.g., Adolphs, 2002).

In this respect, Sperber and Wilson (1986) depart from the classic TT account of Fodor, where mind-reading is due to a central thought process, with a sharp distinction between a relatively undifferentiated central processes supported by modular input processes modular view of the mind. However, they endorse even more modular accounts of inference, supported by special-purpose inferential procedures, attuned to the properties of this particular domain, such as the Eye Direction Detector and the Intentionality Detector (Baron-Cohen, 1997). Since, inferential comprehension typically involves several layers of metarepresentation, while in regular mind-reading a single level is generally enough, they argue that such discrepancy might be accounted for by a specialised sub-module dedicated to comprehension, which might have evolved within

the overall mind-reading module, with its own proprietary concepts and mechanisms. One example is language development. In their view, children come with a substantial innate endowment, so they do not have to learn what ostensive-inferential communication is, but come with a substantial innate endowment. However, two-year-old children fail on regular first-order false belief tasks, and have no chance to recognise and understand the peculiar multi-levelled representations involved in verbal comprehension. Along development and learning, a child with limited metarepresentational capacity might start out as a Naively Optimistic interpreter, who accepts the first interpretation he finds relevant enough regardless of whether it is one the speaker could plausibly have intended. Subsequently, a child might pass through different developmental stages: the Cautious Optimist, with enough metarepresentational capacity who can pass first-order false belief tasks; the the Sophisticated Understander endowed with the metarepresentational capacity to deal simultaneously with mismatches and deception.

Eventually, besides such specific assumptions, which are questionable or at best incomplete with respect to most recent advances in social neuroscience, we can state that Relevance Theory has the merit to qualify as a cognitive psychological theory, with experimentally testable predictions.

D.0.4 At the origins of the communication act

There are some lessons that we can learn from the above discussion. First, the ontology underlying natural language is not the one that underlies the standard modern approach to logic and its application to natural language.

Second, the communicative acts involve, whatever the actual mechanisms, an inference process in the listener(s) which yields an interpretation of the speaker's meaning (mind-reading). This relies on context, mutually assumed conceptual common ground (Clark and Brennan, 1991) and mutually assumed cooperative motives. Note that context and common ground are different concepts. If, for example, a speaker utters "It's there" while pointing to a bicycle in a car park (the context) and smiling, the listener might reach different conclusions whether both know (common ground) that the bike was stolen two days ago, or that the listener's ex-boyfriend owns a bike.

Third, as in the example above, the communicative act is based either on verbal utterances and non verbal signals such as prosody, gestures (pointing to the bike), facial expression and so on.

Fourth, and markedly in the Gricean and the Relevance Theory approaches, every speech act creates an accountability relation, a socially binding force, no matter how trivial or insignificant, between the speaker and the listener. As Seuren (2009) puts it:

> all speech acts ... are performative in that they create a socially binding relation or state of affairs ... The primary function of language is not communication, in the sense of a transfer of information about the world, but social binding, that is, the creation of specific interpersonal, socially binding relations with regard to the proposition expressed by an utterance or speech act. It will be clear that this kind of social binding is a central element in the social fabric that is a necessary requirement for human communities.

From a broader perspective, the key point here is that linguistic acts are social acts that one person, the speaker, intentionally directs to another, the listener, in order to

condition her attention and imagination in particular ways so that she will do, know, or feel what he wants her to.

Clearly, beyond common ground, these acts work only under the assumption that participants are both endowed with a psychological infrastructure of skills and motivations of shared intentionality evolved for facilitating interactions with others in collaborative activities. The speaker informs the listener of her ex-boyfriend's likely presence or the location of her stolen bicycle simply because the speaker surmises that the listener would want to know these things; in other terms the speaker acts on prosocial motivation (Tomasello, 2010).

Linguistic communication, is thus not any kind of object, formal or otherwise. It is a form of social action constituted by social conventions for achieving social ends, premised on at least some shared understandings and shared purposes among the communicating agents, which is recognised as shared intentionality. Shared intentionality or "we" intentionality (Searle et al., 1995) is what is necessary for engaging in uniquely human forms of collaborative activity in which a plural subject "we" is involved: joint goals, joint intentions, mutual knowledge, shared beliefs—all in the context of various cooperative motives. It is, in brief, the the cooperative infrastructure of human communication (Tomasello, 2010).

In these perspective, the discussion on pragmatics (and semantics) goes beyond language itself. The origins of such capabilities lies in the evolutionary process by which basic cognitive skills have developed phylogenetically, up to the point of enabling the creation of cultural products historically. This way, children are provided with the biological and cultural tools they need to develop ontogenetically, a process which culminates in the skills of linguistic communication.

The story of how it happened is long and complicated. Thus, for our purposes, it will suffice to draw on Tomasello (2010) account.

Based on a vast literature on developmental psychology and ethology, Tomasello (2010) characterises human cooperative communication as follows:

1. It emerged first in evolution (and the same holds in individual ontogeny) in the natural, spontaneous gestures of pointing and pantomiming.

2. It is crucially grounded in a psychological infrastructure of shared intentionality, which originated in support of collaborative activities. Intentionality rests on: (a) social-cognitive skills for creating with others joint intentions and joint attention (and other forms of common conceptual ground), and (b) prosocial motivations and norms for helping and sharing with others.

3. Conventional communication, as embodied in human language, is possible only when participants already possess: (a) natural gestures and their shared intentionality infrastructure, and (b) skills of cultural learning and imitation for creating and passing along jointly understood communicative conventions and constructions.

The main steps of such evolutionary development are outlined at a glance in Figure D.2.

The road to human cooperative communication begins with great ape intentional communication. Intentional signals allow communicators for attempting to influence

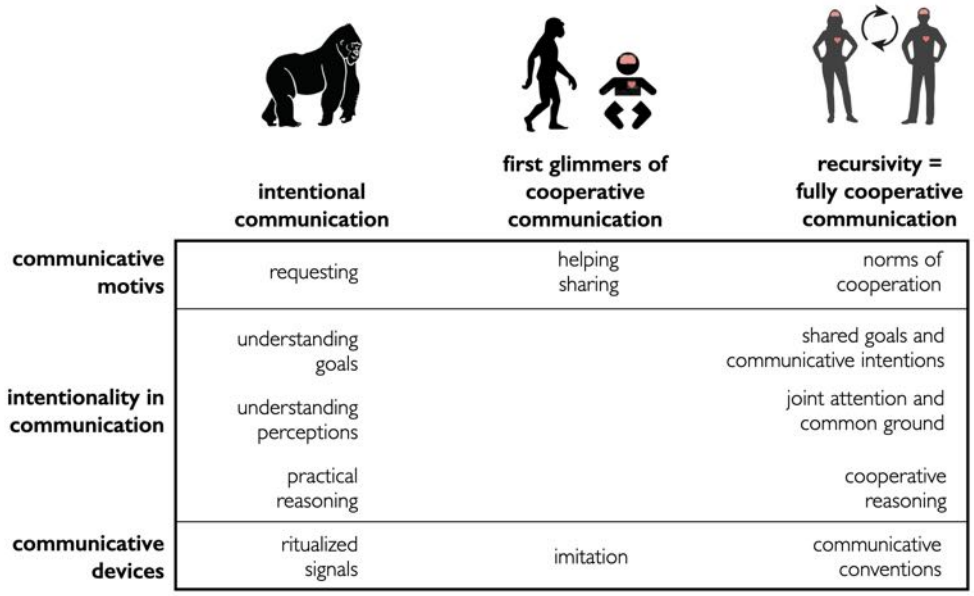


Figure D.2: *The psychological infrastructure of human cooperative communication represented both in terms of phylogeny and ontogeny development. First column: elements already present in great apes. Second column: the new human components. Third column: how the human version is transformed by recursivity. Adapted from Tomasello (2010).*

the behavior or psychological states of recipients intentionally. This is the starting point for communication from a psychological point of view. Non-human primates exhibit vocal displays (e.g., the "snake alarm call," in vervet monkeys), the capability to extract information from vocal calls, and even to learn during ontogeny to respond to novel calls. Yet, the repertoire is rather limited and linked to specific emotional episodes (human attempts to teach new vocalizations to monkeys and apes always fail). Vocal calls seem to be mainly individualistic expressions of emotions, not recipient-directed acts. Indeed, the production of a sound in the absence of the appropriate affective state (and related functional needs, escaping predators, surviving in fights, keeping contact with the group) seems to be an almost impossible task to learn. Also, calls are broadcasted to the group and they cannot be easily directed to selected individuals.

Gestures are the other form of ape intentional signals. Precisely, gesture designates a communicative behavior in the visual channel: mostly bodily postures, facial expressions, and manual gestures. Many of them are genetically fixed (displays), others are individually learned and flexibly used, especially in the great apes. There are two basic types of great ape gesture, based on how they function communicatively. The first are intention-movements (e.g., arm-raise to initiate play and touch-back by infants to moms to request being carried). These dyadic gestures and they are typically learned by imitation.

