UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

Dipartimento di Informatica ‘Giovanni Degli Antoni’
Corso di Dottorato in Informatica
XXXIV Ciclo

Tesi di Dottorato di Ricerca

On wiring emotion to words: a Bayesian model

SSD 01/B1

Autore
Sathya Buršić

Tutor
Prof. Alfio Ferrara and Prof. Giuseppe Boccignone

Coordinatore del Dottorato
Prof. Paolo Boldi

A.A. 2020-2021

Abstract

LANGUAGE and emotion are deeply entangled. In this dissertation we present a theoretical model that addresses how language and emotions intertwine with one another. To such end, we draw on the several results achieved in emotion theory (either at the psychological and the neurobiological levels) that go under the constructivist umbrella of the Conceptual Act Theory and those related to an emerging theoretical framework for pragmatic inference, the Rational Speech Act framework. We connect these theories and spell such connection in the language of probability, namely in that of Bayesian probabilistic modelling.

Our endeavour is addressed to those fields of computer science such as artificial intelligence and machine learning where, in spite of the remarkable progress in the computational processing of language and affect, the study of their intersection is at best at its infancy, in our view. We argue that any further step in such direction only can be afforded by reducing the gap between Affective Science and computational approaches. To pave the way, simulations of the proposed model are presented that account for well known case-studies in pragmatics.

In brief, at a high-level abstract representation we consider two interacting agents-in-context, where each agent performs a conceptual act based on interoceptive and exteroceptive sensation, in order to regulate their body budget. The agents communicate, performing communication acts that in turn regulate the agents' conceptual acts and vice versa, and in this way they create, communicate and share categories, and even add new functions to the world. We implement this framework through two simulations of non-literal language use, namely hyperbole, irony, and a third dealing with politeness, a form of social reasoning. In addition, a fourth simulation concerns the assessment of the stochastic dynamics of the key component of the model, core affect.

Contents

CHAPTER 1

Introduction

How do language and emotions intertwine with one another? Language and emotion are certainly linked. Humans use words to describe how we feel in spoken conversations, when thinking to ourselves, and when expressing ourselves in writing. Yet, a commonly held view is that the sole function of language is to communicate our thoughts. Shakespeare's Juliet famously worded such surmise: "That which we call a rose, by any other name would smell as sweet". But findings from psychology and neuroscience are beginning to suggest otherwise: a flower might indeed be perceived as sweeter by virtue of being categorized as a "rose." Language influences what concepts, in the embodied or grounded cognition sense, are brought to mind when constructing an experience.

In this perspective, language and emotion have more than a mere unidirectional relationship. Language plays a critical role both in communicating thoughts and emotions, but also in constitutively creating emotions. Emotions in the view embraced in this dissertation, a constructionist theoretical perspective, are instances of mental life: experiences in which affective feelings (bodily feelings of arousal, pleasure/displeasure) tend to be a salient feature, and that are organized and understood (made meaningful) with respect to emotion categories such as anger, fear, joy, etc., at least in Western culture. The semantic representation of emotion concepts (the mental representation of categories) is constitutive of emotion. Emotion words and the concepts they signify shape and even help constitute emotional experiences. Strikingly, language does not necessarily require rich narratives to modulate affective feelings: even simple verbal suggestions and individual words can be impactful. Meanwhile, memories of emotion perception are shaped by language: narrative processing can exaggerate memories of others' emotional expressions. Words people use to describe their experiences of emotion and their autobiographical memories play a role in shaping those mental phenomena. Further, there is accumulating evidence that language may even be necessary for the perception of emotions. The term "alexithymia" defines adults who struggle with identifying and describing their emotions with words: adults who are high on alexithymia also exhibit disruptions in emotion perception and experience.

On the other hand, language is a fundamentally social endeavor. It has been argued that it is not language that makes human communication possible, but rather a special underlying communicative ability that makes language possible. In many instances, linguistic expressions are underdetermined with respect to the meaning that they convey: what is said and what is understood are often not the same. Communication goes further than the exchange of explicit propositions. The goal of the speaker is to either change the mind of the listener, or to commit the addressee to the execution of certain actions, such as closing the window in reply to the statement "It is cold here". In other words, a theory of speech acts is needed to understand how we get from coded meaning to inferred speaker meaning. Pragmatics is the study of how speakers and listeners use social reasoning to go beyond the literal meanings of words to interpret language in context. Crucially, context is to be intended in a broad sense, including both linguistic factors (such as the previous discourse or conversation) and extralinguistic elements, from the physical setting to the psychological aspects of the interlocutors: world knowledge, emotions, beliefs, stereotypes, etc. As individuals, we process words, search for hidden meanings or innuendos, and react to sentiment and affect that is embedded in sentences. But in a natural setting, things become more complex. Think for instance to the utmost, though paradigmatic example of a conversation unfolding in a flirting context: spoken words, prosody, coyly smiles and laughter, sing-song voices, hair touch, eager glances, action lining up in mimicry, either voluntary or inescapable, entangle in a unique voice to communicate with one another.

Clearly, the broad spectrum spanned by such mind-blowing problems require a truly interdisciplinary lens that blends questions and tools from fields as far-ranging as linguistics, psychology, neuroscience, biological anthropology, cultural anthropology and sociology. Addressing this question is the purpose of Affective Science.

Our goal. The main concern of the present dissertation is to provide a theoretical model that addresses the initial question of how language and emotions intertwine with one another. Here the term "theoretical model" is precisely intended in the sense of Marr's computational theory and Anderson's rational analysis: the *what/why* level of analysis concerning the individuation of a computable function as a model of a given behavioural phenomenon. To such end, we draw on the several results achieved in emotion theory (either at the psychological and the neurobiological levels) that go under the constructivist umbrella of the Conceptual Act Theory and those related to an emerging theoretical framework for pragmatic inference, the Rational Speech Act framework. We connect these theories and spell such connection in the language of probability, namely in that of Bayesian probabilistic modelling.

Beyond the intrinsic appeal of modelling the conundrum of emotion and language entanglement, our endeavour is addressed to those fields of computer science such as Artificial Intelligence and machine learning where, in spite of the remarkable progress in the computational processing of language and affect, the study of their intersection is at best at its infancy, in our view. We argue that any further step in this direction only can be afforded by reducing the gap between Affective Science and those computational approaches.

Motivations. In recent decades, there has been a veritable explosion of research on both language and emotion in Affective Computing , which should be, in principle, the computer science counterpart of Affective Science. Affective Computing is the interdisciplinary field of study concerned with recognizing, understanding, simulating and stimulating affective states in the design of computational systems. Since the coining of the term by Rosalind Picard, Affective Computing has emerged as a cohesive and increasingly impactful discipline spanning computer science, psychology, neuroscience, philosophy, art and industry. Technology giants such as Apple, Amazon, Google and Facebook, as well as hundreds of smaller companies are deploying Affective Computing methods to predict or influence consumer behavior.

Unfortunately, Affective Computing has been slow to adopt the numerous advances from the Affective Science field; consequently, many Affective Computing models are rooted in outdated affect theories, such as Basic Emotion Theories, and methodology. Clearly, there is a conceptual gap between social and biological scientists who try to understand emotions on the one hand and computer scientists and engineers who try to build emotion savvy applications on the other hand.

In this field we should take breath and turn back, humbly, to William James' unescapable question: What is an emotion? All in all, to compute something, one is expected to know what is being computed. At the light of this minimalist requirement, computers scientists and engineers should understand that it is not sufficient for them to rely on common sense knowledge of what constitutes emotions - or even worst on outdated theories. Breathtaking models, techniques and technologies now available off the machine learning and AI shelf offer modest help: the power of the engine is irrelevant, if the direction of the journey is unsettled. As Noam Chomsky put it : "you do not get discoveries in the sciences by taking huge amounts of data, throwing them into a computer and doing statistical analysis of them: that's not the way you understand things, you have to have theoretical insights." This state of affairs is not specific to Affective Computing, but a general challenge in the emerging context of computational social science.

Consider language. Tremendous progress has been been made in Natural Language Processing (NLP). Though, in principle, NLP should span from syntax to semantics and to pragmatics, much less effort has been devoted to modelling pragmatic reasoning. Advancements mostly relate to deep learning methods for distributional semantic models (e.g., word embeddings and transformers). These models seem to be learning latent representations that capture the same basic ideas as grammars and semantic information. Data-driven models are easier to develop and maintain and score better on standard benchmarks compared to hand-built systems - which explicitly take into account grammars, parsing, and semantic interpretation - that can be constructed using a reasonable amount of human effort. This is a positive trend; however, it has been argued that many recent are surprisingly oblivious of the large body of previous work in fields like cognitive science and computational linguistics. By and large, computational paradigms such as NLP, as daily practiced in the machine learning arena, are uniquely suited to resolve problems with low power in psychological science by incorporating millions of datapoints. At the intersection between NLP and Affective Computing, what is now called "sentic computing" aims to bridge the gap between statistical NLP and many other disciplines that are necessary for understanding human language, such as linguistics, commonsense reasoning, semiotics, and affective computing. The task of detecting emotions in textual conversations leads to a wide range of applications: sentiment analysis, opinion mining and emotion mining are well known examples of such effort. But even in the case of language, we are far from addressing our initial question.

Such state of affairs has indeed provided the main motivation to engage in the challenging journey reported in this thesis.

Thesis plan. The present dissertation unfolds as follows.

- The Prelude (Chapter 2) sets an example for the arguments that follow.
- Chapter 3 discusses the methodological approach guiding our work
- Chapter 4 motivates the assumption that language is essentially a social endeavour and frames the Communication Act. The key problem of conceptualization is discussed to some depth.
- Chapter 5 deals with the fundamental problem of the Conceptual Act. Words, concepts and categorization are revisited at its light.
- Chapter 6 lays down the model. First constraints for both acts are derived at the neurobiological level and mapped onto the psychological level. A functional architecture is outlined, which is then exploited to shape in the language of probability our model. To such end both Probabilistic Graphical Model and Probabilistic Programming representation are used. Further theoretical implications and perspectives related to the model are also discussed
- Chapter 7 presents and discusses model simulations accounting for well known case-studies in pragmatics.
- Conclusions are drawn in Chapter 8

Technical subtleties and apparently less related topics are however included for the sake of completeness, but these have been confined in the Appendices.

CHAPTER 2

Prelude

How much do our perceptions of things depend on our cognitive/affective ability, and how much on our linguistic resources? Where, and how, do these two topics meet?

These are the crucial questions that motivate this thesis.

Umberto Eco, in the endeavour of addressing such fundamental issues, which is epitomized in the brilliant essay “Kant And The Platypus” (Eco, 2000), provides an enlightening example by resorting to the story of Montezuma and the horses.

The story concerns the first Aztecs who hastened to the coast and witnessed the landing of the conquistadors. Among the various things that must have completely amazed them, Eco surmises that, in particular, the horses might have been perceptually puzzling:

> At first (maybe also because they did not distinguish the animals from the pennants and armor that covered them), the Aztecs thought that the invaders were riding deer. Oriented therefore by a system of previous knowledge but trying to coordinate it with what they were seeing, they must have soon worked out a perceptual judgment. *An animal has appeared before us that seems like a deer but isn’t* [...] They must therefore have got a certain idea of that animal, which at first they called *maçatl*, which is the word they used not only for deer but for all quadrupeds in general. Later, since they began adopting and adapting the foreign names for the objects brought by the invaders, their Nahuatl language transformed the Spanish *caballo* into *cauayo* or kawayo.

The Aztecs decided to send messengers to Montezuma to tell him of the landing and of the marvels they were witnessing.

One scribe gave the news in pictograms, and he explained that the invaders were riding deer as high as the roofs of the houses:

> I imagine that the messengers (worried about the fact that in their neck of the woods, if the news was not to the hearer's liking, there was a tendency to punish the bearer of it) screwed up their courage and integrated the report with more than just words, since it seems that Montezuma was wont to require his informers to provide him with all the possible expressions for one and the same thing. And so they must have used their bodies to hint at the movements of the *maçatl*, imitating its whinnying, trying to show how it had long hair along its neck, adding that it was most terrifying and ferocious, capable in the course of the fray of overwhelming anyone who tried to withstand it.

Thus, Montezuma received some descriptions, on the basis of which he tried to get some idea of that as yet unknown animal, albeit it is difficult to tell how he imagined it. He probably understood that it was a worrisome animal:

> according to the chronicles, at first Montezuma did not ask other questions but withdrew into a distressing silence, with head bowed and wearing an absent, sorrowful air.

Eventually, the encounter between Montezuma and the Spaniards. As Eco puts it:

> I would say that, no matter how confused the messengers' description may have been, Montezuma must have easily identified those things called *maçaoa*[the plural of *maçatl*]. Simply, faced with the direct experience of the *maçatl*, he must have adjusted the tentative idea he had conceived of them. Now, like his men, every time he saw a *maçatl*, he too would recognize it as such, and every time he heard talk of *maçaoa*, he would understand what his interlocutors were talking about. Then, as he gradually got to know the Spaniards, he would learn many things about horses, he would begin to call them *cauayo*, he would learn where they came from, how they reproduced, what they ate, how they were reared and trained, what other uses they could be put to, and to his regret he would very soon understand how useful they could be in battle.

The story puts on the table all the ingredients we are dealing with in this thesis:

- the initial, private but contextual construction of meaning from a *referent object* situated in the world up to its mental representation, the *concept*, as an *inference* grounded in the referent's perceptual and affective features (terrifying and ferocious) by contrast to previously acquired knowledge (dear, quadruped, denoted by *maçatl*), and its subsequent labelling based on a new *word* (*cauayo*);
- the multimodal communication with more than just words (gestures, body movements, prosody and drawings) driven by the *goal* to optimally convey the meaning sufficient to support Montezuma's learning of the same concept, even in the absence of the physical referent, but on the base of a *common ground* both *lexical* and conceptual (the animal named *maçatl*);

- the subsequent broadening and updating of prior knowledge, via social communication, of the *cauayo* mental representation into a more general horse *category*.

The recipe we advocate to amalgamate such ingredients relies on the rationale that humans are social animals born with a brain system that has evolved to support social affiliation. A social species is one where animals regulate one another's fundamental physiological processes (or allostasis), so that their survival depends on social bonds (Atzil et al., 2018).

Under such circumstances, many human psychological features can be best understood in a social context. Language is one such case. It is a fundamentally social endeavor: speakers and listeners use social reasoning to go beyond the literal meanings of words to interpret language in context (Bohn and Frank, 2019).

Throughout early language development, social communication is the central organizing principle of language use. Learning occurs in the context of use and communication is central to learning as well.

Over and over, social animals, by using social communication, gradually learn to regulate their own and others' allostasis, namely the ongoing adjustment of the individual's internal milieu, which is necessary for survival, growth and reproduction (Atzil et al., 2018), and which is at the root of the storm of feelings that the individual experiences everyday in life. During development, infants learn social concepts and skills to prepare for allostatic needs, as caretakers introduce all the culturally relevant concepts, using language. Meaning itself is largely grounded in the structure of the overarching social interaction (Bohn and Frank, 2019).

The communicative act and the conceptual act, rooted in the entangled perceptions of the external world and of the internal milieu, are but two sides of the same coin. The modeling framework we present here is an attempt to incorporate both aspects in a straightforward though principled way.

CHAPTER 3

A methodological foreword and a roadmap

As we stated from the beginning, in this study we are addressing the relationship between language and affect in the context of a dyadic interaction. The general idea is outlined in Figure 3.1

Two agents, a speaker and a listener, are involved in a dyadic social interaction. A speaker, who is observing an event/object in a context, utters a sentence/word U passing on information concerning the perceived state $\mathcal{O}$ of the outside world. The listener will reason about speaker's intents and actions within that context and may either take some action and/or, by switching roles, reply to the speaker. Both agents share common perceptual and affective mechanisms, neurobiologically grounded in a common brain/body structure, for inferring and conceptualize the external world and a common language to communicate with one another. Also, beyond the propositional meaning of the uttered sentence, the listener, to recover the speaker's intended meaning, can rely on speaker's non-verbal signalling, such as gestures, prosody, facial expressions and so on. Non-verbal signals might be intentionally conveyed by the speaker or might unveil speaker's affective state, which can thus be appraised by the listener. In a sense, the speaker's non-verbal behavior is part of the events E occurring in the external world as perceived by the listener, i.e. $\mathcal{O} = \{\mathcal{O}_{ext}, \mathcal{O}_{speaker}\}$, while the social interaction unfolds.

In a subject such as this, it is perhaps best to start by establishing models of an apt generality, so to avoid *ad hoc* heuristics, while considering relevant and yet well defined case studies, in order not to complicate an already difficult problem.

Thus, in this chapter we first make clear the methodological framework we have adopted. This can be characterised as a multilevel analysis framework, which aims at devising a theoretical model but informed and constrained by knowledge that we have

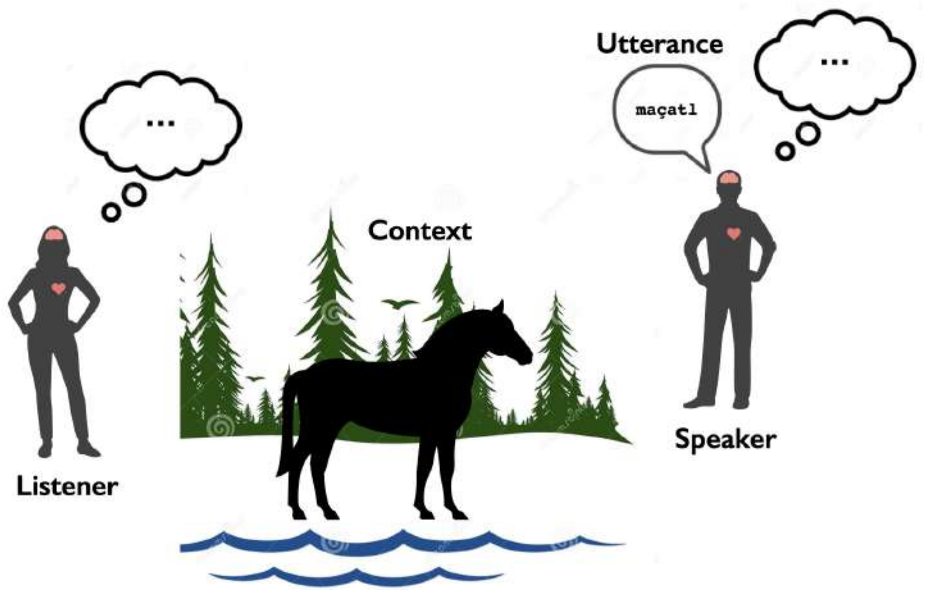


Figure 3.1: *A dyadic social interaction between a speaker and a listener, much like in the Montezuma's story. Note that in both speaker and listener we have highlighted either the brain and the internal body milieu - represented by the heart - as intertwined components enabling the dyadic communication act and the conceptual act that lies behind*

available both at the psychological and at the neuroscience explanation levels.