The second type concerns attention-getters (ground-slap, poke-at, throw-stuff, etc.) that are are used quite often by youngsters. In the prototypical case, the youngster is in a play mood—which is apparent from her mood-induced play face and posture display — and the attention-getter serves to draw attention to the display. In some cases, the communicator offers to another individual either a body part, typically for grooming,

or an object. In either cases, these might involve triadic intentional communication.

There are difference among apes too. Pollick and De Waal (2007) considered two captive bonobo groups, a total of 13 individuals, and two captive chimpanzee groups, a total of 34 individuals. The study distinguished 31 manual gestures and 18 facial/vocal signals. It was found that homologous facial/vocal displays were used very similarly by both ape species, yet the same did not apply to gestures. Both within and between species gesture usage varied enormously. Moreover, bonobos (the most emphatic ape species) showed greater flexibility in this regard than chimpanzees and were also the only species in which multi-modal communication (i.e., combinations of gestures and facial/vocal signals) added to behavioral impact on the recipient.

In gestures, great apes reveal social intention and some basic referential intention, and most important they mark the capability of paying attention to the attention of others. Further, apes raised in rich human contexts, similar to the way human children are raised, have been observed to request things imperatively by pointing (e.g., pointing to a locked door when they want access behind it, so that the human will open it for them). Tomasello argues that human-raised apes have a fairly flexible understanding that humans control many aspects of their world, and that these humans can be induced to do things that help them reach their goals in this human environment with some kind of attention-directing behavior. Interestingly, apes point for humans, but not for one another.

To conclude, a large body of research has demonstrated that great apes understand much about how others work as intentional, perceiving agents. Specifically, great apes understand something of the goals and perceptions of others and how these work together in individual intentional action in ways very similar to young human children (cfr. Figure D.2).

There is some debate on whether apes do have a true ToM (the ability to recognise the mental states of others). For instance, De Waal et al. (2006) are convinced that apes take one another's perspective, and that the evolutionary origin of this ability is not to be sought in social competition, even if it is readily applied in this domain but in the need for cooperation. At the core of perspective-taking is emotional linkage between individuals—widespread in social mammals—upon which evolution (or development) builds ever more complex manifestations, including appraisal of another's knowledge and intentions.

In any case, apes and young human children both understand in the same basic way (in simple situations) that individuals pursue a goal in a persistent manner until they have reached it—and they understand the goal not as the result produced in the external environment, but rather as the actor's internal representation of the state of the world she wishes to bring about.

These primitive codes provides the means for establishing language. Indeed, If we want to understand human communication, we cannot begin with language. Rather, we must begin with unconventionalised, uncoded communication, and other forms of mental attunement, as foundational. Candidates for this role are natural gestures such as pointing and pantomiming. Human gestures, in fact: direct the attention of a recipient spatially to something in the immediate perceptual environment (deictically); direct the imagination of a recipient to something that, typically, is not in the immediate perceptual environment by behaviorally simulating an action, relation, or object

(iconically)

However, human cooperative communication is more complex than ape intentional communication because its underlying social-cognitive infrastructure comprises not only skills for understanding individual intentionality but also skills and motivations for shared intentionality.

At some point basic signalling integrates in more complex processes that allow human beings for being able to communicate with one another: human beings cooperate with one another in species-unique ways involving processes of shared intentionality. As said, the latter denotes behavioral phenomena that are both intentional and irreducibly social, in the sense that the agent of the intentions and actions is the plural subject "we" (Searle et al., 1995).

Shared intentionality, when employed in certain social interactions, generates joint goals and joint attention, which provide the common conceptual ground within which human communication most naturally occurs.

The basic cognitive skill of shared intentionality is recursive mindreading and its basic motives are helping and sharing. When employed in interactions, these generate the three basic motives of human cooperative communication: requesting (requesting help), informing (offering help in the form of useful information), and sharing emotions and attitudes (bonding socially by expanding common ground).

Such phylogenetical path is recapitulated in the ontogeny of human infants' gestural communication, especially pointing, which provides evidence for the various components of the hypothesised cooperative infrastructure and a connection to shared intentionality. All these components must be present for onset of language acquisition.

Infants' iconic gestures emerge on the heels of their first pointing, requiring a communicative intention to be effective (otherwise they are just empty actions). Iconic gestures represent symbolic ways of indicating referents. They are promptly replaced by conventional language, while basic pointing is not displaced by the emergence of language.

The ontogenetic transition from gestures to conventional forms of communication, including language, also relies crucially on the shared intentionality infrastructure — especially joint attention in collaborative activities — to create the common ground necessary for learning "arbitrary" communicative conventions.

The ontogenetic transition from gestures to language demonstrates the common function of (i) pointing and demonstratives (e.g., this and that); and (ii) iconic gestures and content words (e.g., nouns and verbs).

To sum up, sharing emotions and attitudes with others may have arisen as ways of social bonding and expanding common ground within the social group (tied to cultural group selection)— with the actual norms that govern cooperative communication originating from group sanctions for not cooperating.

D.0.5 State of the art in computational pragmatics

Computational pragmatics is a branch of computational linguistics, located between computer science and technology and pragmatics in theoretical linguistics. Bunt and Black (2000) introduced computational pragmatics as the study of the relationship between utterances and contextual information from a computational standpoint, by relying on abduction, belief and context. Jurafsky (2004) points out, in the same vein of

Bunt and Black (2000), that the basic problems can be cast as an inference task, one of somehow filling in information that isn't actually present in the utterance at hand. In this effort, one approach is inferential models are based on belief logics and use logical inference to reason about the speaker's intentions (for example BDI - belief, desire, and intention or o plan-based model, proposed in AI by Allen (1995)). A second approach, is represented by cue-based models thinking of the surface form of the sentence as a set of cues to the speaker's intentions. The cue-based models tend to be probabilistic machine learning models, in particular Stolcke et al. (2000) proposed a Hidden Markov Model for the purpose of decoding dialogue acts. They see interpretation as a classification task, and solve it by training statistical classifiers on labelled examples of speech acts. Despite their differences, these models have in common the use of a kind of abductive inference (Jurafsky, 2004).

Following the statistical strand, more recent developments are oriented towards the solution of problems in a bewildering variety of application fields by exploiting machine learning techniques and in particular deep learning techniques. For instance, due to the extensive use of slangs, bashes, flames, and non-literal texts, tweets are a great source of figurative language, such as sarcasm, irony, metaphor, simile, hyperbole, humor, and satire (Abulaish et al., 2020). Another area of interest is that of Conversational recommender systems (CRS) that are conceived support a richer set of interactions than classic RS. These interactions can, for example, help to improve the preference elicitation process or allow the user to ask questions about the recommendations and to give feedback (Jannach et al., 2021). Automatic generation of stories with minimum effort and customization of stories for the users' education and entertainment needs (Alhussain and Azmi, 2021) is also an active field of investigation. Specific pragmatic problems are of current interest such as sarcasm detection (Joshi et al., 2017), metaphor detection (Rai and Chakraverty, 2020) or hate in speech (Fortuna and Nunes, 2018). A wide panorama of approaches are used, ranging from a hand-coded rule system to more recent deep learning techniques. The latter basically rely on the neural approaches developed in distributional semantics that we have previously touched on (for instance, by extending word embeddings to sentence embeddings). These approaches are in fact often characterised as "neural text generation" (Clark et al., 2018), "neural metaphor processing" (Tong et al., 2021) and so on.

Also, a great deal of such approaches, markedly for solving detection problems, rely on the classic pattern recognition paradigm (feature extraction $\rightarrow$ classification) which is modernly declined in the end-to-end training or fine-tuning of deep nets. This is somehow a consequence of the fact that like much of NLP, distributional semantics is largely bottom-up: the goals are usually to improve performance on particular tasks, or particular datasets. Yet, when contrasted against the truth-conditional approaches which are largely top-down, where the goal is known, one has to admit that those theories haven't reached the goal. But for an enlightening and critical discussion of the field, the reader might refer to Emerson (2020).

On the other hand, these approaches witness the probabilistic turn in semantics and pragmatics (Erk, 2021). Word and sentence meanings as fluid and flexible. Probabilistic and graded approaches can then be used to describe similarities between meanings, as well as degrees of influence of context on sense choice. Beliefs, and preferences of speakers and listeners are often best described in graded or probabilistic terms.

Also, it is worth noting that neural models, currently the most widely used form of machine learning model, used to be considered a framework distinct from and incompatible with Bayesian models, but the boundary has been blurred on both theoretical and practical levels (Goodfellow et al., 2016).

A prominent example is the Rational Speech Act (RSA) model (Goodman and Frank, 2016).

RSA is an agent-based approach to formalizing pragmatic reasoning. Listeners are modeled as reasoning recursively about the goals of speakers, and vice versa. Although the framework is explicitly designed to capture the back-and-forth of Gricean reasoning, it is consistent with much newer theorizing as well (for example, it explicitly incorporates a relevance distribution over possible messages).

The basic architecture is the following. The task of the listener L is to estimate the probability of a particular intended message m given the observed utterance u by the speaker, which we notate $P_L(m \mid u)$. Here, the m conveys information over the states of affairs (generically, the "world") as conceptualised by the speaker. By convention, the utterance u comprises linguistic as well as nonlinguistic components.

The listener is assumed to compute the posterior probability P_L via Bayesian inference through the integration of two components, the likelihood of the utterance given the message and the prior probability of the message:

$$P_L(m \mid u) \propto P_S(u \mid m)P(m).$$

The characteristic feature of RSA is the way that the likelihood term P_S (representing the speaker) is computed. The listener L is assumed to have an internal model of the speaker S, who is modeled as choosing their utterance by maximizing their own utility $U_S(u; m)$:

$$P_S(u \mid m) \propto \exp \alpha U_S(u; m).$$

The scalar value α can be interpreted as an indicator of how rational the speaker is in choosing utterances (i.e., how strongly they prefer the higher utility option). The speaker's utility is higher the more information they transmit through their utterance. Utility maximization through cooperative communication reflects the central idea that humans communicate in a relevant (Sperber and Wilson, 1986) and cooperative (Clark and Brennan, 1991; Grice, 1989; Tomasello, 2010) way.

The utility of an utterance in turn depends on how much epistemic certainty it provides to the listener:

$$U_S(u; m) = \log P_{Lit}(m \mid u)$$

To avoid infinite recursion, the listener is taken to be a literal listener, say P_{Lit} who interprets utterances in accordance with their literal semantics:

$$P_{Lit}(m \mid u) \propto \delta_{[[u]](m)} P(m).$$

Here, $[[u]]$ is a semantic denotation for each sentence, concerning whether or not the utterance is true of a given message. $P(m)$ is the prior probability of the conveyed message. This prior term can be considered a distribution over relevant messages in

context: it represents evidence for or against a particular message, independent of the utterance.

Through this recursive reference back to a listener, the model captures the interdependence of speaker and listener in communicative interactions. The combination of these two terms — speaker likelihood and prior — the listener's belief represents the outcome of a social-cognitive inference about the likely intended meaning of an utterance in context.

The RSA framework, which builds upon and synthesises a number of formal traditions in the study of human inference, from game theory to models of human reasoning it is well suited for our purposes. It is a description of the computational problem being solved by agents rather than being a model of a psychological process.

Also, it is suitable to capture and formalise most relevant inferential theories of pragmatics that we have discussed in this Appendix (Grice, 1989; Clark and Brennan, 1991; Sperber and Wilson, 1986; Tomasello, 2010). These theories have been immensely influential, but they are verbal descriptions of the psychological processes involved in communication, and the actual computations that lead to inference are not further specified.

RSA and its variants have now been used successfully to describe and predict a wide variety of phenomena, including implicature (Goodman and Stuhlmüller, 2013), hyperbole (Kao et al., 2014), vagueness (Lassiter and Goodman, 2017), generic language (Tessler and Goodman, 2019), and politeness (Yoon et al., 2020).

Importantly, RSA can offer a pragmatic perspective on language development. Bohn and Frank (2019) have argued for pragmatic reasoning supporting children's learning, comprehension, and use of language, providing evidence for developmental continuity between early nonverbal communication, language learning, and linguistic pragmatics.

Bibliography

Abulaish, M., Kamal, A., and Zaki, M. J. (2020). A survey of figurative language and its computational detection in online social networks. *ACM Transactions on the Web (TWEB)*, 14(1):1–52.

Adolphs, R. (2002). Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behavioral and cognitive neuroscience reviews*, 1(1):21–62.

Adolphs, R. (2013). The biology of fear. *Current biology*, 23(2):R79–R93.

Ahrens, L. M., Pauli, P., Reif, A., Mühlberger, A., Langs, G., Aalderink, T., and Wieser, M. J. (2016). Fear conditioning and stimulus generalization in patients with social anxiety disorder. *Journal of Anxiety Disorders*, 44:36–46.

Aitchison, L. and Lengyel, M. (2017). With or without you: predictive coding and bayesian inference in the brain. *Current opinion in neurobiology*, 46:219–227.

Alhussain, A. I. and Azmi, A. M. (2021). Automatic story generation: A survey of approaches. *ACM Computing Surveys (CSUR)*, 54(5):1–38.

Allen, J. (1995). Natural language understanding.

Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al. (2016). Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2):20.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, 98(3):409.

Andrews-Hanna, J. R., Smallwood, J., and Spreng, R. N. (2014). The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. *Annals of the new York Academy of Sciences*, 1316(1):29–52.

Antunes, R. and Moita, M. A. (2010). Discriminative auditory fear learning requires both tuned and nontuned auditory pathways to the amygdala. *Journal of Neuroscience*, 30(29):9782–9787.

Apkarian, A. V., Bushnell, M. C., Treede, R.-D., and Zubieta, J.-K. (2005). Human brain mechanisms of pain perception and regulation in health and disease. *European journal of pain*, 9(4):463–484.

Appelhans, B. M. and Luecken, L. J. (2008). Heart rate variability and pain: associations of two interrelated homeostatic processes. *Biological psychology*, 77(2):174–182.

Argüello, E. J., Silva, R. J., Huerta, M. K., and Avila, R. S. (2015). Computational modeling of peripheral pain: a commentary. *Biomedical engineering online*, 14(1):1–8.

Armony, J. L., Servan-Schreiber, D., Romanski, L. M., Cohen, J. D., and LeDoux, J. E. (1997). Stimulus generalization of fear responses: effects of auditory cortex lesions in a computational model and in rats. *Cerebral Cortex (New York, NY: 1991)*, 7(2):157–165.

Asmundson, G. J. and Katz, J. (2009). Understanding the co-occurrence of anxiety disorders and chronic pain: state-of-the-art. *Depression and anxiety*, 26(10):888–901.

Asmundson, G. J. and Wright, K. D. (2004). Biopsychosocial approaches to pain. In *Pain*, pages 35–57. Psychology Press.

Austin, J. L. (1962). *How to do things with words*. Harvard University Press, Cambridge, MA.

Baetu, T. M. (2020). Pain eliminativism. *Journal of Mental Health & Clinical Psychology*, 4(3).

Bain, D. (2003). Intentionalism and pain. *The Philosophical Quarterly*, 53(213):502–523.

Baliki, M. N., Geha, P. Y., Fields, H. L., and Apkarian, A. V. (2010). Predicting value of pain and analgesia: nucleus accumbens response to noxious stimuli changes in the presence of chronic pain. *Neuron*, 66(1):149–160.

Bandt, C. and Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102.

Bargary, G., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Hogg, R. E., and Mollon, J. D. (2017). Individual differences in human eye movements: An oculomotor signature? *Vision research*, 141:157–169.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. MIT press.

Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46.

Barrett, L. F. (2014). The conceptual act theory: A précis. *Emotion review*, 6(4):292–297.

Barrett, L. F. (2016). Navigating the science of emotion. In Meiselman, H. L., editor, *Emotion Measurement*, pages 31–63. Woodhead Publishing.

Barrett, L. F. (2017a). Categories and their role in the science of emotion. *Psychological inquiry*, 28(1):20–26.

Barrett, L. F. (2017b). *How emotions are made: The secret life of the brain*. Pan Macmillan.

Barrett, L. F. and Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in experimental social psychology*, 41:167–218.

Barrett, L. F., Lewis, M., and Haviland-Jones, J. M. (2016). *Handbook of emotions*. Guilford Publications.

Barrett, L. F. and Satpute, A. B. (2019). Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience letters*, 693:9–18.

Barrett, L. F., Wilson-Mendenhall, C. D., and Barsalou, L. W. (2015). The conceptual act theory: A roadmap.

Bartumeus, F. and Catalan, J. (2009). Optimal search behavior and classic foraging theory. *Journal of Physics A: Mathematical and Theoretical*, 42:434002.

Bella-Fernández, M., Suero Suñé, M., and Gil-Gómez de Liaño, B. (2022). Foraging behavior in visual search: A review of theoretical and mathematical models in humans and animals. *Psychological research*, 86(2):331–349.

Bengio, Y. and Frasconi, P. (1996). Input-output hmms for sequence processing. *IEEE Transactions on Neural Networks*, 7(5):1231–1249.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, 20(28):1–6.

Birdwell, B. G., Herbers, J. E., and Kroenke, K. (1993). Evaluating chest pain: the patient's presentation style alters the physician's diagnostic approach. *Archives of Internal Medicine*, 153(17):1991–1995.

Blanchard, D. C. and Blanchard, R. J. (2008). Defensive behaviors, fear, and anxiety. *Handbook of behavioral neuroscience*, 17:63–79.

Boccignone, G., Conte, D., Cuculo, V., D'Amelio, A., Grossi, G., and Lanzarotti, R. (2018). Deep construction of an affective latent space via multimodal enactment. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):865–880.

Boccignone, G. and Cordeschi, R. (2015). Coping with levels of explanation in the behavioral sciences.

Boccignone, G., Cuculo, V., D'Amelio, A., Grossi, G., and Lanzarotti, R. (2019). Give ear to my face: Modelling multimodal attention to social interactions. In Leal-Taixé, L. and Roth, S., editors, *Computer Vision – ECCV 2018 Workshops*, pages 331–345. Springer International Publishing, Cham.