3.1 Levels of explanation

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

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

Such distinction has been further developed by Marr (1982), who proposed three levels of description/explanation (see Fig. 3.2, left):

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

2. the *how* level (algorithm),

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

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

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

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

A critical point here concerns the constraints assumed by computational theory (Anderson's"rational explanation") when aiming to reduce the underdetermination and the arbitrariness or ad-hocness or mimicry, as Marr put it, of cognitive models at the algorithmic level (Anderson's "mechanistic explanation"). In particular, the "heuristics for searching an explanation" can be seen as a guide in the choice of a cognitively plausible mechanism, given that there are usually many mechanisms instantiating the same performance. Briefly, this is the model underdetermination problem, which although being a general problem in scientific explanation, has proven to be remarkably acute for cognitive explanation. In classical Cognitive Science, this issue was deeply discussed in depth by Pylyshyn (1984) in terms of the specific constraints the cognitive scientist has to assume in order to guarantee the "psychological reality" or plausibility of computational models. This issue is also dealt with by embodied cognitive science too, this time introducing plausible constraints stemming from the environment and the body. Further, this seems to be an issue raised at the time of Cybernetics and early Artificial Intelligence (AI); all these topics have been pointed out by recent analyses: see Cordeschi (2002).

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

In the light of the growing exploitation of Bayesian methods in the cognitive sciences, it has been argued (Chater et al., 2006; Knill et al., 1996; Boccignone and Cordeschi, 2007) that Marr's three-fold hierarchy could be reorganised into two levels: the *computational theory* level, which can be formalised precisely in terms of Bayesian theory, and the *implementation theory* level, embedding both Marr's algorithmic and physical realisation levels (see Fig. 3.2, right).

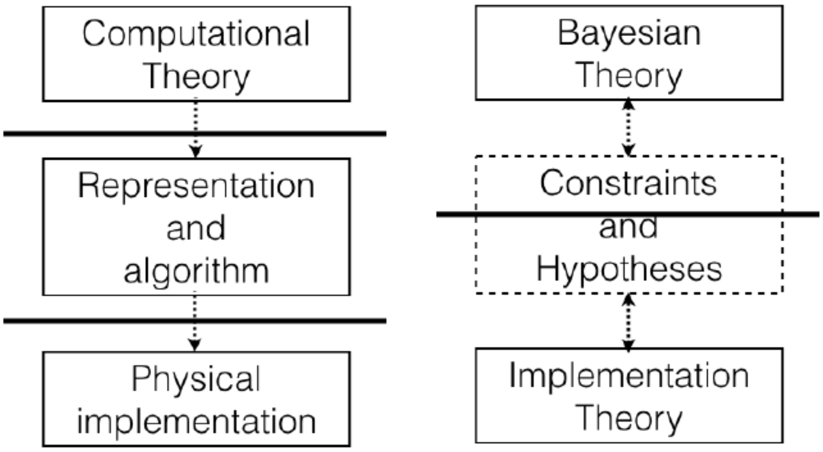


Figure 3.2: *The levels of explanation in cognitive/behavioural sciences. Left: Marr's original proposal (Marr, 1982). Right: Marr's revision according to Yuille and Kersten (adapted from Knill et al., 1996).*

As Figure 3.2, shows, within the Bayesian approach issues about constraints can be settled in a way quite different from Marr's, particularly in relation to his three-fold hierarchy of levels of explanation.

Note that both levels are denoted "theories" here and, differently from Marr, a close interaction between the computational (here Bayesian) theory and the implementation theory level is assumed. Further, hypotheses and constraints are somehow shared between the two levels (see broken-line box in Fig. 3.2).

In this thesis we do endorse this two-level view. Also, we use the term "model" to qualify the two theory levels in Fig. 3.2: briefly, in what follows we will refer to such levels as the *theoretical model*[1]and the *implementation model*. This well reflects the fact that, at both levels, the cognitive scientist is devising models embodying constraints related to a number of physical or biological laws and theoretical hypotheses that are relevant to the explanation of a given phenomenon. This is a point regarding both models explaining behavioural regularities (at the Bayesian theory level) and models explaining neural regularities (at the implementation theory level).

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

3.1.1 The theoretical model

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

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

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

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

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

3.1.2 Subtleties of the implementation model

The most straightforward implementation model to "put into work" the theoretical model can be obtained by specifying the probability distributions defining the conditional dependencies and by applying suitable PGM-based algorithms such as Belief Propagation, Variational Bayes learning, etc. (but see Koller and Friedman, 2009 for an in-depth introduction, and Erk, 2021 for a brief introduction oriented to language modelling). This can be seen as the coarsest-grain implementation model. But it might be the case that an implementation model at a finer grain is needed to be addressed.

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

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

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

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

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

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

3.1.3 Putting all together: multilevel analysis

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

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

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

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

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

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

3.2 Roadmap

Taking stock of the above discussion, at the most general level, a theoretical model of the interacting agents in its bare essential should be able to cope with the following requirements:

- both agents are capable of a *conceptual act*, that is they can perceive and conceptualize the states of the world $P(\mathcal{C}(\texttt{world})|\mathcal{O}(\texttt{world}))$ given a collection of events or *outcomes* $\mathcal{O}(\texttt{world})$ occurring in the world; the world includes both the environment and agents acting in the world: thus, crucially, the conceptualization $\mathcal{C}(\texttt{world})$ depends on both the perception of the external world and the internal perception of agent's body, which brings in the game the affective value of what is externally perceived and of what is communicated (either intentionally or not)

- the speaker can perform a *communicative act* by uttering a sentence or a word to convey some meaning M concerning the current situation (state of the world and speaker's mental state) given a lexicon $\mathcal{L}$ and a language model $\mathcal{LM}$, i.e. $P(u \mid \mathcal{C}(\texttt{world}), M, \mathcal{L}, \mathcal{LM})$.

- the listener, by hearing the utterance U and by observing world events (external events and the speaker's non-verbal behavior), can jointly reason about the states of the world and the meaning the speaker intended to convey, $P(\mathcal{C}(\texttt{world}), M \mid U, \mathcal{O}(\texttt{world}), \mathcal{L}, \mathcal{LM})$.

The scheme outlined in Figure 3.1 together with the probabilistic inferences introduced above provide a blueprint (Tsuji et al., 2021) for setting out the model proposed in this Thesis.

Beyond its apparent simplicity, a number of mind-blowing questions hide behind.

How do agents ground language in actual events in the world (semantics)? How do they convey meaning under certain circumstances, context and goals (pragmatics) along a social interaction? How do they conceptualize events in the world? What is the role played by affect at the different stages?

In the following, we will first present and discuss theories and proposals that engage with such questions. Given the terribly vast literature on these topics, we will of necessity focus on aspects most relevant for our perspective in order to distill the fundamental elements to detail the model's blueprint.

CHAPTER 4

The communication act

We discuss and motivate here the assumption that language is a fundamentally social endeavor, where speakers and listeners use social reasoning to go beyond the literal meanings of words in order to interpret language in context. The big picture is outlined in Figure 4.1 which expands Figure 3.1.

We will briefly touch on those aspects of semantics that are of interest for our research. We also introduce the notions of concept and categorization, which we will be re-examined and extended in Chapter 5, at the light of emotion theories, in particular the Conceptual Act Theory (CAT).

Then we turn on pragmatics that frames communication acts.

While introducing aspects of semantics and pragmatics most relevant for us, we will also touch on computational approaches that have been proposed to address problems that, by and large, arise in both realms.

4.1 Semantics: The Relationship Between Words and Concepts

Semantics is the study of the meaning of words, phrases and sentences. In linguistics and philosophy, semantics is taken to stand for the relation between language and the world.

Historically, as a theory, it denotes the branch of linguistics and logic concerned with meaning (Yule, 2020). The two main areas are logical semantics, concerned with matters such as sense and reference and presupposition and implication, and lexical semantics, concerned with the analysis of word meanings and relations between them (but for a broad introduction, cfr. Speaks, 2021; Yule, 2020. For the purposes of this thesis, we will be mostly involved with the latter aspect.

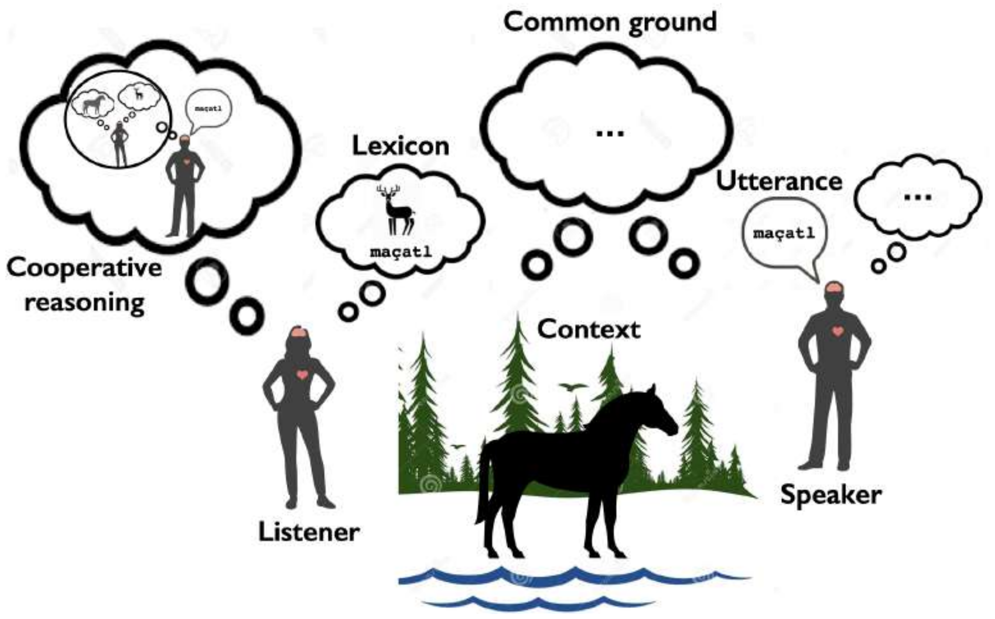


Figure 4.1: *Schematic overview of the communication process between a speaker and a listener during which different sources of information are integrated. Observable variables are the utterance (*"maçatl"*), the context, and additional social cues provided by the speaker. Unobserved psychological variables are lexicon (and concepts), common ground, the cooperative reasoning process, and the inner perceptual processes behind conceptualization. Modified after (Bohn and Frank, 2019)*

In semantics there is always an attempt to focus on what the words conventionally mean, rather than on what a speaker might want the words to mean on a particular occasion. When linguists investigate the meaning of words in a language, they are interested in characterising the conceptual meaning or, less frequently, the associative meaning of words.

Conceptual meaning covers those basic, essential components of meaning which are conveyed by the literal use of a word. Simply put, some of the basic components of a word like "horse" in English might include "large plant-eating domesticated mammal with solid hoofs and a flowing mane and tail". These components would be part of the conceptual meaning of "horse", namely the mental representation of the *horse* category. In the sequel, we will adopt the following broad distinction between category and concept (Murphy, 2004):

category : a population of events or objects that are treated as similar because they all serve a particular goal in some context;

concept : the population of representations that correspond to those events or objects.

One may have "associations" or "connotations" attached to a word like "horse" that lead to think of "hazard" in relation to racing bets, but the association usually is not treated as a part of the conceptual meaning of "horse".

Concepts and words seem to be closely related. One can talk about children learning the concept of horse, say, but one can also talk about their learning the word "horse". In fact, much of the literature uses these two terms interchangeably (Murphy, 2004).

Here, in the same vein of Murphy (2004) we will make a distinction between the two. By *concept*, we denote a nonlinguistic psychological (mental) construct of a class of entities in the world, namely an agent's knowledge of what kinds of things there are in the world, and what properties they have. By *word meaning*, we denote quite generally the aspect of words that gives them significance and relates them to the world. Clearly, based on evidence provided by the psychological literature, words gain their significance in virtue of being connected to concepts (Murphy, 2004). Figure 4.2 outlines at a glance these relationships in terms of a Peircean semiotics triangle (Peirce, 1991).

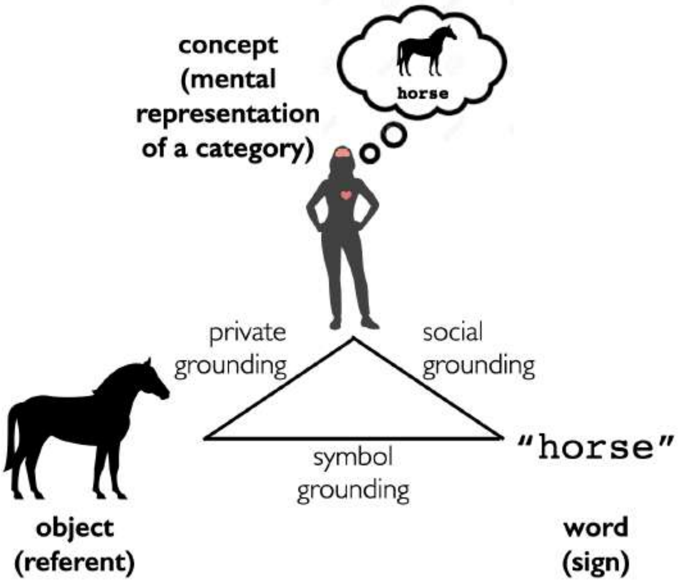


Figure 4.2: *The conceptual view of word meaning. The scheme can also be more generally read in terms of Peircean semiotics where a symbol is defined as a process having three elements. The first is the sign (*representamen*), which describes the form that the sign takes; the sign, e.g., words, visual signs, or pointing, is not a symbol itself. The second is the* object*, which is something that the sign refers to. The third is the* interpretant*, which is the sense made of the sign; the interpretant mediates between the sign and the object. In the Peircean definition, a symbol is not a static material, but a dynamic process of interpretation. Peirce calls this process "semiosis" (Peirce, 1991).*

As said, we will not discuss in general linguistic meaning, which is an enormous topic that has a huge literature in philosophy and linguistics (Speaks, 2021; Yule, 2020), and the meaning of larger linguistic structures (sentence, discourse, or story). For our purposes it will suffice to focus on how word meanings are psychologically represented.

It is however worth, for the sake of completeness, to briefly mention two milestones in these fields that have been largely influential for subsequent studies.

4.1.1 A brief history of semantics

By the time Chomsky's "Aspects of the Theory of Syntax" was published in 1965 (Chomsky, 1965), generative grammars were understood to have a semantic component in addition to a syntactic and phonological component. It was assumed that a speaker's knowledge of his language required her to have tacit knowledge of a generative grammar of it. A speaker of a natural language has the ability to understand indefinitely many sentences of his language that he has never previously encountered. Indeed, his ability to understand any sentence of his language does not depend on his having a prior acquaintance with it. What explains this remarkable ability?

One answer can be shaped in the form of the generative grammar hypothesis (GGH): the ability of a speaker of a natural language L to understand sentences of L requires him to have tacit knowledge of a generative grammar of L, that being a finitely specifiable theory of L that generates one or more syntactic structures for each sentence of L and interprets those structures both phonologically and semantically.

The question that defined semantics in linguistics was the form that the internally represented semantic theory should take, and that is what the defining question was taken to be in Katz and Fodor's seminal 1963 manifesto, "The structure of a semantic theory," (Katz and Fodor, 1963) the first serious effort to do semantics in generative linguistics. Katz and Fodor posited that understanding a sentence was the ability to determine the number and content of the readings of a sentence, to detect semantic anomalies, and to decide on paraphrase relations between sentences. The semantic theory they claimed was needed to explain such ability, and thus to be the semantic component of a generative grammar that verifies GGH, must have two components: (i) a dictionary of the language and (ii) projection rules that select the appropriate sense of each lexical item in a sentence in order to provide the correct readings for each distinct grammatical structure of that sentence.

The impact of philosophy and logic on semantics in linguistic work of the 50's and 60's was limited; many linguists knew some first-order logic, aspects of which began to be borrowed into linguists semantic representations, and there was gradually increasing awareness of the work of some philosophers of language. Generative semanticists in the late 1960's and early 1970's in particular started giving serious attention to issues of "logical form" in relation to grammar, and to propose ever more abstract underlying representations intended to serve simultaneously as unambiguous semantic representations and as input to the mapping from meaning to surface form (see, for instance, Lakoff, 1971)

Then around 1970 linguistic semantics took a curious turn. Without rejecting the claim that speaking a language requires tacit knowledge of a semantic theory of it, linguists turned away from the project of characterizing the nature of that theory in order to pursue instead the Montague-inspired project of providing for the languages we speak the same kind of formal semantics that logicians devise for the artificial languages of formal systems of logic. The external influence that set linguistic semantics on its course was the model-theoretic approach to the metalogic of formal systems of logic that emerged from the work of logicians Löwenheim, Skolem, Gödel, Tarski, Church, Kripke and others in the years between 1915 and 1965.