Boccignone, G., Cuculo, V., D'Amelio, A., Grossi, G., and Lanzarotti, R. (2020). On gaze deployment to audio-visual cues of social interactions. *IEEE Access*, 8:161630–161654.

Boccignone, G., De'Sperati, C., Granato, M., Grossi, G., Lanzarotti, R., Noceti, N., Odone, F., et al. (2020). Stairway to elders: bridging space, time and emotions in their social environment for well-being. In *ICPRAM 2020-Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, pages 548–554. SciTePress.

Bohn, M. and Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, 1:223–249.

Borsook, D., Edwards, R., Elman, I., Becerra, L., and Levine, J. (2013). Pain and analgesia: the value of salience circuits. *Progress in neurobiology*, 104:93–105.

Brischoux, F., Chakraborty, S., Brierley, D. I., and Ungless, M. A. (2009). Phasic excitation of dopamine neurons in ventral vta by noxious stimuli. *Proceedings of the national academy of sciences*, 106(12):4894–4899.

Britton, N. and Skevington, S. M. (1989). A mathematical model of the gate control theory of pain. *Journal of theoretical biology*, 137(1):91–105.

Britton, N. F. and Skevington, S. M. (1996). On the mathematical modelling of pain. *Neurochemical research*, 21:1133–1140.

Brooks, J. and Tracey, I. (2005). From nociception to pain perception: imaging the spinal and supraspinal pathways. *Journal of anatomy*, 207(1):19–33.

Brooks, J. and Tracey, I. (2007). The insula: a multidimensional integration site for pain.

Brunswik, E. (1952). The conceptual framework of psychology. *(No Title)*.

Bucci, D. J., Saddoris, M. P., and Burwell, R. D. (2002). Contextual fear discrimination is impaired by damage to the postrhinal or perirhinal cortex. *Behavioral neuroscience*, 116(3):479.

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the new York Academy of Sciences*, 1124(1):1–38.

Budaev, S., Jørgensen, C., Mangel, M., Eliassen, S., and Giske, J. (2019). Decision-making from the animal perspective: bridging ecology and subjective cognition. *Frontiers in Ecology and Evolution*, 7:164.

Buhle, J. T., Silvers, J. A., Wager, T. D., Lopez, R., Onyemekwu, C., Kober, H., Weber, J., and Ochsner, K. N. (2014). Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. *Cerebral cortex*, 24(11):2981–2990.

Bunt, H. and Black, B. (2000). The abc of computational pragmatics. *Abduction, Belief, and Context in Dialogue: Studies in Computational Pragmatics*, 1:1.

Burke, K. (1966). *Language as symbolic action: Essays on life, literature, and method*. Univ of California Press.

Bushnell, M. C., Čeko, M., and Low, L. A. (2013). Cognitive and emotional control of pain and its disruption in chronic pain. *Nature Reviews Neuroscience*, 14(7):502–511.

Cain, M. S., Vul, E., Clark, K., and Mitroff, S. R. (2012). A bayesian optimal foraging model of human visual search. *Psychological science*, 23(9):1047–1054.

Calvillo, E. R. and Flaskerud, J. H. (1993). Evaluation of the pain response by mexican american and anglo american women and their nurses. *Journal of Advanced Nursing*, 18(3):451–459.

Calvo, R. A. 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.

Cano, A., Barterian, J. A., and Heller, J. B. (2008). Empathic and nonempathic interaction in chronic pain couples. *The Clinical journal of pain*, 24(8):678.

Carew, T. J., Walters, E. T., and Kandel, E. R. (1981). Associative learning in aplysia: Cellular correlates supporting a conditioned fear hypothesis. *Science*, 211(4481):501–504.

Carreiras, C., Alves, A. P., Lourenço, A., Canento, F., Silva, H., Fred, A., et al. (2015). Biosppy: Biosignal processing in python. *Accessed on*, 3(28):2018.

Cauda, F., D'Agata, F., Sacco, K., Duca, S., Cocito, D., Paolasso, I., Isoardo, G., and Geminiani, G. (2010). Altered resting state attentional networks in diabetic neuropathic pain. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(7):806–811.

Cavada, C., Compañy, T., Tejedor, J., Cruz-Rizzolo, R. J., and Reinoso-Suárez, F. (2000). The anatomical connections of the macaque monkey orbitofrontal cortex. a review. *Cerebral cortex*, 10(3):220–242.

Cerf, M., Harel, J., Einhäuser, W., and Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. *Advances in neural information processing systems*, 20.

Chambers, C. T., Hardial, J., Craig, K. D., Montgomery, C., et al. (2005). Faces scales for the measurement of postoperative pain intensity in children following minor surgery. *The Clinical journal of pain*, 21(3):277–285.

Chapman, C. R. and Nakamura, Y. (1999). A passion of the soul: an introduction to pain for consciousness researchers. *Consciousness and Cognition*, 8(4):391–422.

Chapman, C. R., Oka, S., Bradshaw, D. H., Jacobson, R. C., and Donaldson, G. W. (1999). Phasic pupil dilation response to noxious stimulation in normal volunteers: relationship to brain evoked potentials and pain report. *Psychophysiology*, 36(1):44–52.

Chen, Z. S. and Wang, J. (2023). Pain, from perception to action: A computational perspective. *Iscience*, 26(1).

Cheng, Y., Lin, C.-P., Liu, H.-L., Hsu, Y.-Y., Lim, K.-E., Hung, D., and Decety, J. (2007). Expertise modulates the perception of pain in others. *Current Biology*, 17(19):1708–1713.

Choinière, M., Melzack, R., Girard, N., Rondeau, J., and Paquin, M.-J. (1990). Comparisons between patients' and nurses' assessment of pain and medication efficacy in severe burn injuries. *Pain*, 40(2):143–152.

Chuk, T., Chan, A. B., and Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden markov models. *Journal of vision*, 14(11):8–8.

Churchland, P. M. and Churchland, P. S. (1992). Intertheoretic reduction: A neuroscientist's field guide. In *Neurophilosophy and alzheimer's disease*, pages 18–29. Springer.

Clark, E., Ji, Y., and Smith, N. A. (2018). Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260.

Clark, H. and Brennan, S. (1991). Grounding in communication', 127-149 in resnick lb, levine jm and teasley sd. In Resnick, L., B., L., John, M., Teasley, S., and D., editors, *Perspectives on Socially Shared Cognition*, pages 259–292. American Psychological Association.

Cooper, W. E. and Blumstein, D. T. (2015). *Escaping from predators: an integrative view of escape decisions*. Cambridge University Press.

Corcoran, A. W., Pezzulo, G., and Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, 35(3):1–45.

Coutrot, A. and Guyader, N. (2017). Learning a time-dependent master saliency map from eye-tracking data in videos. *arXiv preprint arXiv:1702.00714*.

Coviello, E., Chan, A. B., and Lanckriet, G. R. (2014). Clustering hidden markov models with variational hem. *The Journal of Machine Learning Research*, 15(1):697–747.

Craig, A., Bushnell, M., Zhang, E.-T., and Blomqvist, A. (1994). A thalamic nucleus specific for pain and temperature sensation. *Nature*, 372(6508):770–773.

Craig, A. D. (2003). Interoception: the sense of the physiological condition of the body. *Current opinion in neurobiology*, 13(4):500–505.

Craig, K. D. (2004). Social communication of pain enhances protective functions: a comment on deyo, prkachin and mercer (2004).

Craig, K. D. (2007). *Credibility, Assessment*, pages 491–493. Springer Berlin Heidelberg, Berlin, Heidelberg.

Craig, K. D. (2015). Social communication model of pain. *Pain*, 156(7):1198–1199.

Craig, K. D. and Patrick, C. J. (1985). Facial expression during induced pain. *Journal of personality and social psychology*, 48(4):1080.

Craig, K. D. and Versloot, J. (2022). Psychosocial perspectives on chronic pain. *Clinical pain management: A practical guide*, pages 40–49.

Cuculo, V., D'Amelio, A., Lanzarotti, R., and Boccignone, G. (2018). Personality gaze patterns unveiled via automatic relevance determination. In *Federation of International Conferences on Software Technologies: Applications and Foundations*, pages 171–184. Springer.

Cuculo, V. and D'Amelio, A. (2019). Openfacs: an open source facs-based 3d face animation system. In *Image and Graphics: 10th International Conference, ICIG 2019, Beijing, China, August 23–25, 2019, Proceedings, Part II 10*, pages 232–242. Springer.

Cutter, B. and Tye, M. (2011). Tracking representationalism and the painfulness of pain. *Philosophical Issues*, 21:90–109.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., and Friston, K. (2020). Active inference on discrete state-spaces: a synthesis. *Journal of Mathematical Psychology*, 99:102447.

D'Amelio, A. and Boccignone, G. (2021). Gazing at social interactions between foraging and decision theory. *Frontiers in neurorobotics*, 15:31.

Damien, J., Colloca, L., Bellei-Rodriguez, C.-É., and Marchand, S. (2018). Pain modulation: from conditioned pain modulation to placebo and nocebo effects in experimental and clinical pain. *International review of neurobiology*, 139:255–296.

De Waal, F., Macedo, S. E., and Ober, J. E. (2006). *Primates and philosophers: How morality evolved.* Princeton University Press.

Dennett, D. (1987). *The Intentional Stance*. MIT Press, Cambridge, MA.

Dennett, D. C. (1978). Why you can't make a computer that feels pain. *Synthese*, 38(3):415–456.

Descartes, R. (1969). *De homine*, volume 6. Hack.

Dorr, M., Martinetz, T., Gegenfurtner, K., and Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10).

Downar, J., Crawley, A. P., Mikulis, D. J., and Davis, K. D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. *Nature neuroscience*, 3(3):277–283.

Duck, S. and McMahan, D. T. (2010). *Communication in everyday life*. Sage Publications.

Dudley, S. R. and Holm, K. (1984). Assessment of the pain experience in relation to selected nurse characteristics. *Pain*, 18(2):179–186.

Dunsmoor, J. E. and Paz, R. (2015). Fear generalization and anxiety: behavioral and neural mechanisms. *Biological psychiatry*, 78(5):336–343.

Dworkin, R. H. (1994). Pain insensitivity in schizophrenia: a neglected phenomenon and some implications. *Schizophrenia bulletin*, 20(2):235–248.

Dymond, S., Dunsmoor, J. E., Vervliet, B., Roche, B., and Hermans, D. (2015). Fear generalization in humans: systematic review and implications for anxiety disorder research. *Behavior therapy*, 46(5):561–582.

D'Amelio, A., Patania, S., Bursic, S., Cuculo, V., and Boccignone, G. (2023). Using gaze for behavioural biometrics. *Sensors*, 23(3):1262.

Eccleston, C. and Crombez, G. (1999). Pain demands attention: A cognitive–affective model of the interruptive function of pain. *Psychological bulletin*, 125(3):356.

Eck, J., Richter, M., Straube, T., Miltner, W. H., and Weiss, T. (2011). Affective brain regions are activated during the processing of pain-related words in migraine patients. *Pain*, 152(5):1104–1113.

Eckert, A.-L., Pabst, K., and Endres, D. M. (2022). A bayesian model for chronic pain. *Frontiers in Pain Research*, 3:966034.

Ehinger, K. A. and Wolfe, J. M. (2016). When is it time to move to the next map? optimal foraging in guided visual search. *Attention, Perception, & Psychophysics*, 78(7):2135–2151.

Eisenberger, N. I., Lieberman, M. D., and Williams, K. D. (2003). Does rejection hurt? an fmri study of social exclusion. *Science*, 302(5643):290–292.

Ekman, P., Sorenson, E. R., and Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88.

Elzinga, B. M. and Bremner, J. D. (2002). Are the neural substrates of memory the final common pathway in posttraumatic stress disorder (ptsd)? *Journal of affective disorders*, 70(1):1–17.

Emerson, G. (2020). What are the goals of distributional semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7436–7453.

Engel, G. L. (1977). The need for a new medical model: a challenge for biomedicine. *Science*, 196(4286):129–136.

Erk, K. (2021). The probabilistic turn in semantics and pragmatics. *Annual Review of Linguistics*, 8.

Fanselow, M. S. (1994). Neural organization of the defensive behavior system responsible for fear. *Psychonomic bulletin & review*, 1(4):429–438.

Fields, H. L. (2006). A motivation-decision model of pain: the role of opioids. In *Proceedings of the 11th world congress on pain*, pages 449–459. IASP Press, WA, USA.

Fishbain, D. A. (1982). Pain insensitivity in psychosis. *Annals of emergency medicine*, 11(11):630–632.

Fishbain, D. A., Cutler, R., Rosomoff, H. L., and Rosomoff, R. S. (1997). Chronic pain-associated depression: antecedent or consequence of chronic pain? a review. *The Clinical journal of pain*, 13(2):116–137.

Fordyce, W. E. (1996). Response to thompson/merskey/teasell. *Pain*, 65(1):112–114.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Frey, M. v. (1895). Beitrage zur sinnesphysiologie der haut. *Saechsischen Akademie der Wissenschaften zu Leipzig. Math-Phys Cl*, 47:166–184.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural computation*, 29(1):1–49.

Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences*, 364(1521):1211–1221.

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E., and Bestmann, S. (2012). Dopamine, affordance and active inference. *PLoS computational biology*, 8(1):e1002327.

Gallese, V. (2007). Embodied simulation: from mirror neuron systems to interpersonal relations. In *Novartis Found Symp*, volume 278, pages 3–12.

Gatchel, R. J., Peng, Y. B., Peters, M. L., Fuchs, P. N., and Turk, D. C. (2007). The biopsychosocial approach to chronic pain: scientific advances and future directions. *Psychological bulletin*, 133(4):581.

Gerlee, P. and Lundh, T. (2016). *Scientific models: red atoms, white lies and black boxes in a yellow book*. Springer.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459.

Ghirlanda, S. and Enquist, M. (2003). A century of generalization. *Animal Behaviour*, 66(1):15–36.

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.

Goodman, N. D. (2013). The principles and practice of probabilistic programming. *ACM SIGPLAN Notices*, 48(1):399–402.

Goodman, N. D. and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184.

Goubert, L., Craig, K. D., Vervoort, T., Morley, S., Sullivan, M. J., de CAC, W., Cano, A., and Crombez, G. (2005). Facing others in pain: the effects of empathy. *Pain*, 118(3):285–288.

Grice, H. P. (1975). Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Grice, P. (1989). *Studies in the way of words*. Harvard University Press, Cambridge, MA.

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

Guy, N., Azulay, H., Kardosh, R., Weiss, Y., Hassin, R. R., Israel, S., and Pertzov, Y. (2019). A novel perceptual trait: Gaze predilection for faces during visual exploration. *Scientific reports*, 9(1):1–12.

Hadjistavropoulos, H. D., Ross, M. A., and Von Baeyer, C. L. (1990). Are physicians' ratings of pain affected by patients' physical attractiveness? *Social Science & Medicine*, 31(1):69–72.

Hadjistavropoulos, T. and Craig, K. D. (2002). A theoretical framework for understanding self-report and observational measures of pain: a communications model. *Behaviour research and therapy*, 40(5):551–570.

Hadjistavropoulos, T., Craig, K. D., Duck, S., Cano, A., Goubert, L., Jackson, P. L., Mogil, J. S., Rainville, P., Sullivan, M. J., Williams, A. C. d. C., et al. (2011). A biopsychosocial formulation of pain communication. *Psychological bulletin*, 137(6):910.

Hardcastle, V. G. (2015). Perception of pain. In *The Oxford handbook of philosophy of perception*.

Hayes, T. R. and Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods*, 48:510–527.

Head, H. and Holmes, G. (1911). Sensory disturbances from cerebral lesions. *Brain*, 34(2-3):102–254.

Heimberg, R. G., Mueller, G. P., Holt, C. S., Hope, D. A., and Liebowitz, M. R. (1992). Assessment of anxiety in social interaction and being observed by others: The social interaction anxiety scale and the social phobia scale. *Behavior therapy*, 23(1):53–73.

Heins, R. C., Mirza, M. B., Parr, T., Friston, K., Kagan, I., and Pooresmaeili, A. (2020). Deep active inference and scene construction. *Frontiers in Artificial Intelligence*, 3:509354.

Henderson, J. M. (2017). Gaze control as prediction. *Trends in cognitive sciences*, 21(1):15–23.

Henderson, J. M. and Luke, S. G. (2014). Stable individual differences in saccadic eye movements during reading, pseudoreading, scene viewing, and scene search. *Journal of Experimental Psychology: Human Perception and Performance*, 40(4):1390.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., and Ramstead, M. J. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural computation*, 33(2):398–446.

Hill, M. L. and Craig, K. D. (2002). Detecting deception in pain expressions: the structure of genuine and deceptive facial displays. *Pain*, 98(1-2):135–144.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science*, 30(1):3–41.

Hoffman, K. M., Trawalter, S., Axt, J. R., and Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16):4296–4301.

Hu, Z., Li, S., Zhang, C., Yi, K., Wang, G., and Manocha, D. (2020). Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1902–1911.

Iannetti, G. D., Hughes, N. P., Lee, M. C., and Mouraux, A. (2008). Determinants of laser-evoked eeg responses: pain perception or stimulus saliency? *Journal of neurophysiology*, 100(2):815–828.

Iannetti, G. D. and Mouraux, A. (2010). From the neuromatrix to the pain matrix (and back). *Experimental brain research*, 205(1):1–12.

Ingvar, M. (1999). Pain and functional imaging. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 354(1387):1347–1358.

Insua, D., Ruggeri, F., and Wiper, M. (2012). *Bayesian analysis of stochastic process models*. John Wiley & Sons.

Izard, C. E. (1993). Four systems for emotion activation: cognitive and noncognitive processes. *Psychological review*, 100(1):68.