Montague made three remarkable claims pertaining to natural language semantics (but see Montague, 2019, for a recent account), namely: (1) there is no important theoretical difference between natural languages and the uninterpreted formal languages of systems of logic; (2) it's possible to treat a natural language as an uninterpreted formal language and to construct for it a model-theoretic semantics of exactly the same meta-mathematical kind that a logician would provide for the formal language of a system of intensional logic that captured the logical entailments expressible in the natural language; (3) the construction of such a semantics should be the goal of any serious semantics for natural language

Formal semantics originally signified semantics for formal languages devised for the mathematical study of formal systems of logic, but the expression now has a meaning akin to "analytical philosophy" and signifies the Montague-inspired approach to the semantical study of natural languages. At the heart of the approach is Montague's *Principle of Compositionality*: The meaning of any complex expression is a function of the meanings of its parts and of the way they are syntactically combined.

Indeed, the Fregean principle of compositionality was central to Montague's theory and remains central in formal semantics. The nature of the elements of both the syntactic and the semantic algebras is open to variation; what is constrained by compositionality is the relation of the semantics to the syntax. The crucial structure for syntax is the "derivation tree", showing what parts have been combined at each step, by what syntactic rule. In this perspective, syntax is an algebra of forms, semantics is an algebra of meanings, and there is a homomorphism mapping the syntactic algebra into the semantic algebra.

Nowadays, there are many theories of syntax, and many theories of semantics, and the interface questions look different for all of them. For instance, taking a radical stance, Jackendoff and Jackendoff (2002) suggests a view on which semantic structures and syntactic structures are independently generated, and the interface conditions may be quite complex.

An epistemological remark on computational ontologies

Computational ontologies are a means to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation, and which are useful to our purposes Guarino et al. (2009). Informally, an ontology is a formal, explicit specification of a shared conceptualization. More formally, an ontology is a logical theory designed to account for the intended meaning of the vocabulary used by a logical language:

> Let C be a conceptualization, and L a logical language with vocabulary V and ontological commitment K. An ontology O_K for C with vocabulary V and ontological commitment K is a logical theory consisting of a set of formulas of L, designed so that the set of its models approximates as well as possible the set of intended models of L according to K

In this case the predicate symbol Horse has both an extensional interpretation (through the usual notion of model, or extensional first-order structure) and an intensional interpretation (through the notion of ontological commitment, or intensional first order structure)

In this view, ontologies are collections of axioms that intend to capture the semantics of the terms used in a certain domain of discourse and bring the text that belong to the domain within the reach of standard, model-theoretic semantic approaches.

There are many critical aspects of this approach. Even a "perfect" ontology may fail to exactly specify its target conceptualization, if its vocabulary (from informal glossaries and data dictionaries to formal logical languages) and its domain of discourse are not suitably chosen. A complete discussion is out of the scope of this thesis but we leave to Santini and Dumitrescu (2008) for a poignant review of these topics. It will suffice to highlight that the perspective on meaning given by computational ontologies is very different from the contextual one that is necessary in order to create meaning, and herein lies its main limitation. This limitation goes beyond the use of a specific logic system: it derives from the disregard of interpretation as a creator of meaning and, consequently, from the idea that meaning is a thing rather than a process (Santini and Dumitrescu, 2008).

To conclude this brief review, many theorists -including many formal semanticists- recognize that the theories semanticists construct under the formal semantics rubric cannot plausibly be regarded as theories of the kind needed to explain a speaker's knowledge of her language.

This bifurcation raises concerns the relation between, on the one hand, the psychologically explanatory semantic theories still thought to be needed but no longer the object of study in linguistic semantics and, on the other hand, the theories formal semanticists are concerned to construct. A cogent issue since a psychologically explanatory theory can supersede even the best formal semantic theory.

To conclude this brief account, we observe that, by and large, in the various approaches that have flourished from the sixties, the theory of semantics is often spelled in terms of *referential semantics*, arguing that, in the end, words get their meanings by referring to real objects and events. Thus, a statement is true just in case it corresponds to a situation in the world: word meaning is simply a relation between a word and the world, i.e. a reference. In brief, the meaning of "horse" is the set of horses in the world.

Is this tenable as a psychological theory?

4.1.2 Word meaning, concepts and psychology: a first glance

Clearly, people do not know or have access to the sets all the horses in the world.

The psychological approach surmises that people do not know about every example of each word they know. Instead, it assumes that people have some sort of mental description, the concept, that allows them to pick out examples of the word and to understand it when they hear it. Word meanings are psychologically represented by mapping words onto conceptual structures.

Yet, in turn, this assumption raises a number of issues (Murphy, 2004).

The first relates to the kind of mapping. In the simplest view, we have a one-to-one mapping (word = concept). Beyond the problem of accomodating synonyms and ambiguous words, the real challenge is that there are many concepts that do not have a word to go with them (for instance, "The moist residue left on a window after a dog presses its nose to it"). In this case we have to admit that word < concept, in the sense that some concepts are not labeled by words.

Further, even unambiguous words often have a number of different, related senses. For example, "theater" can refer to the institution which puts on plays and the building in which one views the plays: they are related but they are not the same thing. To sum up, the word/concept mapping is not a simple one, if one has to account polysemy (e.g. "Put it on the table" vs. "The entire table shared pizza") and contextual modulation (but again, cfr. Murphy, 2004 for an in-depth discussion)

In the end it must be admitted, that a complete model of word meaning cannot be represented in a simple mapping, because the number of variables and complexity of the structures involved precludes a simple depiction. If one abstracts the conceptual structure in the shape of a graph, where nodes and arcs (their connections), words are likely to be connected to a node or subgraph. A word usually has a number of different senses, like word 2 in Figure 4.3. The meaning of the word then consists in coherent chunks of conceptual knowledge that are picked out by the lexical item. However, note that the conceptual component also includes other information from the general domain, including superordinates, coordinates (nodes that share an immediate superordinate), and other related concepts. Clearly, the substructure being picked out by the word is not a simple list of components, but a possibly elaborate structure. Further, the conceptual structure being picked out by the word is closely integrated with related structures, and these probably have considerable influence on the use of the word. For example, the word "horse" is embedded in biological knowledge of animals in general as well as knowledge of horse riding and so forth, and this knowledge might be activated to various degrees on some occasions in which one hears the word "horse" though it is not an official part of the meaning.

On the one hand, a sub-graph being selected by the word is not a simple list of components, but a possibly elaborate structure. It is in general hard to know how to determine the number of nodes, links, and their relations in order to assess each word's overall complexity. On the other hand, although one needs not assume that every aspect of conceptual knowledge is retrieved when a word is understood, adjacent related conceptual chunks are important for specifying the meaning by providing contrast, background knowledge, and underlying assumptions.

This very fact has important theoretical implications. First it is very difficult to provide a simple set of components that is a word's meaning. This undermines classic theories of semantic representation stating that word meaning can be cast as a list of semantic components (Katz and Fodor, 1963) that can be gained via a look-up in a dictionary.

Second, and most important for us, relates to the problem of *how* to exploit the conceptual structure, namely which process is put into action on the structure. When a sentence emphasizes one aspect of a word meaning, that aspect is more activated, and other aspects may not be activated. This implies that in order to exploit the structure some kind of constructive/inferential process is needed in order to constrain possible senses.

At this point, it is clear that the issues we have discussed, the merits and faults of the different approaches for solving the semantic problem, crucially depend on how we define a concept and its structure.

The problem here is that the literature on word meaning has not been directed towards distinguishing theories of concept representation Murphy (2004).

The concept conundrum: prototypes or many exemplars?

We have defined a concept as the mental representation of a category. This raises a question: what is a mental representation and how it is shaped? Early psychological approaches to concepts took a definitional approach, a view that has been dominant one since Aristotle. The *classical view* Murphy (2004), which can be dated back to

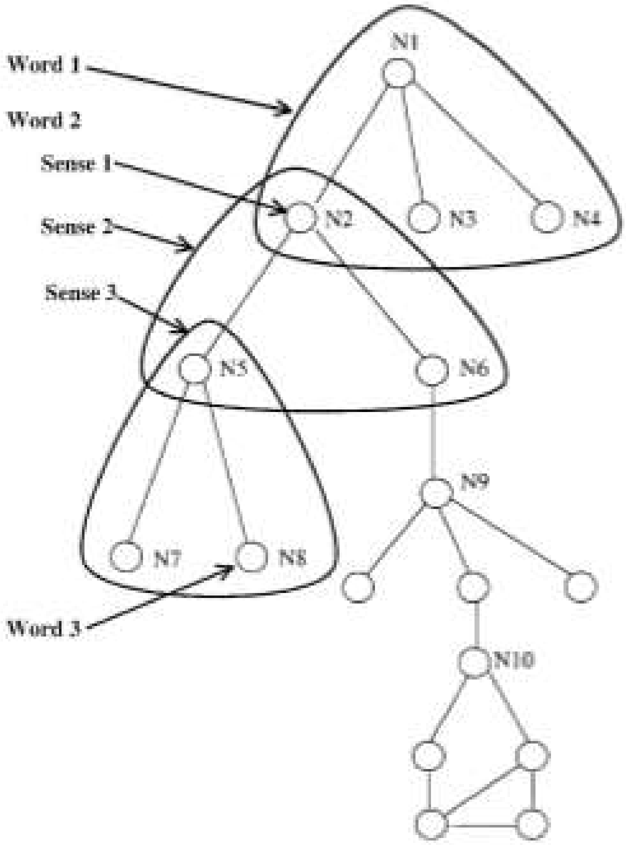


Figure 4.3: *Word meaning as a complex mapping of words on a conceptual structure shaped as a graph. Individual, unambiguous, words usually have a number of different senses, for example word 2. These senses are rendered as overlapping sub-graphs of the structure. Sense 1 of word 2 picks out a single node, N2, in the conceptual structure. Yet, note that N2 has superordinate N1, coordinates N3 and N4, and subordinates N5 and N6. Adjacent sub-graphs might be suitable to provide background knowledge. Adapted from Murphy (2004)*

Hull's early work, posits that: 1) concepts are mentally represented as definitions that provide characteristics that are a) necessary and b) jointly sufficient for membership in the category; 2) every object is either in or not in the category, with no in-between cases.

If description of a horse picks out all horses and nothing else, one has given a successful definition. An animal that has the feature common to all horses is thereby a horse, just the same as any other thing that has that feature. Clearly, even most definitions in dictionaries do not meet this criterion. .

There are many problems with the equation "concept = definition", either theoretical and empirical. A remarkable argument was proposed by the philosopher Wittgenstein in the definition of what is a game.

As a matter of fact, the work of Eleanor Rosch in the 1970s essentially ruled out the classical view from the field of psychology. What was proposed by Rosch it is known as the *prototype view*. It is usually defined in terms of every category being best represented by a single prototype or best example: the category of horses is represented by a single ideal horse, which best embodies all the attributes normally found in horses.

Clearly, this view is questionable. If from horses, one moves to birds, what single "best bird" could work for penguins, ostriches, pelicans, hummingbirds, turkeys, parrots, and sparrows? Also, a single prototype is unsuitable to account for the variability within a large category. Due to such drawbacks, the prototype view has most often been interpreted as a summary representation, namely a description of the category as a whole, rather than describing a single, ideal member. The representation itself could be described in terms of family-resemblance, where the concept is conceived as the ensemble of features that are usually found in the category members, with some features being more relevant than others. A similarity computation can be used to identify similar exemplars, in the form of an additive function of matching and mismatching features.

Clearly, by substituting the single best example with an ensemble of features, raises in turn the issue of how to determine and organize the feature list. Different proposals have been advanced such as feature combinations and schemata, that also have been influential in the early Artificial Intelligence community.

A radical alternative to the prototype view (and to previous theories of concept) is the *exemplar view* proposed by Medin and Schaffer (1978), named context model. Here, an individual is not supposed to have an entire representation of a category. Instead, a person's concept of horses is the set of horses that the person remembers. In a sense, there is no actual concept, because there is no summary representation that stands for all horses. The decision on distinguishing a horse from a non-horse depends on the similarity to other horses one has seen in the past. Medin and Schaffer (1978) also proposed a computational procedure based on adding up the similarity scores for each exemplar in a category. If one has 50 exemplars of horses in memory, the similarities of the observed animal to these 50 items are added up, the result providing evidence for the animal being a horse; the same is performed with respect to other animal categories, say deer, and so on. The category with the most similarity to the item will win the contest.

The evolution of the Medin and Shaffer context model is the Generalized Context Model (GCM). In the GCM the distance (dissimilarity) between two items indexed i and j, where i indexes the object to be categorized, and j one of the remembered exemplars, is calculated as $d_{ij} = \sqrt{\sum_m w_m (x_{mi} - x_{mj})^2}$, where x_{mk} is the m-th "feature" of the vector x_k, and w_m represents the item weight for dimension m. The similarity score is then computed as $s_i j = \exp -cd_{ij}$, and eventually the Luce Choice Axiom is used to turn this similarity into a response, by cacalculating the posterior probability that the object i, is to be assigned to category, J, $P(J \mid i) = \frac{\sum_{j \in J} s_i j}{\sum_K \sum_{k \in K} s_i k}$. The more similar the item is to known exemplars in J, the higher this probability. The denominator is the similarity of i to members of all known categories K.

GCM has been an extremely influential categorization model. It is easy to note the relation between the GCM and many models used in modern machine learning and pattern recognition (e.g., clustering). To sum up, according to exemplar models of categorization, people decide what a new item is by accessing already-known examples and relying on them, which, in some sense, can be thought of as relying on analogy. For an in-depth discussion of theoretical and empirical evaluations of this model, and subsequent variations, that stand on a voluminous literature on comparisons of exemplar models and prototype models of categorization one should refer to Murphy (2004).

There is *prima facie* an enormous difference between prototype and exemplar models. The prototype view argues that people learn a summary representation of the whole category and use that to decide category membership. Category learning involves the formation of that prototype, and categorization involves comparing an item to the prototype representation. The exemplar view posits that people's category knowledge is represented by specific exemplars, and categorization involves comparing an item to all (or many) such exemplars. Thus, conceptual representations and the processes of learning and categorization all differ between these two models.

Some scholars have concluded that one cannot readily resolve the exemplar-prototype question. Most notably, Barsalou (2014) provided a very detailed analysis, concluding that there is no specific pattern of performance that could be accounted for by only one kind of theory.

Yet, a synthesis is possible at the computational theory level as it has been shown by (Anderson, 1991).

Towards a synthesis: rational analysis of human categorization

Anderson (1991) aims at providing a rational analysis of categories (rational model of categorization, RMC), that is a theoretical model in the sense we have specified in Chapter 3.

He considers three views of the origins of categories:

1. Linguistic: A linguistic label provides a cue that a category exists, and people proceed to learn to identify it

2. Feature overlap: People notice that a number of objects overlap substantially and proceed to form a category to include these items

3. Similar function. People notice that a number of objects serve similar functions and proceed to form a category to include them.

These three views need not be in opposition. They are all special cases of the predictive nature of categories. Categorization is justified by the observation that objects tend to cluster in terms of their attributes, be these physical features, linguistic labels, functions, or whatever.

Thus, according to the RMC, categorization is a special case of feature induction, in which the learner uses the observed features of a stimulus to predict its unobserved features, using the previous stimuli to guide the prediction. Since the model treats category linguistic labels as features, these labels are the obvious features to predict, but other features can be predicted as well.

The basic goal of categorization is to predict the probability of various unexperienced features of objects. The situation can be characterized as one in which n objects have been observed, they have an observed feature structure F_n, and one wants to predict whether a particular object will display some value j on dimension i unobserved for that object.

This amounts to the marginalization

$$P_i(x \mid F_n) = \sum_k P(k \mid F_n) P_i(x \mid k)$$

The sum is across all possible partitionings/categories k of the n objects into disjoint sets, $P_i(x \mid k)$ is the probability that the object in question would display value x on dimension i if k were the partition.

$P(k \mid F_n)$ is the posterior probability of category k given the objects display feature structure F_n, to be inferred via Bayes rule:

$$P(k \mid F_n) = \frac{P(F_n \mid k)P(k)}{\sum_{k'} P(F_n \mid k')P(k')}$$

The sum in the marginalization and the denominator of Bayes'equation are intractable for large n, as the number of partitions grows rapidly with the number of stimuli. Consequently, an approximate inference algorithm is needed.

Anderson identified two desiderata for an approximate inference algorithm: that it be incremental, assigning a stimulus to each cluster as it is seen, and that these assignments, once made, be fixed. These desiderata were based on beliefs about the nature of human category learning: that people need to be able to make predictions all the time not just at particular junctures after seeing many objects and much deliberation, and that people tend to perceive objects as coming from specific categories. To such end, he developed a simple iterative inference algorithm that satisfies these desiderata.

Impressively, Anderson independently discovered one of the most celebrated models in nonparametric Bayesian statistics, the Dirichlet process mixture model (DPMM (Sanborn et al., 2010)), deriving this distribution from first principles. Namely, the problem of predicting an arbitrary feature of a stimulus can be solved by estimating the joint probability of the features of a set of stimuli. This is the statistical problem of density estimation. In Bayesian statistics, this problem is addressed by defining a prior distribution over a set of possible densities, and then updating this distribution with the observed data to obtain a posterior distribution over densities. In nonparametric Bayesian statistics, the goal is to define a prior that includes as broad a range of densities as possible, so that complex densities can be inferred if they are warranted by the data.

The RMC uses a flexible representation that can interpolate between prototypes and exemplars by clustering stimuli into groups, adding new clusters to the representation as required. When a new stimulus is observed, it can either be assigned to one of the pre-existing clusters, or to a new cluster of its own.

Viewing category learning in this way helps to clarify the assumptions behind the two main classes of psychological models, the exemplar and prototype views. As we have seen, exemplar models assume that a category is represented by a set of stored exemplars, and categorizing new stimuli involves comparing these stimuli to the set of exemplars in each category (e.g., Medin and Schaffer, 1978). Prototype models assume that a category is associated with a single prototype and categorization involves comparing new stimuli to these prototypes. These approaches to category learning correspond to different strategies for density estimation used in statistics, being nonparametric and parametric density estimation respectively. RMC takes a third approach, modeling category learning as Bayesian density estimation. This approach encompasses both prototype and exemplar representations, automatically selecting the number of clusters to be used in representing a set of objects. To sum up the RMC is an example of a successful Bayesian model of cognition. It provides a reasonable explanation of how objects should be grouped into clusters and the result of this clustering can be used to explain many categorization experiments (Sanborn et al., 2010).

4.1.3 Grounding concepts: The knowledge based approch

In contrast to prototype and exemplar representations, the *knowledge approach* argues that concepts are part of our general knowledge about the world. We do not learn concepts in isolation from everything else; rather, we learn them as part of our overall understanding of the world around us. When we learn concepts about animals, this information is integrated with our general knowledge about biology, about behavior, and other relevant domains. This relation works both ways: concepts are influenced by what we already know, but a new concept can also effect a change in our general knowledge (consider again the initial Montezuma example). In general, this approach emphasizes that concepts are part and parcel of the individual's general knowledge of the world, and so there is pressure for concepts to be consistent with whatever else the individual knows.

This bears a consequence that is central in our investigation: in order to maintain such consistency, part of categorization and other conceptual processes may be a reasoning process that infers properties or constructs explanations from general knowledge. People use their *prior knowledge* to reason about an example in order to *infer* what category it is, or in order to *learn* a new category. This aspect of concepts was referred to as "mental theories about the world". In this framework, knowledge of how each category fits in with other parts of our lives shapes the *ideal* of the category: for instance, vehicles are made so that people can be moved from place to place; thus, the most typical vehicles would behave this way in the best possible way.

The importance of such knowledge can be illustrated even more by a kind of category that Barsalou (1983, 1985) called *goal-derived categories* or *ad-hoc categories*. These are categories that are defined solely in terms of how their members fulfill some desired goal or plan (e.g., the category of things to take from one's house during a fire). This is an important assumption: as we will see in Chapter 5, emotions will be exactly defined in terms of ad-hoc categories. Barsalou found that the most typical examples of goal-derived categories were the ones that were closest to the ideal.

One of the themes of the knowledge approach, then, is that people do not rely on simple observation or feature learning in order to learn new concepts. They pay attention to the features that their prior knowledge says are the important ones, under the given goal. They may make inferences and add information that is not actually observed in the item itself. Their knowledge is used in an active way to shape what is learned and how that information is used after learning.

The grounding problem

One important topic is the tight connection between conceptualization/categorization and perception as depicted in Figure 4.2.

In most theories, the concept-learning device takes in category exemplars as described by a preexisting vocabulary of features, and it then outputs a category description in terms of those features. However, in a different vein it has been argued that learning must occur at the perceptual level as well; the features themselves must be constructed, often in parallel with category learning.

Indeed, categorization can change the boundaries of perception. For example, perceptual discrimination is heightened along category boundaries, meanwhile perceptual effects such as the "magnet effect" (e.g., the continuum of the color spectrum perceived as bands of segmented colors) arise as a consequence of categorization. Such effects can be elegantly accounted for by the knowledge-based approach shaped in the form of a Bayesian model: the prior term models available knowledge and the likelihood term models the incoming perceptual stimulus, the final effect being summarised by the posterior distribution.

A deep issue behind this findings is whether concepts and knowledge are really perceptual or symbolic: should concept representations be thought of as perceptual or symbolic? Indeed, if our symbolic representations are not all innately given, it is argued, then they may well come from perceptual representations. After all, humans can learn the use of language through physical interaction with their environment and semiotic communication with other people. It is very important to obtain a computational understanding of how humans can form a symbol system and obtain semiotic skills through their autonomous mental development.

Unfortunately, early AI mutuated from logic the idea of symbols. In predicate logic, which is a representative of symbolic logic, predicates and variables that represent real-world phenomena are given as discrete representations in a top-down manner. The fundamental assumption is that our world can be distinguished and segmented into a discrete "symbol" system, and that the system is deterministic and static. The Physical symbol system hypothesis, proposed by (Newell, 1980) was no exception to such assumtpion. As a consequence, the meaning of a symbol is syntactically determined in relation with other symbols.

However, a relationship between two signifiers can never provide the relationship between a signifier and a signified object (cfr. Figure 4.2). Harnad (1990) put on the table the symbol grounding problem, which is one of the most famous problems in AI:

> What is the representation of a zebra? It is just the symbol string "horse & stripes". But because "horse" and "stripes" are grounded in their respective iconic and categorical representations, "zebra" inherits the grounding, through its grounded symbolic representation. In principle, someone who had never seen a zebra (but had seen and learned to identify horses and stripes) could identify a zebra on first acquaintance armed with this symbolic representation alone (plus the nonsymbolic – iconic and categorical – representations of horses and stripes that ground it) (Harnad, 1990)

.

In robotics, Brooks (1991) representatively criticized and insisted that sensory-motor coupling with the environment is primarily important for robots to achieve everyday tasks in our daily environment. For a critical survey of the problem in robotics, see Taniguchi et al. (2016).

In cognitive science, physical symbol systems have also been criticized. Barsalou (1999) proposed the concept of Perceptual Symbol System (PSS), to place an emphasis on perceptual experiences for theories of knowledge and categorization

In its bare essentials (but see Barsalou, 2008 for an overview), a perceptual state can contain two components: an unconscious neural representation of physical input, and an optional conscious experience. Once a perceptual state arises, a subset of it is extracted via selective attention and stored permanently in long-term memory (cfr. Figure 4.4).

On later retrievals, this perceptual memory can function symbolically, standing for referents in the world, and entering into symbol manipulation. As collections of perceptual symbols develop, they constitute the representations that underlie cognition. Perceptual symbols are modal and analogical. They are modal because they are represented in the same systems as the perceptual states that produced them. The neural systems that represent color in perception, for example, also represent the colors of objects in perceptual symbols, at least to a significant extent. On this view, a common representational system underlies perception and cognition, not independent systems.

The very notion of concept in this framework, which is radically different than those previously discussed is that of simulation: a concept is a dynamic entity a *simulator*.

Viewing concepts as simulators suggests a different way of conceiving categorization. Whereas many theories assume that relatively static, amodal structures determine category membership (e.g., definitions, prototypes, exemplars, theories), simulators suggest a more dynamic, embodied approach: if the simulator for a category can produce a satisfactory simulation of a perceived entity, the entity belongs in the category. If the simulator cannot produce a satisfactory simulation, the entity is not a category member (Barsalou, 1999). For example, the perceptual simulations used to categorize chairs approximate the actual perceptions of chairs. In contrast, amodal theories assume that amodal features in concepts are compared to perceived entities to perform categorization (Barsalou, 1999, 2008). Cogently, it is this idea of simulation that paves the way to the notion of conceptual act, which will be discussed in the next chapter.

The PSS hypothesis currently animates hot research fields that are beyond psychology, for instance robotics (Taniguchi et al., 2016). Robotics indeed makes a real case for learning of abstract words such as "use" and "make" in humanoid robot experiments, and the acquisition of numerical concepts via gesture and finger counting strategies. The current approaches share a strong emphasis on *embodied cognition* aspects for the grounding of abstract concepts, and a continuum, rather than dichotomy, view of concrete/abstract concepts differences (see, Cangelosi and Stramandinoli, 2018 for a review).

4.1.4 Other approaches to meaning: associative models

We have been focusing on a conceptual approach to word meaning here, but this has not always been the major approach to word meaning in linguistics and psychology. However, one general approach that has been important in the history of psychology has been based on *word associations*. That is, the meaning of a word is the set of other words (or perhaps words plus other mental entities) it is associated to. This is the bulk of the Distributional Hypothesis (DH) proposed by Harris (1954)

> The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear.

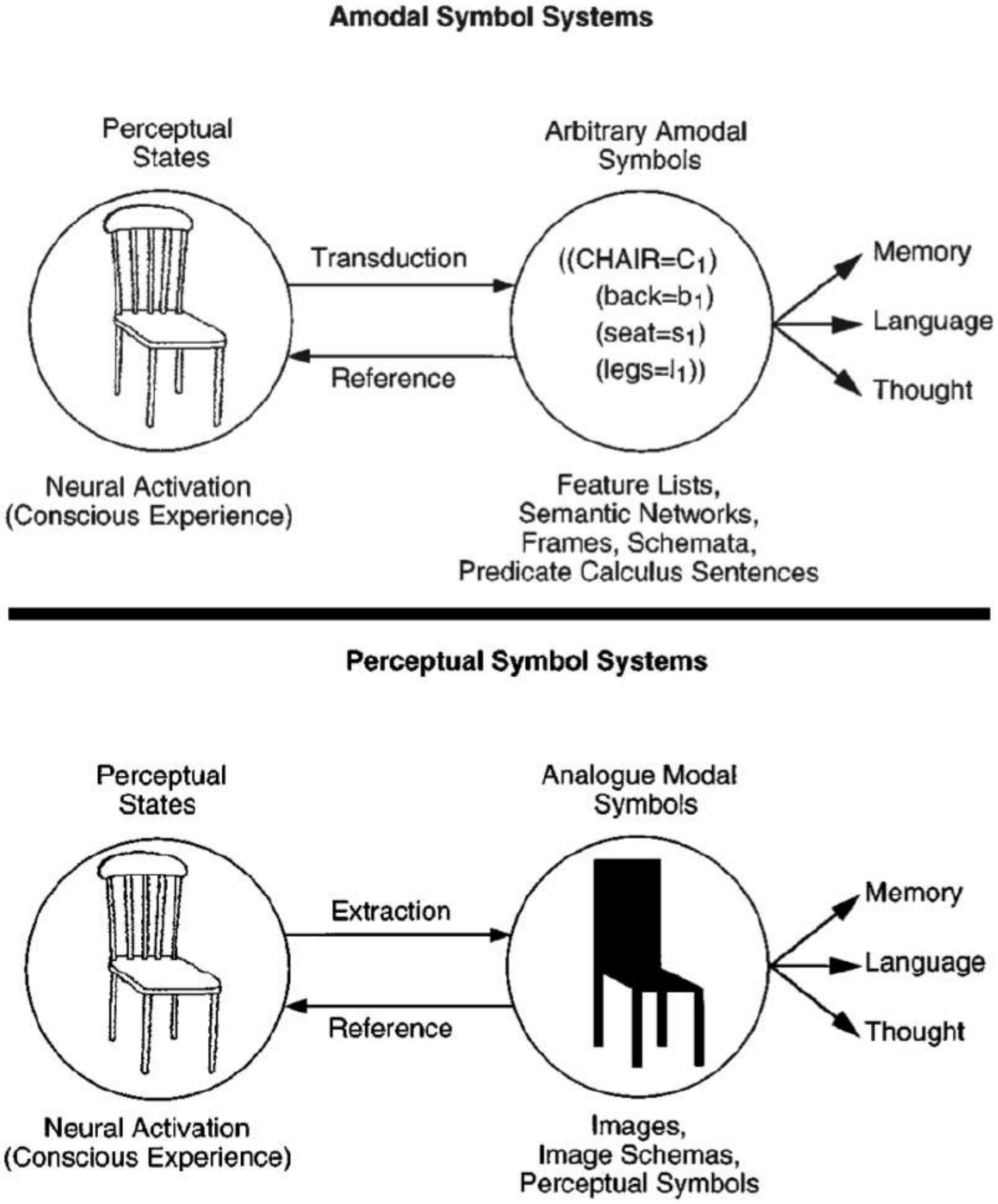

Amodal Symbol Systems
Perceptual States
Arbitrary Amodal Symbols
Transduction
Reference
((CHAIR=C1) (back=b1) (seat=s1) (legs=l1))
Memory
Language
Thought
Neural Activation (Conscious Experience)
Feature Lists, Semantic Networks, Frames, Schemata, Predicate Calculus Sentences
Perceptual Symbol Systems
Perceptual States
Analogue Modal Symbols
Extraction
Reference
Memory
Language
Thought
Neural Activation (Conscious Experience)
Images, Image Schemas, Perceptual Symbols

Figure 4.4: *The PSS hypothesis. Upper panel: the basic assumption underlying amodal symbol systems. Perceptual states are transduced into a completely new rep- resentational system that describes these states amodally. As a result, the internal structure of these symbols is unrelated to the perceptual states that produced them, with conventional associations establishing reference instead. Bottom panel: the basic assumption underlying perceptual symbol systems. Subsets of perceptual states in sensory-motor systems are extracted and stored in long-term memory to function as symbols. As a result, the internal structure of these symbols is modal, and they are analogically related to the perceptual states that produced them. Adapted from Barsalou (1999).*

According to Harris, the semantic similarity between two words is, in fact, a function of the degree of the similarity of their "linguistic environments", i.e. of the degree to which they can occur in similar contexts. For example, the near-synonymy between oculist and eye-doctor depends on the possibility to use these words interchangeably in most linguistic contexts . Harris inherits from Bloomfield the refusal of meaning as an *explanans* in linguistics. However, at the same time he reverses the direction of the methodological arrow, and claims that similarity in distributions should be instead taken as an *explanans* for meaning itself, and therefore used to build paradigmatic classes out of distributionally semantic similar linguistic expressions Lenci (2008).

This approach has become more prominent again, due to some high-profile computational models that use text association to specify the meaning of a word. These approaches may seem at odds with the conceptual view.

A general and potentially useful attempt to represent meaning came about with the development of scaling procedures in the 1960s, such as multidimensional scaling (MDS) and various clustering techniques. These procedures require information about the similarity of different words. This has most often been obtained by asking subjects to rate the similarity of pairs of words. For example, how similar are "horse" and "deer"? "horse" and "donkey"? "donkey" and "deer"? and so on. These similarities between all pairs of words being studied would be input into a program, which would output a structure. In the case of MDS the latter would be a similarity space.

When such techniques were developed, there was some hope that they could provide a more quantitative basis for representing word meaning. However, they also have severe limitations as representations of meaning. Solutions obtained simply do not represent much of the meaning. Word meaning is extremely complex, and no small set of dimensions (2 o 3) can account for how words are used.

Further, the solutions obtained depend greatly on the particular set of items tested. For example, when only mammals are considered that size and predacity are important dimensions. However, when plant-eating animals are tested, it is likely that predacity would no longer be a dimension, and some other difference among the animals would have become more salient and would have appeared in the solution.

More recent versions of associative models have gained great currency, especially within the computational modelling realm.

One seminal approach is Latent Semantic Analysis (LSA, Landauer and Dumais, 1997). By using very large text corpora (over millions of words), one can derive via LSA that certain words tend to occur together in a document and other words hardly ever co-occur. These patterns of co-occurrence, the matrix of word-document co-occurrences represent meaning: words with similar patterns have similar meanings. One can compute the dimensions of this latent space, via a reduced-rank Singular-value decomposition (SVD) in which the k largest singular values are retained so that the resulting reduced-dimension SVD representation is the best k-dimensional representation to the original matrix (in the least-squares sense). Each word then has a value on each latent dimension and word position (now a k-dimensional vector) in the latent space serves as the new kind of semantic indexing. The semantic similarity between two words can eventually be calculated as the dot-product or cosine distance between the two vectors.

State of the art in statistical semantics

From the original LSA model a number of associative approaches have flourished, that might be characterised as the field of "statistical semantics" (Sikström and Garcia, 2020).

Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) is the probabilistic version of LSA that models each word in a document as a sample from a mixture model of conditionally independent multinomial distributions. Each document consists of topics, and each topic consists of words. pLSI has an improvement over LSA in terms of the interpretability of word-topic and topic-document relations. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) was proposed to overcome the overfitting problem in pLSI by introducing the Dirichlet prior over the multinomial topic distribution. Interestingly enough

The basic rationale of this research program is the one that has motivated LSA. In a nutshell:

- Meaning is created by co-occurrences of concepts in the world (namely, the DH). These can be used to create semantic representations of concepts.
- Semantic representations can be generated by applying data-compression algorithms on co-occurrence of words in text corpora.

When we see a dog, we also see a tail, paws, eyes, legs, fur, and and so on. In this context, we may also see an owner, that goes for a walk in the park with the dog that is attached to a leach, and the dog might bark. Thus, dog is connected to other concepts, and when these concepts reliable co-occur with each other, then the meaning of dog as a concept is created (Sikström and Garcia, 2020).

In this perspective, a variety of text representation methods, and model designs have blossomed in the context of Natural Language Processing (NLP), including state of the art Language Models (LM). Here an LM basically means trying to predict the next word w_{t+1} given the previous words $w_t, w_{t-1}, \cdots, w_1$ in a sentence (in probabilistic terms, $P(w_{t+1} \mid w_t, w_{t-1}, \cdots, w_1)$)

While LSA takes a geometric approach to extracting meaning from co-occurence statistics, and LDA uses a probabilistic generative approach to the problem, recent models use a neural network approach to extract the semantic vectors.

These models are trained to either predict the most likely word given some context, or, in reverse direction, the most likely context for a given word. The word vectors that emerge as a side-effect of this prediction task, have turned out to be of much higher quality than the word vector from classical distributional semantics.

The origins of neural word embeddings can be traced back to the proposals of Hinton et al. (1986) and Bengio et al. (2003), who used neural architectures in the derivation of word meanings.