Jackson, P. L., Meltzoff, A. N., and Decety, J. (2005). How do we perceive the pain of others? a window into the neural processes involved in empathy. *Neuroimage*, 24(3):771–779.

James, W. (1884). What is an emotion? *Mind*, pages 188–205.

Jänig, W. (2012). Autonomic reactions in pain. *Pain*, 153(4):733–735.

Jannach, D., Manzoor, A., Cai, W., and Chen, L. (2021). A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

Jarrell, T. W., Gentile, C. G., Romanski, L. M., McCabe, P. M., and Schneiderman, N. (1987). Involvement of cortical and thalamic auditory regions in retention of differential bradycardiac conditioning to acoustic conditioned stimuli in rabbits. *Brain research*, 412(2):285–294.

Jensen, M. P. and Karoly, P. (2011). Self-report scales and procedures for assessing pain in adults.

Jensen, M. P., Turner, J. A., Romano, J. M., and Karoly, P. (1991). Coping with chronic pain: a critical review of the literature. *Pain*, 47(3):249–283.

Johnson, E. O., Kamilaris, T. C., Chrousos, G. P., and Gold, P. W. (1992). Mechanisms of stress: a dynamic overview of hormonal and behavioral homeostasis. *Neuroscience & Biobehavioral Reviews*, 16(2):115–130.

Joshi, A., Bhattacharyya, P., and Carman, M. J. (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Jurafsky, D. (2004). Pragmatics and computational linguistics. *Handbook of pragmatics*, pages 578–604.

Kaczkurkin, A. N., Burton, P. C., Chazin, S. M., Manbeck, A. B., Espensen-Sturges, T., Cooper, S. E., Sponheim, S. R., and Lissek, S. (2017). Neural substrates of overgeneralized conditioned fear in ptsd. *American Journal of Psychiatry*, 174(2):125–134.

Kao, J. T., Wu, J. Y., Bergen, L., and Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Kappesser, J., Williams, A. C. d. C., and Prkachin, K. M. (2006). Testing two accounts of pain underestimation. *Pain*, 124(1-2):109–116.

Kartynnik, Y., Ablavatski, A., Grishchenko, I., and Grundmann, M. (2019). Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*.

Katsumi, Y., Quigley, K., and Barrett, L. F. (2021). Situating allostasis and interoception at the core of human brain function.

Katz, M. J. (1988). Fractals and the analysis of waveforms. *Computers in biology and medicine*, 18(3):145–156.

Klein, C. (2007). An imperative theory of pain. *The Journal of Philosophy*, 104(10):517–532.

Kloeden, P. and Neuenkirch, A. (2013). Convergence of numerical methods for stochastic differential equations in mathematical finance. In *Recent Developments in Computational Finance: Foundations, Algorithms and Applications*, pages 49–80. World Scientific.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press, Cambridge, MA.

Kouyanou, K., Pither, C. E., Rabe-Hesketh, S., and Wessely, S. (1998). A comparative study of iatrogenesis, medication abuse, and psychiatric morbidity in chronic pain patients with and without medically explained symptoms. *Pain*, 76(3):417–426.

Kowler, E. (2011). Eye movements: The past 25 years. *Vision Research*, 51(13):1457–1483. 50th Anniversary Special Issue of Vision Research - Volume 2.

Kringelbach, M. L. (2005). The human orbitofrontal cortex: linking reward to hedonic experience. *Nature reviews neuroscience*, 6(9):691–702.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of machine learning research*.

Kunz, M. and Lautenbacher, S. (2014). The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain. *European Journal of Pain*, 18(6):813–823.

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, 25(3):296 – 324.

Lang, V. A., Lundh, T., and Ortiz-Catalan, M. (2021). Mathematical and computational models for pain: A systematic review. *Pain Medicine*, 22(12):2806–2817.

Lassiter, D. and Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 194(10):3801–3836.

LeDoux, J. E. (2012). Evolution of human emotion: a view through fear. *Progress in brain research*, 195:431–442.

Lee, M. C., Mouraux, A., and Iannetti, G. D. (2009). Characterizing the cortical activity through which pain emerges from nociception. *Journal of Neuroscience*, 29(24):7909–7916.

Lindquist, K. A. and Barrett, L. F. (2008). Constructing emotion: The experience of fear as a conceptual act. *Psychological science*, 19(9):898–903.

Lissek, S., Bradford, D. E., Alvarez, R. P., Burton, P., Espensen-Sturges, T., Reynolds, R. C., and Grillon, C. (2014). Neural substrates of classically conditioned fear-generalization in humans: a parametric fmri study. *Social cognitive and affective neuroscience*, 9(8):1134–1142.

Liversedge, S. P. and Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in cognitive sciences*, 4(1):6–14.

Loggia, M. L., Berna, C., Kim, J., Cahalan, C. M., Martel, M.-O., Gollub, R. L., Wasan, A. D., Napadow, V., and Edwards, R. R. (2015). The lateral prefrontal cortex mediates the hyperalgesic effects of negative cognitions in chronic pain patients. *The Journal of Pain*, 16(8):692–699.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., et al. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., and Chen, S. H. A. (2021). NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696.

Marek, S. and Dosenbach, N. U. F. (2018). The frontoparietal network: function, electrophysiology, and importance of individual precision mapping. *Dialogues in Clinical Neuroscience*, 20(2):133–140.

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X., Petrides, M., Jefferies, E., and Smallwood, J. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 113(44):12574–12579.

Marr, D. and Vaina, L. (1982). Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 214(1197):501–524.

Marstaller, L., Burianová, H., and Reutens, D. C. (2017). Adaptive contextualization: A new role for the default mode network in affective learning. *Human brain mapping*, 38(2):1082–1091.

McCaffery, M. (1968). *Nursing practice theories related to cognition, bodily pain, and man-environment interactions*. University of California Print. Office.

McCaffery, M., Ferrell, B. R., and Pasero, C. (2000). Nurses' personal opinions about patients' pain and their effect on recorded assessments and titration of opioid doses. *Pain Management Nursing*, 1(3):79–87.

McCulloch, W. S. (1945). A heterarchy of values determined by the topology of nervous nets. *The bulletin of mathematical biophysics*, 7:89–93.

McEwen, B. S. (1998). Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York academy of sciences*, 840(1):33–44.

McGuire, D. B. (1992). Comprehensive and multidimensional assessment and measurement of pain. *Journal of pain and symptom management*, 7(5):312–319.

Meerman, E. E., Verkuil, B., and Brosschot, J. F. (2011). Decreasing pain tolerance outside of awareness. *Journal of psychosomatic research*, 70(3):250–257.

Melzack, R. (1983). Pain measurement and assessment. *(No Title)*.

Melzack, R. (1989). Phantom limbs, the self and the brain (the do hebb memorial lecture). *Canadian Psychology/Psychologie Canadienne*, 30(1):1.

Melzack, R. (1999). From the gate to the neuromatrix. *Pain*, 82:S121–S126.

Melzack, R. and Casey, K. L. (1968). Sensory, motivational, and central control determinants of pain: a new conceptual model. *The skin senses*, 1:423–43.

Melzack, R., Wall, P. D., et al. (1965). Pain mechanisms: a new theory. *Science*, 150(3699):971–979.

Mesulam, M. (2008). Representation, inference, and transcendent encoding in neurocognitive networks of the human brain. *Annals of neurology*, 64(4):367–378.

Miller, G. A., Galanter, E., and Pribram, K. H. (1960). Plans and structure of behavior. holt, reinhart and winston. *Inc., New York*.

Mirza, M. B., Adams, R. A., Mathys, C., and Friston, K. J. (2018). Human visual exploration reduces uncertainty about the sensed world. *PloS one*, 13(1):e0190429.

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016a). Scene construction, visual foraging, and active inference. *Frontiers in computational neuroscience*, 10:56.

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016b). Scene construction, visual foraging, and active inference. *Frontiers in Computational Neuroscience*, 10:56.

Mouraux, A., Diukova, A., Lee, M. C., Wise, R. G., and Iannetti, G. D. (2011). A multisensory investigation of the functional significance of the "pain matrix". *Neuroimage*, 54(3):2237–2249.

Mouraux, A. and Iannetti, G. D. (2009). Nociceptive laser-evoked brain potentials do not reflect nociceptive-specific neural activity. *Journal of neurophysiology*, 101(6):3258–3269.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press, Cambridge, MA.

Ochsner, K. N. and Gross, J. J. (2005). The cognitive control of emotion. *Trends in cognitive sciences*, 9(5):242–249.

Ong, D. C., Zaki, J., and Goodman, N. D. (2018). Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in Cognitive Science*, 11(2):338–357.

Ortony, A. and Turner, T. J. (1990). What's basic about basic emotions? *Psychological review*, 97(3):315.

Ossipov, M. H., Dussor, G. O., Porreca, F., et al. (2010). Central modulation of pain. *The Journal of clinical investigation*, 120(11):3779–3787.

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.

Patil, A., Huard, D., and Fonnesbeck, C. J. (2010). Pymc: Bayesian stochastic modelling in python. *Journal of statistical software*, 35(4):1.