Neural word embeddings can transform large volumes of text into effective vector representations capturing the same semantic information. Further, such representations can be utilized by various machine learning (ML) algorithms for a variety of NLP-related tasks. Based on the association hypothesis, text representation methods have evolved from manually selecting the features called feature engineering to representational learning methods that leverage deep neural networks to discover relevant embeddings, most notable example being Word2vec (Mikolov et al., 2013) The resulting representation is encoded with word meaning, so similar words will have similar vector representations, for example, the vector('car') will be similar to the vector('driver'). Moreover, the relationship between words is also preserved in term of displacement between points such that basic vector operations on these points will be meaningful, for example, vector('Paris') - vector('France') + vector('Italy') will result in a vector very similar to the vector('Rome'). Also, the displacement can capture syntactic relations, for instance, vector('sweet') - vector('sweetest') + vector('fastest') will result in a vector very similar to the vector('fast')

For an in-depth, comprehensive analysis and survey, which is out of the scope of this thesis, the reader is urged to refer to Naseem et al. (2021). It will suffice to say that, based on the idea of continuous word embedding in which text from the corpus is mapped as vectors under the distributional hypothesis, such methods have evolved from non-contextual embeddings (e.g., Word2vec, GloVe, FastText), to contextual word representations (e.g., Context2Vec, CoVe, and ELMo, a weakly bi-directional model) based on LSTMs / Recurrent approaches in deep learning. Continuous word representation models have drastically improved text classification results (Naseem et al., 2021). The next step has been represented by truly bi-directional contextual representations relying on deep learning transformer architectures such as GPT-OpenAI Transformer, BERT and variants (XLNet, RoBERTa, ALBERT, DistilBERT, MegatronLM, ERNIE, SpanBERT, BART, etc.) (Naseem et al., 2021).

The recent neural word embedding models built on those successes, and moreover showed that automatically learned word representations encode many linguistically relevant relations be- tween words. But distributional and neural word vectors have little to say about how sentence meanings can be constructed (Repplinger et al., 2018). On the other hand, The symbolic models of language in the tradition of Montague emphasized the compositional semantics of sentences, but largely ignored how the meaning of words can be modelled and relied on hand-built grammars specifying the syntactic and semantic properties of words. In the last few years, much research is devoted to bringing together insights about compositionality from the symbolic tradition, and insights from vector-space models of word meaning from the distributional and neural traditions, giving rise to compositional distributional semantics (Repplinger et al., 2018).

The seminal work by Mitchell and Lapata (2010) has shown how to use different types of composition functions , such as additive, multiplicative and tensor-based composition, that can be used in vector space models to compose sentence meaning from the meaning of smaller units, such as words.

Current research on compositional distributional semantics exists in two flavors. One class of models, type-based tensor approaches, combines a powerful compositional mechanism with a robust distributional foundation of word meaning at the cost of very high computational complexity. The other class of models consists of deep neural network architectures that implic- itly or explicitly—account for the demands of semantic compositionality (Repplinger et al., 2018).

As technological advances have emerged over time, these have been globally used in many domains such as medicine, social sciences, healthcare, psychology, law, engineering, and so on. In particular, these LMs have been used in many different areas of text classification tasks such as Information Retrieval, Sentiment Analysis, Recommender Systems, Summarization, Question Answering, Machine Translation, Named Entity Recognition, Adversarial Attacks and Defenses, and so on (Naseem et al., 2021).

4.1.5 Final remarks on computational models

The success of statistical semantics models could be taken as undermining the conceptual approach to meaning taken in this chapter. If these models can truly account for how words are used, then word meaning may be a large associative network that merely encodes patterns of co-occurrence.

However, the general problem with nets of associations is that knowing what words are associated to one another does not specify what the meaning of an individual word is. Words must be connected to our knowledge of things in the world, not just other words. Although “jugs” may be related to both “vinegar” and “bottles”, one would not know from the scores that “Put it in the jugs” is similar to “Put it in the bottles” but not “Put it in the vinegar.”

Since concepts are our non-linguistic representation of the world, by connecting words to these representations, we can explain how people can connect sentences and words to objects and events in the world (in the vein outlined in Figure 4.2). Concepts are just the things that are evoked by our perceptual systems and that control our actions. Thus, by hooking up words to concepts, we can break out of the circle of words connected to words and tie language to perception and action. And, as we will see, to emotions.

This challenge is vividly present when the problem of word (and language) acquisition in children is addressed: a question that lies at the very heart of cognitive science. In a sense, modern NLP models, such as BERT, ELMo and GPT-3, are already close to simulating children’s language acquisition. So what is missing? The main thing they currently lack is exactly a real world representation of semantics that allows them to map from form to meaning and vice-versa, with “meanings” represented solely as contextualized word embeddings. Indeed, if our goal is to translate from one natural language to another, to develop a predictive-text application, or to generate passages of text given a prompt (e.g., GPT-3), contextualized word embeddings will probably do a better job. But if our goal is to simulate children’s language acquisition, we have develop real-world semantic representations in the sense discussed above. Again, this fundamental problem is one of the most challenging and intriguing issues in current robotics Taniguchi et al., 2016; Cangelosi and Stramandinoli, 2018.

Yet, in the current computational modelling practice, markedly in robotics the story is much more nuanced than it would suggest the in principle stark contrast between statistical semantics and conceptual/grounded approaches.

For instance, in Taniguchi et al. (2016) one can find a detailed discussion of how, starting from a paradigmatic statistical semantics tool such as LDA, one can extend the original model by incorporating multi-modal perceptual information (vision, touch, speech) to a multi-modal LDA thus providing an appropriate grounding.

For instance, pairing words with “words” that represent object features and that are taught by humans along human-robot interaction, the robot can acquire word meanings. In turn, the robot can recall the multimodal information that can be represented by the word, much like Anderson suggested in his RMC. Thus, we may say, at the light of previous discussion, that the robot has formed an object concept and learned speech recognition similarly to infants by using multi-modal information obtained from objects and teaching utterances given by humans. Moreover, by connecting the recognized words and concepts, the robot can also acquire word meanings To sum up, the robot is considered to have “understood” word meanings through its own body.

Beyond robotics, a similar avenue has been taken in those research fields lying at the intersection between Computer Vision and NLP, such as image captioning, video captioning, visual question answering, visual retrieval (Wiriyathammabhum et al., 2016).

To sum up the whole discussion, whatever the modelling approach is chosen, it should be able to account for the following fundamental characteristics of the human symbol system (Taniguchi et al., 2016):

Grounded : A symbol does not have any meaning without being grounded or interpreted.

Dynamic : There does not exist an objectively true symbol system that can be determined in top-down manner in our human society.

Social : An individual representation system and the socially shared symbol system are not same.

But, as to the problem of the social grounding of meaning, we need to jump from semantics to pragmatics, which will be the topic of the next Section.

Interestingly enough, this view, is not a novel one. It was already present in Augustine (1876) “Confessions”, where he describes how he learnt to use language (and acknowledged by Wittgenstein, 2009 at the beginning of his “Philosophical Investigations”):

> When grown-ups named some object and at the same time turned towards it, I perceived this, and I grasped that the thing was signified by the sound they uttered, since they meant to point it out. This, however, I gathered from their gestures, the natural language of all peoples, the language that by means of facial expression and the play of eyes, of the movements of the limbs and the tone of voice, indicates the affections of the soul when it desires, or clings to, or rejects, or recoils from, something. In this way, little by little, I learnt to understand what things the words, which I heard uttered in their respective places in various sentences, signified. And once I got my tongue around these signs, I used them to express my wishes

4.2 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 4.5 (though, in many practical cases such distinction tends to blur)

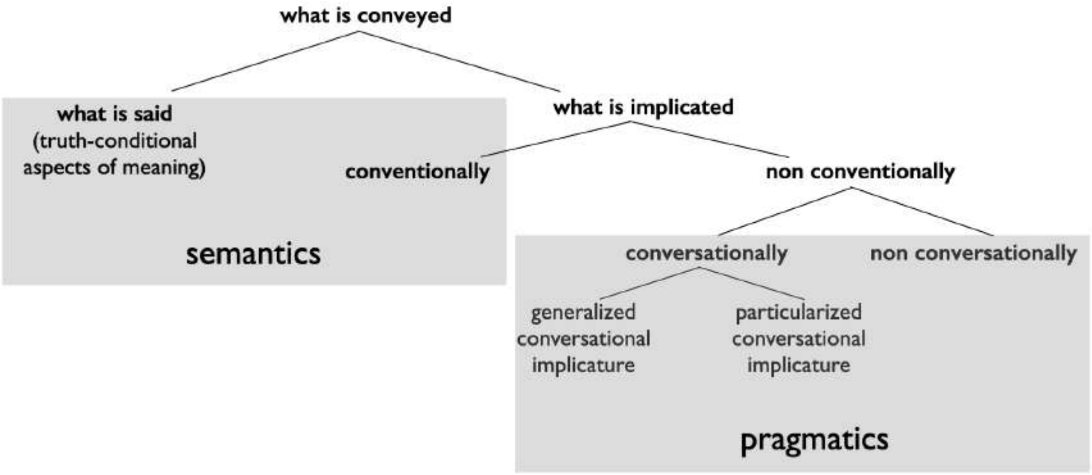


Figure 4.5: *Kinds of meaning that are in principle addressed by semantics and pragmatics. These include both conventional and conversational (to be inferred) aspects of meaning. Adapted from Terkourafi (2021).*

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 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).

4.2.1 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 utilizing 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 systematized and somewhat formalized 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 recognize 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.

4.2.2 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 recognize 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.

4.2.3 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. 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.

4.2.4 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 recognized 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.

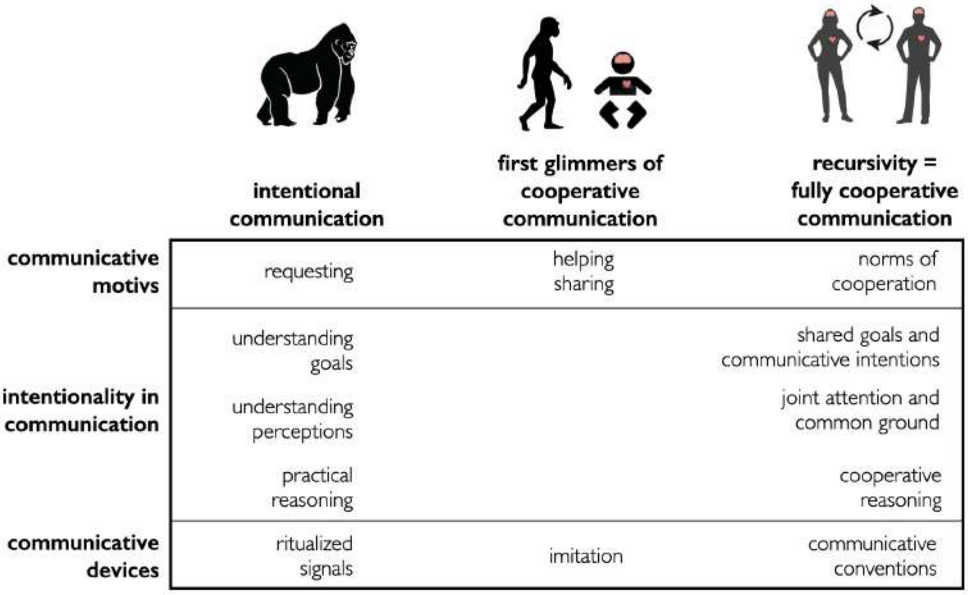
The story of how it happened is long and complicated. Thus, for our purposes, it will suffice to draw on Tomasello (2010) account.

Based on a vast literature on developmental psychology and ethology, Tomasello (2010) characterizes 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.



	intentional communication	first glimmers of cooperative communication	recursivity = fully cooperative communication
communicative motivs	requesting	helping sharing	norms of cooperation
intentionality in communication	understanding goals understanding perceptions practical reasoning		shared goals and communicative intentions joint attention and common ground cooperative reasoning
communicative devices	ritualized signals	imitation	communicative conventions

Figure 4.6: *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; the latter stage is the one previously detailed in Figure 4.1 at the beginning of this chapter. Adapted from Tomasello (2010).*

The main steps of such evolutionary development are outlined at a glance in Figure 4.6.

The road to human cooperative communication begins with great ape intentional communication. Intentional signals allow communicators for attempting to influence 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 (Figure 4.7).

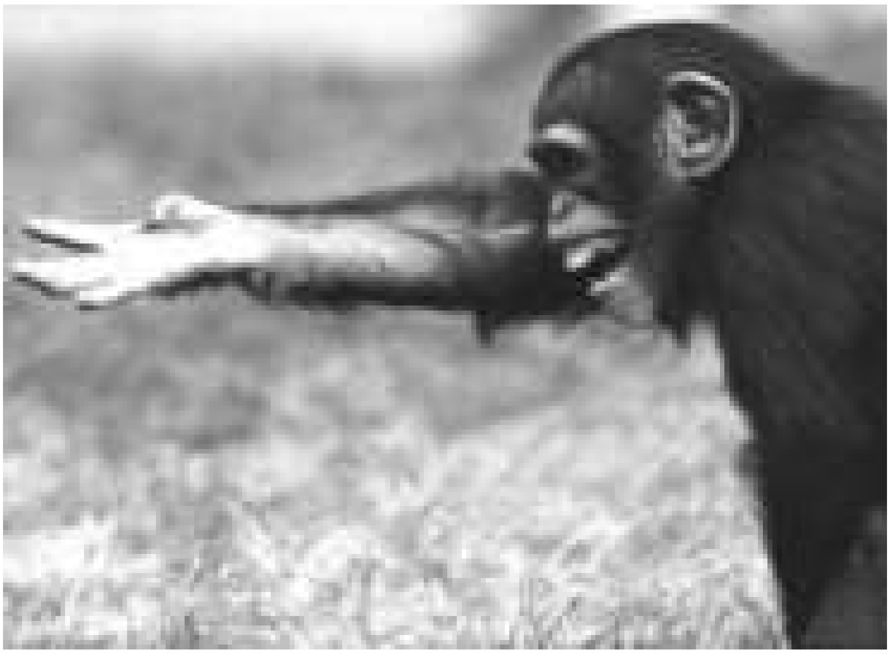


Figure 4.7: *A juvenile chimpanzee tries to reclaim food that a dominant has taken away by combining the reach out up begging gesture with a scream vocalization. Adapted from Pollick and De Waal (2007).*

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 4.7).

There is some debate on whether apes do have a true ToM (the ability to recognize 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 unconventionalized, 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 hypothesized 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.

4.2.5 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 to the purpose of decoding dialogue acts. They see interpretation as a classification task, and solve it by training statistical classifiers on labeled 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 characterized 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 → 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 is 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).

The Rational Speech Act model

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 (cfr., Figure 4.1). 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 synthesizes 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, in the precise sense described in Chapter 3.

Also, it is suitable to capture and formalize 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 de- scribe 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.

CHAPTER 5

The conceptual act

TURNING back to the roadmap outlined in Section 3.2, we have discussed so far how a speaker can perform a communicative act by uttering a sentence/word to convey some meaning and how the listener, by hearing the utterance, might reason about the meaning and the states of the world.

This is possible *prima facie* in reason of the common language and lexicon between the speaker and the listener, but more deeply because the speaker and the listener share the Shakespearean fate of being "made of the same stuff". Under a common neurobiological/cognitive infrastructure and socio-cultural context, both agents share the capability of performing a *conceptual act*, namely the inference $P(\mathcal{C}(\texttt{world})|\mathcal{O}(\texttt{world}))$, to perceive and conceptualize the states of the world, given a collection of events or *outcomes* $\mathcal{O}(\texttt{world})$.

The dividing line between our brain and world is permeable, perhaps nonexistent. The brain's core systems combine in various ways to construct perceptions, memories, thoughts, feelings, and other mental states.

A brain is constantly faced with continuous sensory inputs such as dynamically changing wavelengths of light, air pressure, chemical concentrations, and so on, which are noisy and ambiguous in their meaning. Yet, an agent does not perceive the world as a continuum of "blooming, buzzing confusion", as William James noted (James, 1890), but as an orderly world of discrete objects. And concepts pave the way to achieve such order. Everything that is perceived is represented by concepts in the brain.

Infants awash in sensory input, there is a point in our ontogeny (cfr. Figure 4.7 and Figure 5.1, which expanding on it), where the outside world seeds our earliest concepts, as our brain hardwires itself to the realities of the physical world (Hoemann et al., 2020b, 2019).

As brain develops and we begin learning words, we connect to the social world,

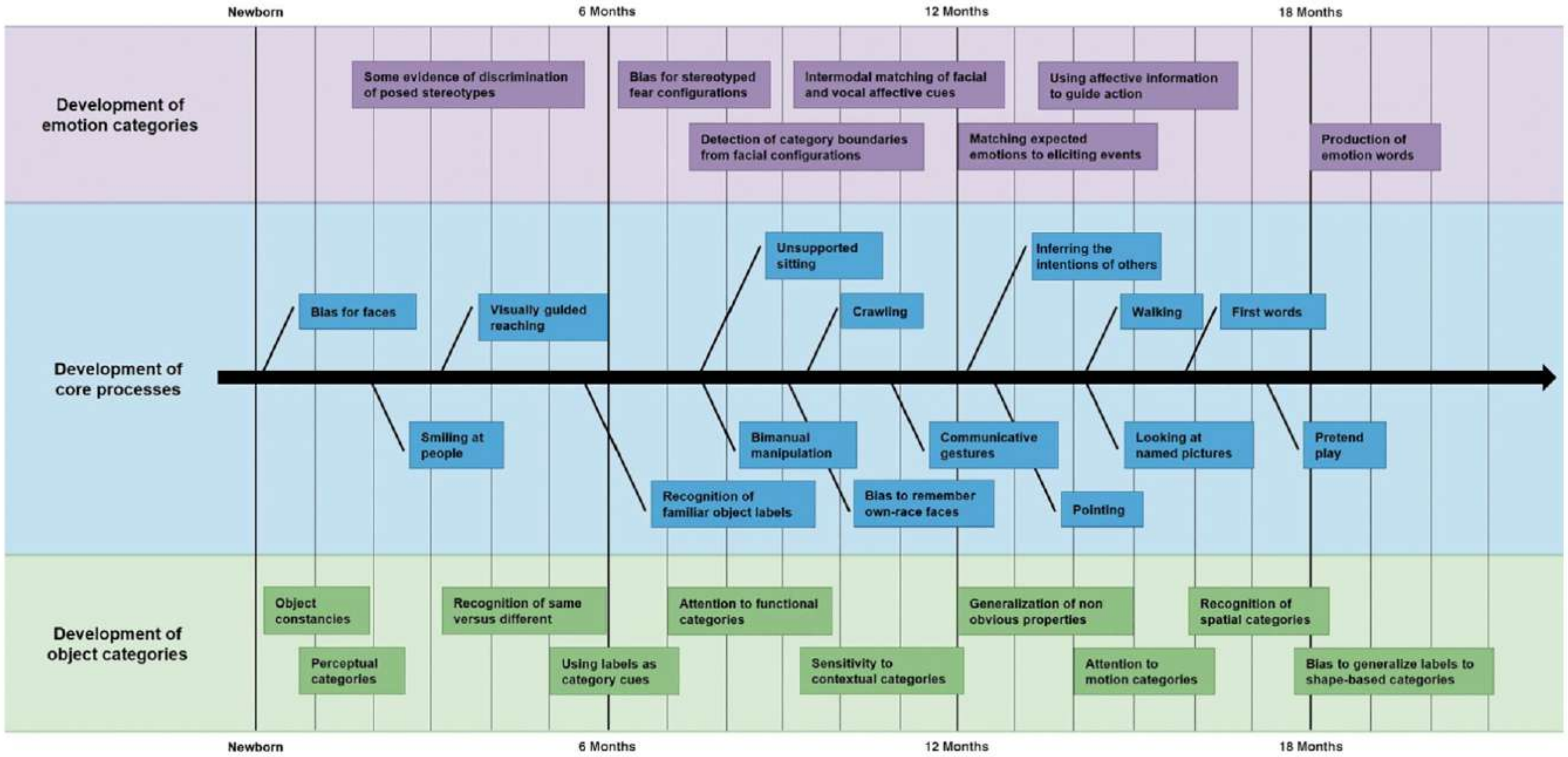

Newborn
6 Months
12 Months
18 Months
Development of emotion categories
Some evidence of discrimination of posed stereotypes
Bias for stereotyped fear configurations
Intermodal matching of facial and vocal affective cues
Using affective information to guide action
Detection of category boundaries from facial configurations
Matching expected emotions to eliciting events
Production of emotion words
Development of core processes
Unsupported sitting
Inferring the intentions of others
Bias for faces
Visually guided reaching
Crawling
Walking
First words
Smiling at people
Bimanual manipulation
Communicative gestures
Looking at named pictures
Pretend play
Recognition of familiar object labels
Bias to remember own-race faces
Pointing
Development of object categories
Object constancies
Recognition of same versus different
Attention to functional categories
Generalization of non obvious properties
Recognition of spatial categories
Perceptual categories
Using labels as category cues
Sensitivity to contextual categories
Attention to motion categories
Bias to generalize labels to shape-based categories
Newborn
6 Months
12 Months
18 Months

Figure 5.1: *Hoemann et al. (2020b) hypothesis of developmental timelines. The core processes (depicted in blue) belong to developmental cascades that might contribute to the development of object categorization (in green) and emotion categorization (in purple). At each point in time, multiple abilities are emerging: for example, at 9 months, infants may be crawling, are sensitive to contextual cues and functional inferences, and can match affective facial and vocal signals. Each of these reflects development within their own domain , but may also reflect potential cascades across rows. For example, the limitations in the newborn visual system, which bias infants to look more at faces than at other stimuli, influence their developing perceptual abilities. Similarly, the increased ability of infants to interact with the world through visual-manual exploration provides them with opportunities to learn new properties of objects, as well as to learn facial expressions, vocalizations, and other cues related to emotion during object explorations with caregivers and others. Adapted from Hoemann et al. (2020b).*

and through cooperative communication we begin creating purely mental concepts. Concepts from our culture appear to be in the outside world, but they are constructions of our conceptual system. Brains grow socially (Atzil et al., 2018): culture helps to wire our brain, and we behave in certain ways that wire the brains of the next generation. As Barrett (2017b) puts it, "It takes more than one brain to create a mind".

What we have learned from the discussion given in Section 4 is that concepts are *grounded*, *dynamic* and *social*. Further, we have seen that a rational analysis of concept formation, in terms of a Bayesian setting Anderson (1991), can be given and meanwhile extended in order to be grounded in sensorimotor interactions. Here we take a step further.

The literature we have so far discussed mostly concerns with concept formation from perception, but limited to external perceptions as gathered through sight hearing and touch, or, at least in the case of robotics, the proprioception of self-movement and body position.

However, along with exteroceptive processing, interoceptive processing give birth to agents' feelings of affect, and influences every action the agents performs, speech acts being no exception. Interoception here denotes the sensory data that collectively describe the constantly changing physiological state of the body, arising from the allostatic regulation of various bodily systems, including the autonomic nervous system, the endocrine system, and the immune system (Barrett, 2017c).

Interoception enables the agent's brain to construct the environment in which the agent lives and, eventually, to give meaning to words. Deprived of interoception, without affect and feelings, the agent would be unlikely to survive for long (Barrett, 2017b).

Yet, at a more fundamental level, the human brain did not evolve to think or feel or see, but to efficiently maintain energy regulation in the body (namely, allostasis, Schulkin and Sterling, 2019; Sterling, 2012). Energy regulation (e.g, metabolism) is likely to be at the core of the human mind, regardless of whether a person is thinking, feeling or perceiving.

To such end, brains do not react to the world, but instead predict and then test their hypotheses against incoming sensory evidence. Their hypotheses constitute internal models of the body in the world. Brain's internal model consists of embodied, whole brain representations that predict what is about to happen in the external environment and the best course of action for dealing with these impending events. Allostasis itself can thus be defined in terms of prediction: a brain maintains energy regulation by anticipating the body's needs and preparing to satisfy those needs before they arise Schulkin and Sterling, 2019; Sterling, 2012.

In brief, brain's internal model runs on concepts. On the one hand, the brain transform sensory inputs from the body and the world in the context of a concept. On the other hand, conceptual representations are tested against the incoming sensory evidence to categorize incoming sensory signals according to past experience (Barrett, 2017a). The resulting categorization enables allostasis, allowing the brain to efficiently predict energy expenditures for motor actions, as well as the benefits that will result.

This idea of the brain as a predictive machine (the Bayesian brain) has gained currency in cognitive and theoretical neuroscience in contrast to traditional stimulus-response "feedforward" frameworks (see, Vilares and Kording, 2011; De Ridder et al., 2014; Aitchison and Lengyel, 2017; Chater et al., 2020; Colombo and Seriès, 2020;

Yon and Frith, 2021; Marino, 2020, for general reviews and problems). According to these theories, the Bayesian brain can be conceptualized as a probability machine that constantly makes predictions about the world and then updates them based on what it receives from the senses (De Ridder et al., 2014 and see Appendix B for a thorough introduction).

Whatever the brain might be doing — thinking, seeing, tasting — it is also predictively regulating the body's physiological systems in the service of allostasis. In every waking moment, the brain gives sensations, either exteroceptive or interoceptive, meaning.

Under such circumstances, there is no specific difference between emotion, vision or audition. When, we focus on some of those sensations that are interoceptive ones, the resulting meaning can be an instance of emotion (Barrett, 2017c).

The interoceptive network issues predictions about the body, tests the resulting simulations against sensory input from the body, and updates the brain's model of the body in the world. Interoceptive sensations are routinely experienced as lower dimensional feelings of affect (valence and arousal, Barrett and Russell, 2014). If interoception plays a role in allostasis, and allostasis is at the core of the brain's computational architecture, then the properties of affect — valence and arousal — are best thought of as basic features of consciousness, rather than properties of emotion *per se* (Barrett and Satpute, 2019).

In this view (originally called the Conceptual Act Theory, CAT, Barrett et al., 2015), emotions are constructed concepts, and words like "happiness" and "fear" labels of the related categories, much like the word "red" stands for the categorization of the perception of a visual stimulus occurring within a certain range of visible light wavelengths (Hoemann et al., 2019; Barrett, 2017b).

The physical changes in the natural world (internal physical changes occurring within a perceiving agent, and sensory changes from the world such as from other people's facial muscle movements, actions, the physical surroundings, etc.) become real as emotion (as fear, anger, etc.) when they are categorized as such using emotion concept knowledge within the perceiver. These concepts have been learned from language, socialization, and other cultural artifacts within the person's day-to-day experience.

Indeed, language and emotions are tightly intertwined (Lindquist et al., 2015a). On the one hand, language has a constitutive role in emotion. Infants and children learn emotion categories the way they learn other abstract conceptual categories – by observing others use the same emotion word to label highly variable events.

On the other hand, language communicates emotional states. Whether spoken or written, words allow the members of a culture to dynamically re-establish category boundaries by labeling instances, reinforcing the social reality that they create. Through dialog during social interaction, people come to use the same words to categorize objects, actions, and events, progressively aligning the associated concepts. This may be one mechanism by which people communicate emotion and help to co-construct each other's emotional experiences.

Eventually, emotions ground the communicative act, and, ostensively or not, contribute to convey some intended affective meaning that the listener might infer from speaker's utterance or from the speaker's non-verbal behaviour as a part of the world state. Emotional expressions can be rich communicative devices: they do much more than simply expressing emotions (Scarantino, 2017). Bodily displays are sophisticated social tools that can communicate the signaler's "intentions" and "requests". Emotional expressions are a means not only of expressing what's inside but also of directing other people's behavior, of representing what the world is like and of committing to future courses of action (Scarantino, 2017). In some cases, it is possible to engage in analogs of speech acts without using language at all.

Yet, emotion categories, need to be made real through shared intentionality. A speaker to communicate to a listener that he feels happy, needs a shared understanding of "happiness" with the listener.

Eventually, putting feelings into words (also known as "affect labelling") can attenuate individual emotional experiences. Research investigating affect labeling has found it produces a pattern of effects like those seen during explicit emotion regulation, suggesting affect labeling is a form of implicit emotion regulation (Torre and Lieberman, 2018).

In the following we shall give a precise form to the above observations and assumptions.

5.1 Word and concepts reloaded: how concepts shape perception

We shall first review categories: they are important because they determine how we see and act upon the world (Harnad, 2003); emotions at the highest level will be defined as categories ("fear", "anger", etc.)

We have previously defined a concept as a mental representation of a category. A category, or *kind*, is a set of things (cfr. Section 4.1).

Membership in the category may be all-or-none (e.g., "apple" or a matter of degree (e.g., "small"). In the latter case, the category is said to be continuous. Concrete sensorimotor categories (items of which can be seen and touched), are a mixture of the two: categorical at an everyday level of magnification, but continuous at a more microscopic level (Harnad, 2003).

An interesting example is represented by color categories : central reds are clearly reds, and not shades of yellow. Yet, in the orange region of the spectral continuum, red/yellow is a matter of degree; context and contrast effects can also move these regions around somewhat. An example is provided in Figure 5.2

When the category of "red" is joined with the concept of "apple" to form the "red apple" category, the population of red items becomes even looser (cfr. Figure 5.3).

5.1.1 Categorical perception

According to the "Whorf Hypothesis" (the *linguistic relativity* hypothesis, Whorf, 1964), colors are determined by how our culture and language happens to subdivide the spectrum: colors are perceived categorically only because they happen to be named categorically. Whorf famously argued that, to an Eskimo, it would be unthinkable to use the same word for all types of snow because of its wide range of types and different uses.

In brief, for the brain to convert a visual sensation into the experience of red, it must possess the concept "red." This concept may come from prior experience with apples,

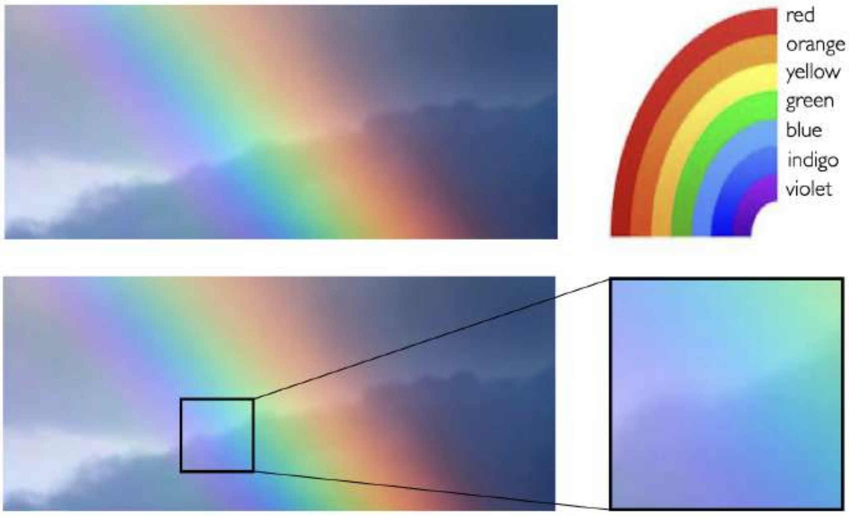


Figure 5.2: *When looking at a rainbow, we perceive discrete bands of color. However in nature, a rainbow has no stripes — it's a continuous spectrum of light, with wavelengths that range from approximately* 400 *to* 750 *nanometers. This is one example of categorical perception where mental concepts for colors like "red," "orange," "yellow" implies a reorganization of incoming stimuli within a common structure, here the color bands*

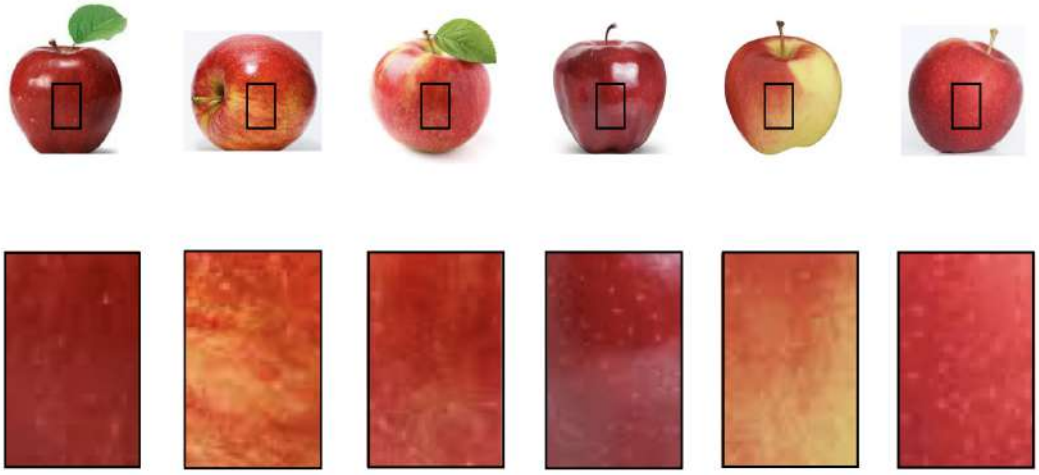


Figure 5.3: *Apple images retrieved on the fly by searching "red apple" on Google images. Under this goal, a red apple is a compound concept/category (apple + red) whose exemplars have a high statistical variation and rely at best upon a "loose",* ad hoc *concept of red*

roses, and other objects one perceives as red, or from learning about red from other people. Without this concept, the apple would be experienced differently.

The study by Davidoff et al. (1999), for instance, has shown that to the Berinmo people of Papua New Guinea, a stone-age culture, apples reflecting light at 600 nanometers are experienced as brownish, because Berinmo concepts for color divide up the continuous spectrum differently.

Categorical effects are found across speech sound categories, with the degree of these effects ranging from extremely strong categorical perception in consonants to nearly continuous perception in vowels (Feldman et al., 2009; Kronrod et al., 2016).

Similar patterns have been observed in the representation of objects belonging to artificial categories that are learned over the course of an experiment as well as in the perception of facial expressions such as facial expressions of emotion in stimuli constructed from line drawings, photograph-quality stimuli. Stimuli for these experiments were drawn from morphed continua in which the endpoints were prototypical facial expressions (e.g., happiness, fear, anger) (Feldman et al., 2009). With few exceptions, results showed discrimination peaks at the same locations as identification boundaries between these prototypical expressions. Figure 5.4 qualitatively illustrates such effect on a continuum of facial expressions, from neutral to angry.

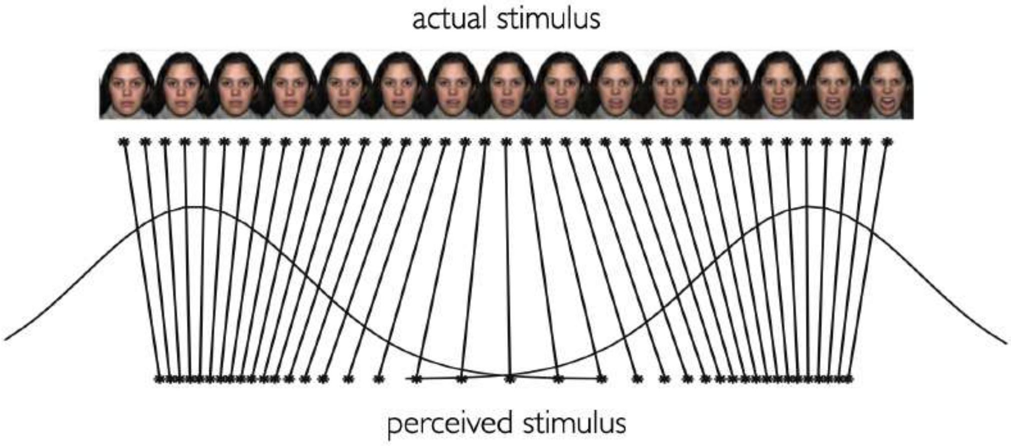


Figure 5.4: *Warping from actual (a continuum of facial expression from neutral to angry, on the top) to perceived stimuli on the bottom is shown in the dispersion of the vertical bars toward category centers (neutral and angry faces) acting as "perceptual magnets" (Feldman et al., 2009). The Gaussian-like curves displayed over the bars, qualitatively represent the distributions of the two target categories that cluster the actual stimuli in the perceived stimuli around the distribution means.*

All of these categorical effects are characterized by better discrimination of between-category contrasts than within-category contrasts (a sort of perceptual "magnet effect"), although the magnitude of the effect varies between domains.