Pekkanen, J. and Lappi, O. (2017). A new and general approach to signal denoising and eye movement classification based on segmented linear regression. *Scientific reports*, 7(1):1–13.

Peng, C.-K., Havlin, S., Stanley, H. E., and Goldberger, A. L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos: an interdisciplinary journal of nonlinear science*, 5(1):82–87.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature reviews neuroscience*, 9(2):148–158.

Peters, A., McEwen, B. S., and Friston, K. (2017). Uncertainty and stress: Why it causes diseases and how it is mastered by the brain. *Progress in neurobiology*, 156:164–188.

Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the national academy of sciences*, 88(6):2297–2301.

Pineles, S. L. and Mineka, S. (2005). Attentional biases to internal and external sources of potential threat in social anxiety. *Journal of abnormal psychology*, 114(2):314.

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information*. Oxford University Press, New York, NY.

Pitcher, G. (1970). Pain perception. *The Philosophical Review*, 79(3):368–393.

Pollick, A. S. and De Waal, F. B. (2007). Ape gestures and language evolution. *Proceedings of the National Academy of Sciences*, 104(19):8184–8189.

Price, D. D. (2000). Psychological and neural mechanisms of the affective dimension of pain. *Science*, 288(5472):1769–1772.

Price, D. D. et al. (1999). *Psychological mechanisms of pain and analgesia*, volume 15. IASP press Seattle.

Price, D. D., Harkins, S. W., and Baker, C. (1987). Sensory-affective relationships among different types of clinical and experimental pain. *Pain*, 28(3):297–307.

Prkachin, K. M. and Craig, K. D. (1995). Expressing pain: The communication and interpretation of facial pain signals. *Journal of Nonverbal Behavior*, 19(4):191–205.

Prkachin, K. M. and Solomon, P. E. (2008). The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274.

Rai, S. and Chakraverty, S. (2020). A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

Rainville, P., Carrier, B., Hofbauer, R. K., Bushnell, M. C., and Duncan, G. H. (1999). Dissociation of sensory and affective dimensions of pain using hypnotic modulation. *Pain*, 82(2):159–171.

Rainville, P., Feine, J. S., Bushnell, M. C., and Duncan, G. H. (1992). A psychophysical comparison of sensory and affective responses to four modalities of experimental pain. *Somatosensory & motor research*, 9(4):265–277.

Raja, S. N., Carr, D. B., Cohen, M., Finnerup, N. B., Flor, H., Gibson, S., Keefe, F. J., Mogil, J. S., Ringkamp, M., Sluka, K. A., et al. (2020). The revised international association for the study of pain definition of pain: concepts, challenges, and compromises. *Pain*, 161(9):1976–1982.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Rees, G., Kreiman, G., and Koch, C. (2002). Neural correlates of consciousness in humans. *Nature Reviews Neuroscience*, 3(4):261–270.

Rescorla, R. A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory*, 2:64–69.

Reutter, M. (2022). Diagnostic facial features & fear generalization.

Reutter, M. and Gamer, M. (2022). Individual patterns of visual exploration predict the extent of fear generalization in humans. *Emotion*.

Reutter, M. and Gamer, M. (2023). Individual patterns of visual exploration predict the extent of fear generalization in humans. *Emotion*, 23(5):1267.

Richardson, D. E. and Akil, H. (1977). Long term results of periventricular gray self-stimulation. *Neurosurgery*, 1(2):199–202.

Richman, J. S. and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American journal of physiology-heart and circulatory physiology*, 278(6):H2039–H2049.

Rizzolatti, G. and Sinigaglia, C. (2016). The mirror mechanism: a basic principle of brain function. *Nature Reviews Neuroscience*, 17(12):757–765.

Roberts, S. J., Penny, W., and Rezek, I. (1999). Temporal and spatial complexity measures for electroencephalogram based brain-computer interfacing. *Medical & biological engineering & computing*, 37:93–98.

Robinson, M. E. and Riley III, J. L. (1999). The role of emotion in pain.

Roesmann, K., Wiens, N., Winker, C., Rehbein, M. A., Wessing, I., and Junghoefer, M. (2020). Fear generalization of implicit conditioned facial features–behavioral and magnetoencephalographic correlates. *Neuroimage*, 205:116302.

Rosati, A. G. (2017). Foraging cognition: reviving the ecological intelligence hypothesis. *Trends in cognitive sciences*, 21(9):691–702.

Rosenthal, S. H., Porter, K. A., and Coffey, B. (1990). Pain insensitivity in schizophrenia: case report and review of the literature. *General hospital psychiatry*, 12(5):319–322.

Sabeti, M., Katebi, S., and Boostani, R. (2009). Entropy and complexity measures for eeg signal classification of schizophrenic and control participants. *Artificial intelligence in medicine*, 47(3):263–274.

Sajid, N., Ball, P. J., Parr, T., and Friston, K. J. (2021). Active inference: demystified and compared. *Neural Computation*, 33(3):674–712.

Salamone, J. D. and Correa, M. (2012). The mysterious motivational functions of mesolimbic dopamine. *Neuron*, 76(3):470–485.

Salomons, T. V., Johnstone, T., Backonja, M.-M., Shackman, A. J., and Davidson, R. J. (2007). Individual differences in the effects of perceived controllability on pain perception: critical role of the prefrontal cortex. *Journal of cognitive neuroscience*, 19(6):993–1003.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. peerj computer science, 2, e55.

Satpute, A. B. and Lindquist, K. A. (2019). The default mode network's role in discrete emotion. *Trends in cognitive sciences*, 23(10):851–864.

Schäfer, G., Prkachin, K. M., Kaseweter, K. A., and de C Williams, A. C. (2016). Health care providers' judgments in chronic pain: the influence of gender and trustworthiness. *Pain*, 157(8):1618–1625.

Schiavenato, M. and Craig, K. D. (2010). Pain assessment as a social transaction: beyond the "gold standard". *The Clinical journal of pain*, 26(8):667–676.

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

Schwartz, N., Temkin, P., Jurado, S., Lim, B. K., Heifets, B. D., Polepalli, J. S., and Malenka, R. C. (2014). Decreased motivation during chronic pain requires long-term depression in the nucleus accumbens. *Science*, 345(6196):535–542.

Schwöbel, S., Kiebel, S., and Marković, D. (2018). Active inference, belief propagation, and the bethe approximation. *Neural computation*, 30(9):2530–2567.

Scott, D. J., Heitzeg, M. M., Koeppe, R. A., Stohler, C. S., and Zubieta, J.-K. (2006). Variations in the human pain stress experience mediated by ventral and dorsal basal ganglia dopamine activity. *Journal of Neuroscience*, 26(42):10789–10795.

Searle, J. R., Willis, S., et al. (1995). *The construction of social reality*. Simon and Schuster.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., and Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience*, 27(9):2349–2356.

Senft, G. (2014). *Understanding pragmatics*. Routledge, New York, NY.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences*, 17(11):565–573.

Seth, A. K. and Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1708):20160007.

Seuren, P. A. (2009). *Language from within: Vol. 1. Language in cognition*. Oxford University Press.

Seymour, B. (2019). Pain: a precision signal for reinforcement learning and control. *Neuron*, 101(6):1029–1041.

Seymour, B. and Mancini, F. (2020). Hierarchical models of pain: Inference, information-seeking, and adaptive control. *NeuroImage*, 222:117212.

Sharp, T. J. (2001). Chronic pain: a reformulation of the cognitive-behavioural model. *Behaviour Research and Therapy*, 39(7):787–800.

Shepherd, S. V. and Platt, M. L. (2007). Spontaneous social orienting and gaze following in ringtailed lemurs (lemur catta). *Animal Cognition*, 11(1):13.

Shi, J., Samal, A., and Marx, D. (2006). How effective are landmarks and their geometry for face recognition? *Computer vision and image understanding*, 102(2):117–133.

Shimo, K., Ueno, T., Younger, J., Nishihara, M., Inoue, S., Ikemoto, T., Taniguchi, S., and Ushida, T. (2011). Visualization of painful experiences believed to trigger the activation of affective and emotional brain regions in subjects with low back pain. *PLoS One*, 6(11):e26681.

Simon, D., Craig, K. D., Gosselin, F., Belin, P., and Rainville, P. (2008). Recognition and discrimination of prototypical dynamic expressions of pain and emotions. *PAIN®*, 135(1-2):55–64.

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., and Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, 303(5661):1157–1162.

Singh, M. K., Giles, L. L., and Nasrallah, H. A. (2006). Pain insensitivity in schizophrenia: trait or state marker? *Journal of Psychiatric Practice®*, 12(2):90–102.

Smith, R., Friston, K. J., and Whyte, C. J. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, 107:102632.

Society, A. P. (1999). *Principles of analgesic use in the treatment of acute pain and cancer pain*. American Pain Society.

Solomon, P. R. and Moore, J. W. (1975). Latent inhibition and stimulus generalization of the classically conditioned nictitating membrane response in rabbits (oryctolagus cuniculus) following hippocampal ablation. *Journal of comparative and physiological psychology*, 89(10):1192.

Solomon, R. C. (2008). *True to our feelings: What our emotions are really telling us*. Oxford University Press.