In speech perception, categorical effects in speech perception are typically studied

]

through behavioral identification and discrimination tasks, which provide data on listeners' ability to classify the sounds (identification) and to differentiate sounds along an acoustic continuum (discrimination). The stimuli that participants hear in each task typically lie along a one-dimensional continuum between two phonemes. As a simple example, consider a continuum between two phonemes (categories), $C = c_1$ and $C = c_2$, with seven equally spaced stimuli $S_1 \cdots S_7$. For example, if c_1 = /b/ and c_2 = /p/, stimuli might be created by varying the voice onset time (VOT) of the signal. The identification task consists of choosing between two competing labels, c_1 and c_2, in a forced choice paradigm. Participants choose one of the two labels for every stimulus heard, even if they are unsure of the proper classification. By examining the frequency with which participants choose each category, one can observe an apparent boundary between the categories and can determine the sharpness of this boundary (Kronrod et al., 2016).

In the tradition of rational analysis Feldman et al. (2009) considered the abstract computational problem posed by speech perception as described above. The theoretical model proposed is a Bayesian generative model which, for its relevance, we analyse to some extent.

The model begins with the listener's knowledge of the two categories, $C = c_1$ and $C = c_2$. The next steps concern the process that the listener presumes to have generated the sounds heard.

The speaker chooses one of the two phonetic categories. Categories can be represented in our model as a distribution $P(C \mid \theta_c)$ around the category mean μ_c with variance σ_c, $\theta_c = \{\mu_c, \sigma_c\}$ (e.g., in their model a Gaussian distribution is used). Formally, the choice can be written as the sampling step

$$c \sim P(C \mid \theta_c)$$

It is assumed that parameters θ_c of the category being used in the generative procedure by the speaker are known by the listener from previous exposure to sounds from this category in the language.

The next step in the generative process is the selection of an intended target production from the previous distribution. Denote T the intended target production. Then,

$$t \sim P(T \mid C = c, \theta_t).$$

Once the intended target production t is chosen, it needs to be articulated by the speaker and perceived by the listener. This process introduces additional articulatory, acoustic, and perceptual noise that distorts the signal. Denote S the possible speech sounds related to the underlying category chosen by the speaker at the beginning of the generative procedure. Hence, sound emission is sampled as:

$$s \sim P(S \mid T = t, \theta_s).$$

The ratio between the category variance and the total variance provides a measure of warping from the actual to the perceived stimulus, that is the dispersion of the vertical bars toward category centers as illustrated in Figure 5.4.

This simple model is important beyond the specific application to categorical speech perception. On the one hand, it can be formally considered as a simplified version of the Anderson RCM we have discussed in Section 4.

On the other hand, in its simplicity, it can be exploited to describe at a more general level how Perceptual Categorization works in more complex perceptual cases. Consider, for instance, the famous dalmatian dog example recapped in Figure 5.5.

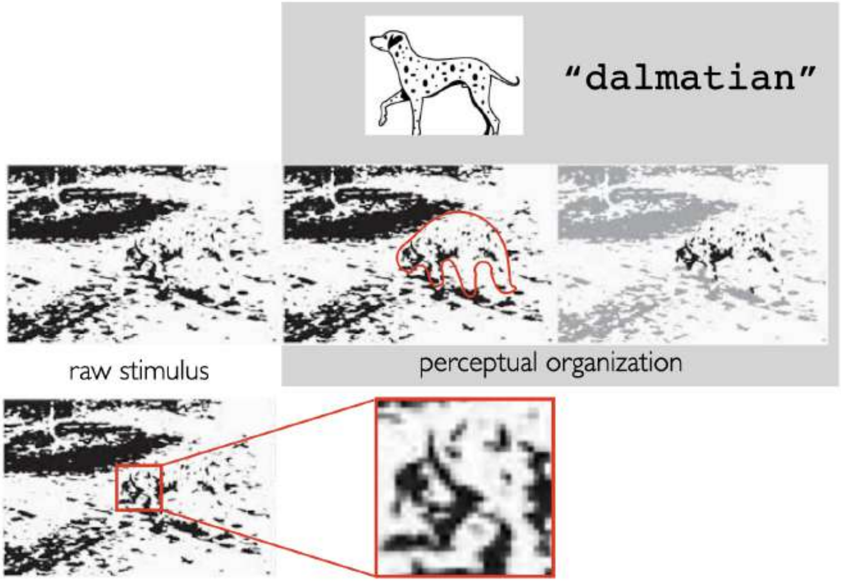


Figure 5.5: *The famous dalmatian dog. When viewing the first time this stimulus, a bare ensemble of blobs, the brain is working hard to make sense of them. Neurons in the visual cortex are processing lines and edges. Sub-cortically, the amygdala is firing rapidly because the input is novel. Other brain regions are sifting through past experiences to determine if any input like this has been encountered before and are conversing with the body to prepare it for an as-yet-undetermined action in a state of experiential blindness. Then, by scrutinizing the red silhouette and coming back to the raw stimulus, blobs are no longer perceived as formless, but a familiar object takes form. The brain has exploited its vast array of prior experiences and constructed the familiar object you now see in the blobs. Neurons in the visual cortex changed their firing to create lines that are not present, linking the blobs into a shape of a dalmatian that is not physically there. The brain changed the firing of its own sensory neurons in the absence of incoming sensory input, performing a simulation of a likely shape under the guide of concepts previously constructed. When the process has reached equilibrium, it is virtually impossible to "unsee" the object*

This is a difficult case in perception, but the perceptual effect experienced when viewing the picture can be suitably explained in terms of the generative model presented above and in terms of predictive activity of the brain. The result can be summarised in the semiotic triangle presented in Figure 5.6.

In a Bayesian generative/predictive account, we might re-appraise the Peircean dynamic process of interpretation as follows:

- a category representing items is labelled through a word, e.g. "dalmatian"; contextual and background knowledge on the population of events/objects that defines the category can be shaped in terms of the prior probability $P(C)$
- the interpretant relies on the concept that mentally represents such category, which

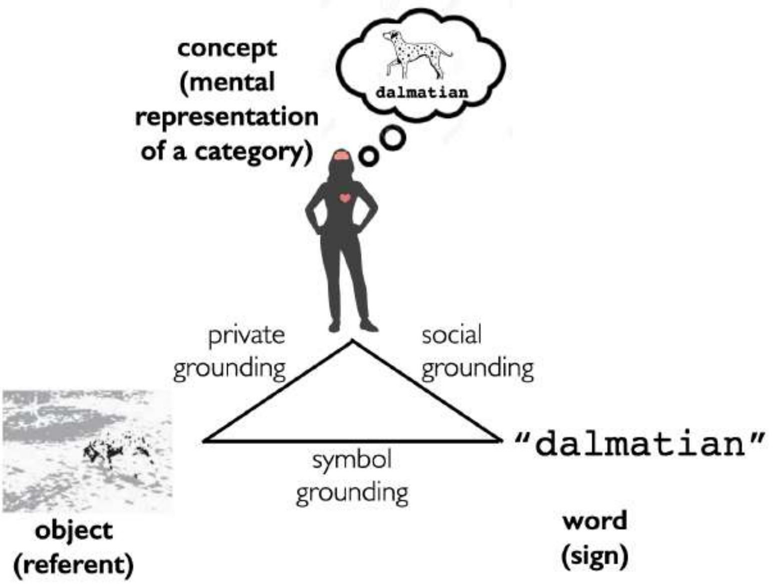


Figure 5.6: *The dalmatian dog experience summarised in the semiotic triangle. At a point, the interpretant is consciously aware of perceiving an dog-like object, which she names "dalmatian", though the entire process of construction is invisible to her. The object itself, the referent, is not a static entity but the result of making sense of the "blooming, buzzing confusion" characterizing the physical world (in this case, meaningless blobs).*

can be defined as predictive model (a dynamic process of interpretation) that mediates between the sign/word and the object / stimulus S: conceptualization amounts to the backward inference $S \rightarrow C$ computed via the posterior distribution $P(C \mid S)$ given the sensory data S; in turn, prediction/simulation of sensory data can be achieved by running the model in a forward mode $C \rightarrow S$ via the conditional probability $P(S \mid C)$ apt to generate the most likely stimulus S given the category C

The details of how this process is actually instantiated depend on the implementation model chosen to approximate optimal Bayesian calculations: predictive coding, active inference, probability coding, direct variable coding, sampling and so on (Aitchison and Lengyel, 2017; Sanborn and Chater, 2016).

Yet, beyond the apparent simplicity of this explanation, one should be aware that the interplay between bottom-up features that guide attention toward a target, and top-down conceptually driven simulation that generates hypothetical shapes can be subtle, as demonstrated by van Tonder and Ejima (2000) in a well-known experiment.

The model can naturally provide a simple and elegant account for these effects as depicted in Figure 5.8, which also outlines at a glance the model structure in terms of the underlying PGM.

To sum up, categorical perception is not just categorization. It implies a reorganization of representations within a common structure that allows a more pronounced boundary between exemplars from one category and exemplars from an otherwise perceptually adjacent category.

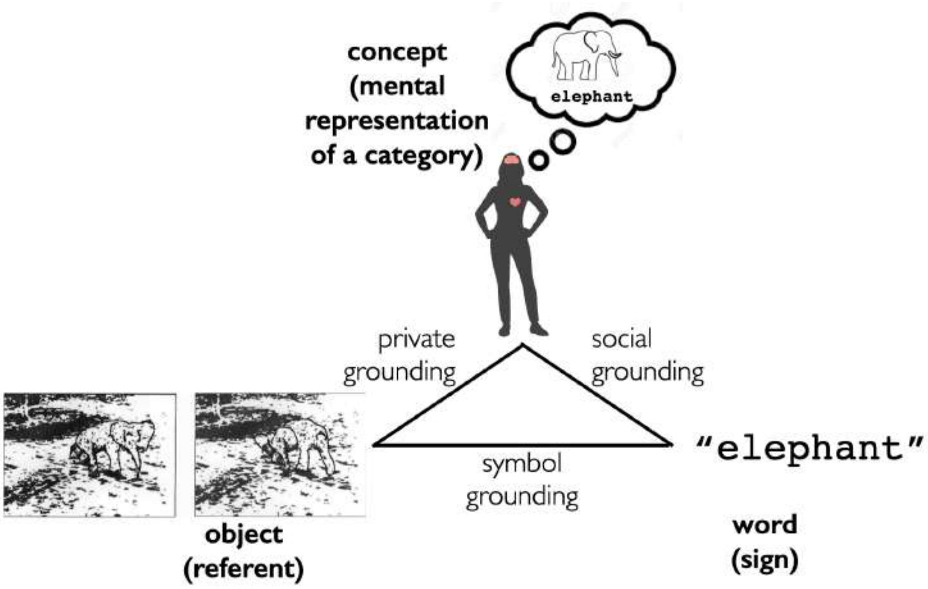


Figure 5.7: *The interplay between bottom-up features that guide attention toward a target, and top-down conceptually driven simulation that generates hypothetical shapes can be subtle. In a famous experiment, van Tonder and Ejima (2000) by cleverly manipulating bottom-up features strongly related to surface interpolation, found that many subjects assigned incorrect head and limbs to the hypothetical body and ended in recovering different objects such as a hulkish lion cub, a dog with a tiny head, a funny bear, a cow with a big head, a jogger stretching out, an iguana, and even two strange elephants.*

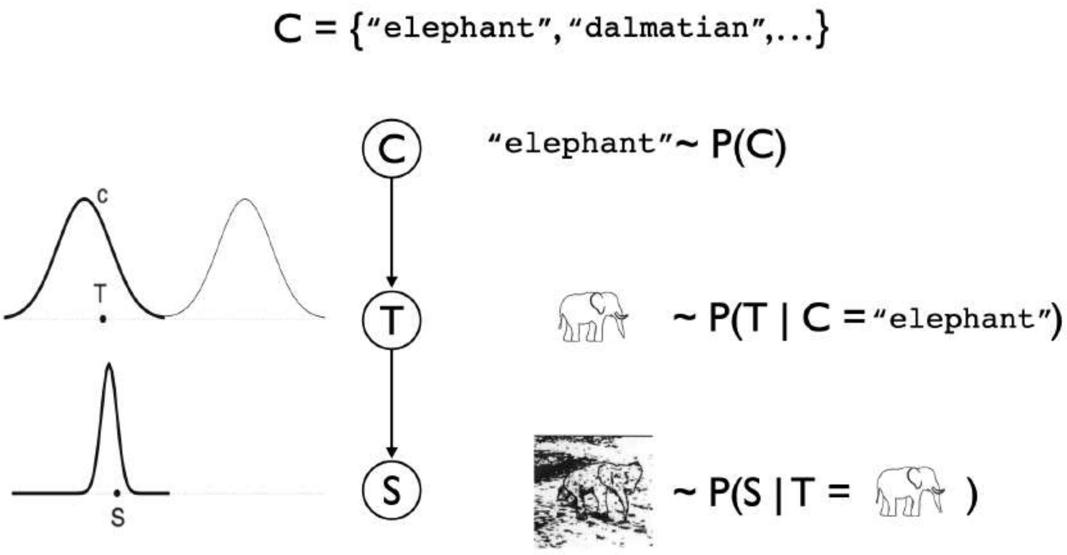


Figure 5.8: *A generative view of categorical perception*

This is an important issue for our work. We will assume that, to support the brain predictive striving, grounded concepts build upon and make sense of both exteroceptive and interoceptive sensations, and that, in the light of Conceptual Act Theory (CAT) emotions themselves, as we name them, are nothing but categories in service of allostasis.

5.2 Categorization and the Conceptual Act Theory

The basic features of Perceptual Categorization that we have discussed above to some extent, can be generalised, conforming to Barrett et al. (2015), as follows:

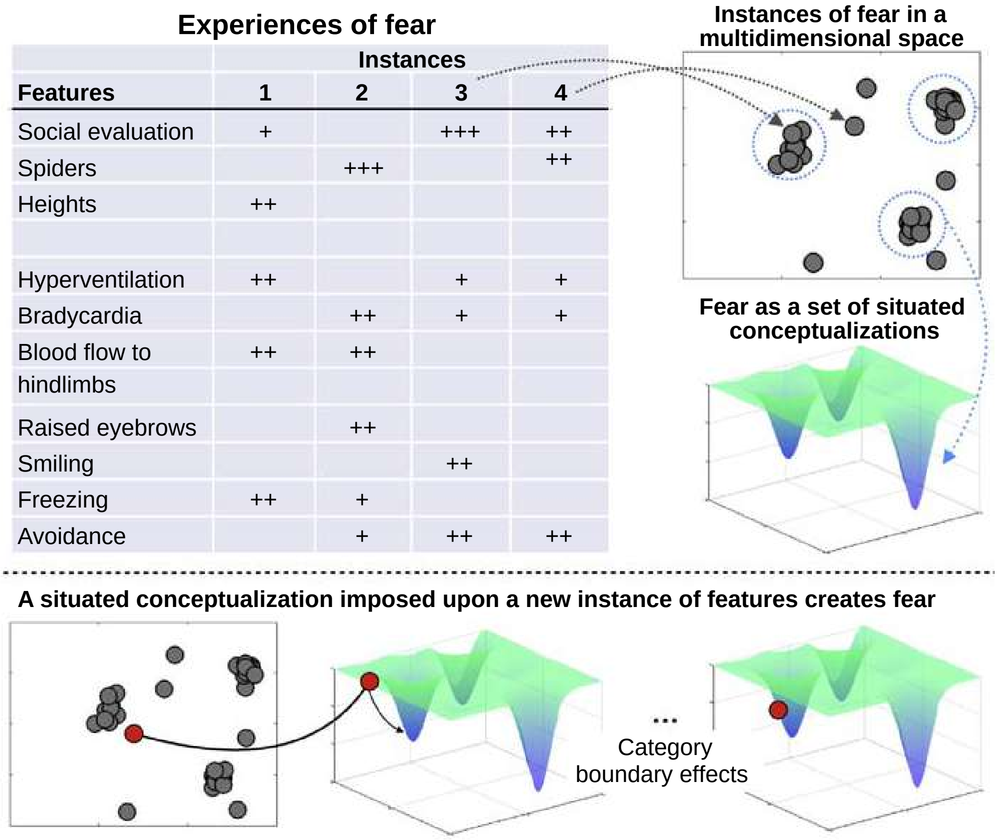
- Sensory input is categorized using conceptual knowledge from past experience. Prior experience is used to predict and make meaning of sensations
- Prior knowledge is *enactive* along perceptual inference: not only elements not immediately present in the visual input are inferred, but also extra experiential detail can be filled in either exteroceptive (odour, touch sensations, etc.) or interoceptive (e.g.,gut feelings); altogether, this is often defined as a simulation in the vein of Barsalou (2008) or categorization *tout court* by Barrett et al. (2015);
- The inferential process induced by categorization prepares for situated action
- Categorization, being enactive and preparing for specific actions, produces some kind of automatic change in the physical state of the agent. Change impacts the internal, interoceptive sensations contributing to the core affective tone (pleasantness, arousal). As such, it is a tool to modify and regulate the body, and in turn to create feelings.
- The process of meaning making rarely happens because of a deliberate, conscious goal to figure things out.

In this framework, to sum up, the process of applying prior knowledge to incoming sensory input is named *conceptual act*. Emotion, *per se* is nothing but an abstract, *ad hoc* category, mentally represented by a situated conceptualization, or more precisely a set or a manifold of situated conceptualizations. One clear example concerning fear is provided in Figure 5.9.

Just like any other form of categorization it is a construction across many levels of abstraction (Figure 5.10.