Sperber, D. and Wilson, D. (1986). *Relevance: Communication and cognition*. Harvard University Press, Cambridge, MA.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., Seth, A. K., et al. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in human neuroscience*, 10:550.

Stephens, D. W. (1986). *Foraging theory*. Princeton University Press, Princeton, NJ.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Storm, H. (2008). Changes in skin conductance as a tool to monitor nociceptive stimulation and pain. *Current Opinion in Anesthesiology*, 21(6):796–804.

Strube, A., Horing, B., Rose, M., and Büchel, C. (2023). Agency affects pain inference through prior shift as opposed to likelihood precision modulation in a bayesian pain model. *Neuron*, 111(7):1136–1151.

Sullivan, M. J. (2008). Toward a biopsychomotor conceptualization of pain: implications for research and intervention. *The Clinical journal of pain*, 24(4):281–290.

Sullivan, M. J., Thibault, P., Savard, A., Catchlove, R., Kozey, J., and Stanish, W. D. (2006). The influence of communication goals and physical demands on different dimensions of pain behavior. *Pain*, 125(3):270–277.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tabor, A., Thacker, M. A., Moseley, G. L., and Körding, K. P. (2017). Pain: a statistical account. *PLoS computational biology*, 13(1):e1005142.

Tait, R. C. and Chibnall, J. T. (1997). Physician judgments of chronic pain patients. *Social science & medicine*, 45(8):1199–1205.

Tatler, B., Hayhoe, M., Land, M., and Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5).

Tatler, B. and Vincent, B. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6-7):1029–1054.

Tatler, B. W. and Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, 2(2).

Teich, A. H., McCabe, P. M., Gentile, C. G., Jarrell, T. W., Winters, R. W., Liskowsky, D. R., and Schneiderman, N. (1988). Role of auditory cortex in the acquisition of differential heart rate conditioning. *Physiology & behavior*, 44(3):405–412.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Terkourafi, M. (2021). Pragmatics as an interdisciplinary field. *Journal of Pragmatics*, 179:77–84.

Tessler, M. H. and Goodman, N. D. (2019). The language of generalization. *Psychological Review*, 126(3):395–436.

Thiam, P., Kessler, V., Walter, S., Palm, G., and Schwenker, F. (2016). Audio-visual recognition of pain intensity. In *IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction*, pages 110–126. Springer.

Todd, P. M. and Hills, T. T. (2020). Foraging in mind. *Current Directions in Psychological Science*, 29(3):309–315.

Tomasello, M. (2010). *Origins of human communication*. MIT press, Boston, MA.

Tomkins, S. (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.

Tong, X., Shutova, E., and Lewis, M. (2021). Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686.

Torta, D., Legrain, V., Mouraux, A., and Valentini, E. (2017). Attention to pain! a neurocognitive perspective on attentional modulation of pain in neuroimaging studies. *Cortex*, 89:120–134.

Toscano, B. J., Gownaris, N. J., Heerhartz, S. M., and Monaco, C. J. (2016). Personality, foraging behavior and specialization: integrating behavioral and food web ecology at the individual level. *Oecologia*, 182(1):55–69.

Tracey, I., Ploghaus, A., Gati, J. S., Clare, S., Smith, S., Menon, R. S., and Matthews, P. M. (2002). Imaging attentional modulation of pain in the periaqueductal gray in humans. *Journal of Neuroscience*, 22(7):2748–2752.

Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., and Blei, D. M. (2016). Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.

Trappes, R. (2022). Individual differences, uniqueness, and individuality in behavioural ecology. *Studies in History and Philosophy of Science*, 96:18–26.

Treister, R., Kliger, M., Zuckerman, G., Aryeh, I. G., and Eisenberg, E. (2012). Differentiating between heat pain intensities: the combined effect of multiple autonomic parameters. *PAIN®*, 153(9):1807–1814.

Tsai, F.-S., Weng, Y.-M., Ng, C.-J., and Lee, C.-C. (2017). Embedding stacked bottleneck vocal features in a lstm architecture for automatic pain level classification during emergency triage. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 313–318. IEEE.

Turk, D. C. (1996). Cognitive factors in chronic pain and disability.

Turk, D. C. and Okifuji, A. (2002). Psychological factors in chronic pain: evolution and revolution. *Journal of consulting and clinical psychology*, 70(3):678.

Turk, D. C. and Rudy, T. E. (1992). Cognitive factors and persistent pain: A glimpse into pandora's box. *Cognitive therapy and research*, 16(2):99–122.

Tye, M. (2005). Another look at representationalism about pain. *Pain: New essays on its nature and the methodology of its study*, pages 99–120.

Urban, M. and Gebhart, G. (1999). Supraspinal contributions to hyperalgesia. *Proceedings of the National Academy of Sciences*, 96(14):7687–7692.

Valentini, E., Vaughan, S., and Clauwaert, A. (2023). Qualia, brain waves, and spinal reflexes: The study of pain perception by means of subjective reports, electroencephalography, and electromyography. In *Somatosensory Research Methods*, pages 129–159. Springer.

van de Meent, J.-W., Paige, B., Yang, H., and Wood, F. (2018). An introduction to probabilistic programming. *arXiv preprint arXiv:1809.10756*.

Vasil, J., Badcock, P. B., Constant, A., Friston, K., and Ramstead, M. J. (2020). A world unto itself: human communication as active inference. *Frontiers in psychology*, 11:480375.

Viswanathan, G. M., Da Luz, M. G., Raposo, E. P., and Stanley, H. E. (2011). *The physics of foraging: an introduction to random searches and biological encounters*. Cambridge University Press, Cambridge, UK.

Vogt, B. A. (2005). Pain and emotion interactions in subregions of the cingulate gyrus. *Nature Reviews Neuroscience*, 6(7):533–544.

Walsh, J., Eccleston, C., and Keogh, E. (2014). Pain communication through body posture: The development and validation of a stimulus set. *PAIN®*, 155(11):2282–2290.

Watt-Watson, J., Stevens, B., Garfinkel, P., Streiner, D., and Gallop, R. (2001). Relationship between nurses' pain knowledge and pain management outcomes for their postoperative cardiac patients. *Journal of advanced nursing*, 36(4):535–545.

Webler, R. D., Berg, H., Fhong, K., Tuominen, L., Holt, D. J., Morey, R. A., Lange, I., Burton, P. C., Fullana, M. A., Radua, J., and Lissek, S. (2021). The neurobiology of human fear generalization: meta-analysis and working neural model. *Neuroscience & Biobehavioral Reviews*, 128:421–436.

Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S., and Traue, H. C. (2016). Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing*, 8(3):286–299.

Werner, P., Al-Hamadi, A., and Walter, S. (2017). Analysis of facial expressiveness during experimentally induced heat pain. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 176–180. IEEE.

Werner, P., Lopez-Martinez, D., Walter, S., Al-Hamadi, A., Gruss, S., and Picard, R. (2022). Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, 13(1):530–552.

Wiech, K. (2016). Deconstructing the sensation of pain: The influence of cognitive processes on pain perception. *Science*, 354(6312).

Wiech, K., Farias, M., Kahane, G., Shackel, N., Tiede, W., and Tracey, I. (2008). An fmri study measuring analgesia enhanced by religion as a belief system. *Pain*, 139(2):467–476.

Wiech, K., Kalisch, R., Weiskopf, N., Pleger, B., Stephan, K. E., and Dolan, R. J. (2006). Anterolateral prefrontal cortex mediates the analgesic effect of expected and perceived control over pain. *Journal of Neuroscience*, 26(44):11501–11509.

Wild, J. and Blampied, N. (1972). Hippocampal lesions and stimulus generalization in rats. *Physiology & behavior*, 9(4):505–511.

Williams, A. C. d. C. (2002). Facial expression of pain: an evolutionary account. *Behavioral and brain sciences*, 25(4):439–455.

Williams, A. C. d. C. and Craig, K. D. (2016). Updating the definition of pain. *Pain*, 157(11):2420–2423.

Williams, D. A. and Keefe, F. J. (1991). Pain beliefs and the use of cognitive-behavioral coping strategies. *Pain*, 46(2):185–190.

Wilson, H. R. and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24.

Wilson, P. H., Henry, J. L., and Nicholas, M. K. (1993). Cognitive methods in the management of chronic pain and tinnitus. *Australian Psychologist*, 28(3):172–180.

Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *Journal of Vision*, 13(3).

Yoon, E. J., Tessler, M. H., Goodman, N. D., and Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, 4:71–87.

Zelinski, E. L., Hong, N. S., Tyndall, A. V., Halsall, B., and McDonald, R. J. (2010). Prefrontal cortical contributions during discriminative fear conditioning, extinction, and spontaneous recovery in rats. *Experimental brain research*, 203:285–297.

Zhang, L., Losin, E. A. R., Ashar, Y. K., Koban, L., and Wager, T. D. (2021). Gender biases in estimation of others' pain. *The journal of pain*, 22(9):1048–1059.

Zhang, M., Cui, Z., Neumann, M., and Chen, Y. (2018). An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.