It is an act rather than a passive event, because the agent is not merely detecting and experiencing what it is out there in the world or what is going on inside the body: agent's prior knowledge plays a role in creating momentary experience. Thus, any conceptual act is embodied, because prior experience, in the form of category knowledge, comes about as the activation of sensory and motor neurons, thereby reaching down to influence bodily states and/or their representations and sensory processing (Barrett et al., 2015). A glimpse of the process, from the Bayesian listener's perspective, is shown in Figure 5.11.

What is novel here, with respect to classic Categorical Perception, is the role that interoceptive sensations play in addition to the exteroceptive ones. Interoception determines which parts of the world are worth caring about in the moment. Without it, an


Experiences of fear
Instances of fear in a multidimensional space
Fear as a set of situated conceptualizations
A situated conceptualization imposed upon a new instance of features creates fear
...
Category boundary effects

	Instances			
Features	**1**	**2**	**3**	**4**
Social evaluation	+		+++	++
Spiders		+++		++
Heights	++			
Hyperventilation	++		+	+
Bradycardia		++	+	+
Blood flow to hindlimbs	++	++		
Raised eyebrows		++		
Smiling			++	
Freezing	++	+		
Avoidance		+	++	++

Figure 5.9: *Conceptualization 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 behavioral 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 conceptualizations 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)*

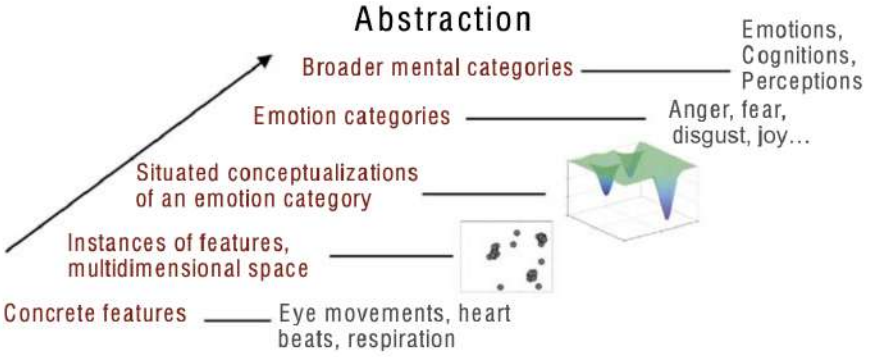

Abstraction
Broader mental categories
Emotions, Cognitions, Perceptions
Emotion categories
Anger, fear, disgust, joy...
Situated conceptualizations of an emotion category
Instances of features, multidimensional space
Concrete features
Eye movements, heart beats, respiration

Figure 5.10: *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 conceptualization 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 conceptualizations 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)*

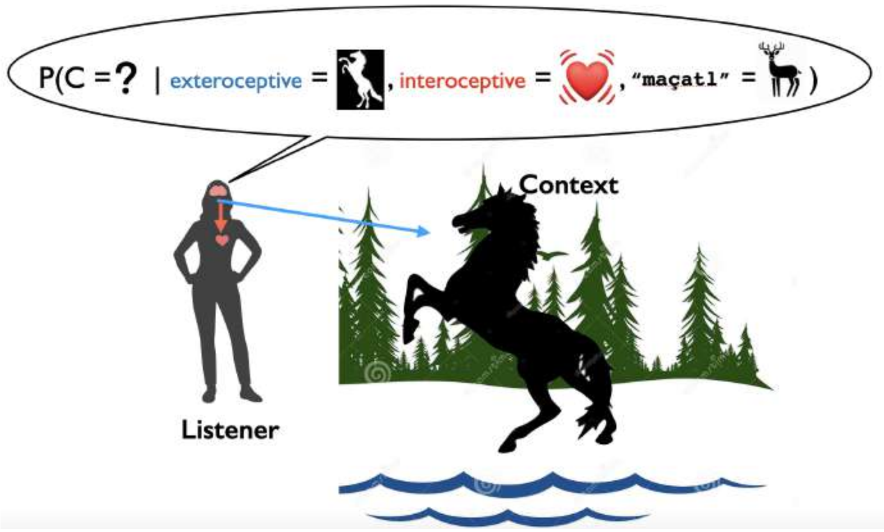


Figure 5.11: *A glimpse of the conceptual act in meaning making from the listener's perspective. In a Bayesian view, the listener engages in the backward process of inferring / categorizing the novel "object" in context, (e.g., a horse in a bad temper). Her inference is based on both exteroceptive and interoceptive sensations, together with prior conceptual knowledge that, after hearing the world "maçatl", is available to her. In the forward, top-down process, predictions are issued by generating / simulating exteroceptive and interoceptive signals to be compared against actual sensations in the effort of making meaning of them.*

actual agent would not appraise relevant features of the physical surroundings or even care for conversational partners.

Interoception is fundamental to build a fundamental psychological primitive named the "core affect". Core affect can be described as a state of pleasure or displeasure (valence) with some degree of arousal (Barrett, 2006b,c; Russell, 2003; Russell Barrett, 1999). Together, valence and arousal form a unified state. Although it is possible to focus on one property or the other, people cannot feel pleasant or unpleasant in a way that is isolated from their degree of arousal. This kind of affect is referred to as "core" because it is grounded in the *internal milieu*, an integrated sensory representation of the physiological state of the body: the somatovisceral, kinesthetic, proprioceptive, and neurochemical fluctuations that take place within the core of the body.

Core affect is thus realized by integrating incoming sensory information from the external world with homeostatic and interoceptive information from the body. The result is a mental state that can be used to safely navigate the world by predicting reward and threat, friend and foe. Indeed, affect is a central feature in many psychological phenomena, obviously including emotion. But by no means affect can be equated to emotion.

Further, these three sources — sensations from the world, sensations from the body, and prior experience — are continually available, and they form three of the fundamental aspects of all mental life.

As Barrett and Bliss-Moreau (2009) put it:

> Core affect [...] represents a basic kind of psychological meaning. The basic acoustical properties of animal calls (and human voices) directly act on the nervous system of the perceiving animal to change its affective state and in so doing conveys the meaning of the sound [...] All words (regardless of language) have an affective dimension of meaning, so that people cannot communicate without also (often inadvertently) communicating something about their affective state. Learning a new language fluently does not merely require making a link between the phonological forms of words and their denotation, but a connection to affective changes must also be forged.

In this perspective, even though CAT is mostly referred to as a psychological construction theory of emotion, emotion is *au pair* with cognition and perception: no longer are these Platonic essences, faculties instantiated in modules of our brain, but names that we use to describe mental states, complex psychological categories, under certain circumstances. So, *perception* is the name for psychological moments in which the focus is on understanding what externally driven sensations refer to in the world.

Cognition is the name for psychological moments in which the focus is on understanding how prior experiences are reinstated in the brain. When a person experiences the act of remembering, this mental activity is called memory. When they do not, it is called thinking. When the mental activity refers to the future, it is called imagining.

Emotion is the name for psychological moments in which the focus is on understanding what the internal sensations from the body represent.

These three complex psychological categories rely on basic processes or psychological primitives: the ingredients in a recipe that produces a psychological moment — what we might call an emotion, or memory, or thought, and so on — although they are not specific to any one kind of moment. Core affect, categorization, executive attention are but three examples of such primitives. Depending on the combination and on the weighting of primitives (and depending on the purposes of the observer), mental states mental might be called seeing or thinking or feeling (cfr. Figure 5.12.

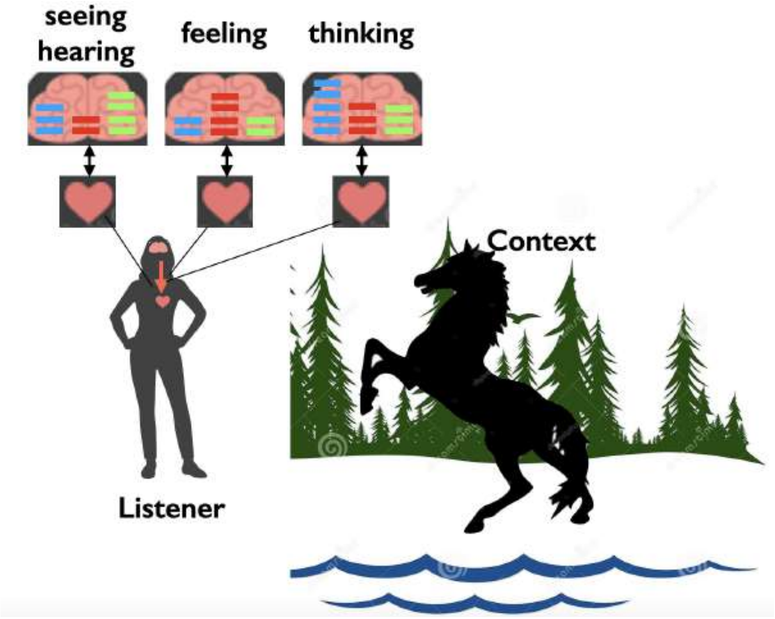


Figure 5.12: *Mental states, iconically depicted as brain states, comprised of different combinations of the same three psychological primitives (represented in red, green, and blue). Depending on the recipe (the combination and relative weighting of psychological primitives in a given instance) and a psychologist's interest, mental states are called seeing/hearing or thinking or feeling.*

As we have said, in every waking moment, the core business of our brain is not perceiving, or thinking or feeling, but, more generally to make predictions in order to give sensations, either exteroceptive or interoceptive, meaning in the service of allostatis,

The history of emotion theories is a longstanding one, and other competing theories have been proposed. It is important to provide a brief overview, because this makes clear what assumptions are made, explicitly but more often implicitly, in computational theories of emotions in general, but markedly in the affective computing field, which is somehow central in Computer Science and Artificial Intelligence.

5.3 What is an emotion? Perspectives and theories

So far, to smooth the way for illustrating the main rationales of this thesis, we have straightforwardly introduced the CAT approach and its view on emotions. Yet, in the study of emotion there are other competing theories, some of which are certainly relevant on the computational modelling side.

As a matter of fact, the attempt at answering the fundamental question posed by James (1884), "What is an emotion?", as the title of an essay he wrote for Mind well over a century ago, has a long-standing and turbulent history in the Western culture. For an in-depth discussion, the reader should refer to the superb accounts by philosophers Leys (2017) and Nussbaum (2003).

Philosophers have been concerned about the nature of emotion since Socrates and the “pre-Socratics” who preceded him, and although the discipline has grown up as the pursuit of reason, the emotions have always lurked in the background. Plato himself wrote that the human psyche consists of three parts: rational thoughts, passions (which today we call emotions), and appetites like the drive for hunger and sex. Rational thought was in charge, controlling the passions and appetites, as a charioteer wrangling two winged horses. This essentialist view might be considered the beginning of the classical view of emotion. By contrast, Aristotle seems to have anticipated most of the main contemporary appraisal theories. His analysis of emotion includes a distinctive cognitive component, a specified social context, a behavioral tendency, and a recognition of physical arousal (Solomon, 2008).

In the classical view, an emotion is understood as a separate and independent ability, or *faculty*, caused by its own separate processes. In this approach, emotions are categorically different phenomena from perceptions and cognitions, and each emotion (eg, anger, sadness, fear, and so on) is categorically different from every other emotion, each being caused by a different mechanism.

The classical view is essentialist by very nature: each emotion faculty is assumed to have its own innate physical essence or “fingerprints” that distinguishes it from all other emotions. Modern versions of the classical view of emotion include Basic Emotion Theory (BET) approaches and causal appraisal approaches: are united by a similar hypothesis regarding emotion “fingerprints”.

In the construction view, either psychological or social, an emotion is not a distinct faculty with its own distinct mechanism. Emotion categories are not natural kinds. Instead, one key hypothesis that unites all constructionist theories is that an emotion word, such as “happiness”, refers to a population of highly variable instances, each of which is tailored to a specific situation or context (Barrett, 2016; Gross and Feldman Barrett, 2011). So an emotion is not an entity with firm boundaries in nature: it is a category of instances. Instances within a category vary because each one is tailored to the environment, that is, there are no Platonic emotion essences or fingerprints.

A number of appraisal theories can be situated between these two contrasting views as illustrated in Figure 5.13. Constitutive appraisal theories (eg, Ortony et al., 1990) have more in common with psychological construction theories than causal appraisal approaches, the latter being in some respect closer to BET approaches.

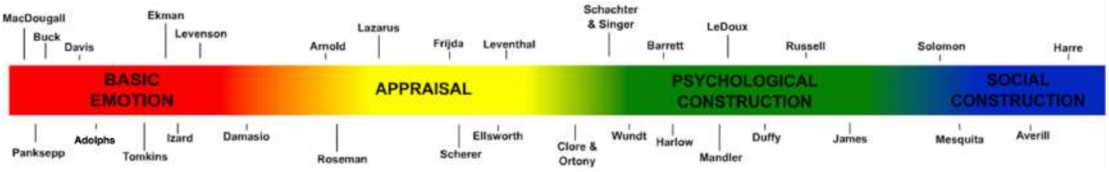


Figure 5.13: *Navigating emotion Theories. Theories are loosely arranged along a continuum, which could be defined in terms of a gradient of essentialism, the highest degree located at the left. Four "zones" are distinguished: (1) Basic Emotion Theories (BET), in red (for example, Anderson and Adolphs, 2014; Damasio, 1999; Ekman and Friesen, 1971; Izard, 1993; Levenson, 1988; Panksepp, 2004; Tomkins, 1962); (2) Appraisal Theories, in yellow (for example, Arnold, 1960; Ortony et al., 1990; Lazarus, 1991; Ellsworth and Scherer, 2003; Frijda, 1986); (3) Psychological Construction Theories, in green (Duffy (1941), James (1884), Barrett, 2009; Russell, 2003; LeDoux and Hofmann, 2018; Mandler et al., 1975; Schachter and Singer, 1962); (4) Social Construction Theories. in blue (e.g, Solomon, 2007; Averill, 1980; Mesquita et al., 2016; Wood and Harré, 1986). Theories in the red band and the left-most portion of the yellow band are much more essentialist than those in the right-most part of the yellow band and the green/blue bands (which are all non-essentialist theories). The greatest heterogeneity in essentialist assumptions lies in the appraisal zone, where classical appraisal theories (eg, Arnold, 1960; Lazarus, 1991 share many similar assumptions with BET, whereas constitutive appraisal theories (eg, Ortony et al., 1990) have more in common with psychological construction theories. Adapted from Barrett (2016); Gross and Feldman Barrett (2011)*

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

Table 5.1: *Core assumptions of four emotion perspectives. Adapted from Gross and Feldman Barrett (2011)*

5.3.1 The classical view

The most representative approach of the classical view currently is the Basic Emotion Theory (BET). In its modern version, it was pioneered by the american psychologist Silvan S. Tomkins. Tomkins posited the existence of a limited number of discrete, primary, or "basic emotions" as part of a universal human nature. These were held to be characterized by signature facial expressions and specific patterns of behavioral and autonomic responses. Tomkins' approach paved the way to Carroll Izard and Paul Ekman, the foremost theorist of the concept of "basic emotions". Here, we focus on Ekman's proposal, because of the currency it has impressively gained as a founding paradigm of ongoing affective computing research.

Following Tomkins's theorizing, Paul Ekman, together with associate Wallace Friesen and anthropologist E. R. Sorensen, published highly cited and influential studies of facial expressions of nonliterate indigenous peoples in exotic locales, such as the Fore tribe in Papua New Guinea (Ekman et al., 1969). The initial studies produced disappointing results when participants were asked to label the faces freely (Fridlund and Russell, 2021). To boost recognition rates, Ekman et al. shifted to Tomkins and McCarter's tactic of having participants pick the closest match from a short list either of emotion labels or emotion-related stories. The list was tailored to fit the six face photos and prescribed emotions which Ekman selected somewhat arbitrarily from Tomkins's eight, with the terms slightly renamed: happiness, sadness, anger, fear, surprise and disgust. Translators asked participants to make the matches, and Sorenson later disclosed his uncertainty about whether the translation included coaching. Ekman codified his claims and presented them as his Neurocultural Theory (Eckman, 1972).

Ekman originally cast his theory as under Darwin's imprimatur: at some early time in history certain facial movements were acquired to serve some biologically adaptive function, but are now vestiges which do communicate feelings, but which do not have as their primary purpose the expression of an inner state to another person. Ekman also claimed that facial movements were nonetheless instigated by a specialized "facial affect program" (FAP), neurally coded instructions that were phyletic and universal, as were many of the triggers that activated them. Thus Ekman's Neurocultural Theory simultaneously portrayed our faces as vestigial, but with brain circuitry evolved to produce them.

To make a long story short, the key features of Ekman's "neurocultural" model (Ekman et al., 1969; Ekman and Friesen, 1971; Eckman, 1972; Ekman and Rosenberg, 1997; Ekman, 1993; Ekman and Rosenberg, 1997,?; Ekman, 1999) can be summarized as follows:

1. There exists a small set of basic emotions, defined as pan-cultural categories or "natural kinds." These include fear, anger, sadness, disgust, joy, and surprise. Basic emotion are evolved, genetically hardwired, reflex-like responses of the organism.

2. Each basic emotion manifests itself in distinct physiological and behavioral patterns of response, especially in characteristic facial expressions that are automatically recognized.

3. The face expresses the emotions, except when expressions are disguised by cul-

tural or conventional norms (*display rules*) for controlling and managing emotions in public, or are masked by deliberate deception. Under pure or unfiltered conditions, facial displays are authentic "readouts" of the discrete internal states that constitute the basic emotions. Ekman called the muscles involved in the facial expression of the emotion "reliable" muscles because they are difficult to control and hence produce expressions that are hard or costly to fake.

4. Each basic emotion is linked to specific neural substrates or subcortical *affect programs*. This assumption involves some degree of modularity and information encapsulation in brain functions (for instance, the amygdala has been pinpointed as the neural seat of fear, while the insula has been implicated in disgust).

5. Although the emotions can and do combine with cognitive systems in the brain, emotion and cognition are essentially separate. Emotional expressions are special because they are involuntary, not intentional: they occur without choice. The communicative value of a signal differs if it is intended or unintended: emotional expressions are unintended.

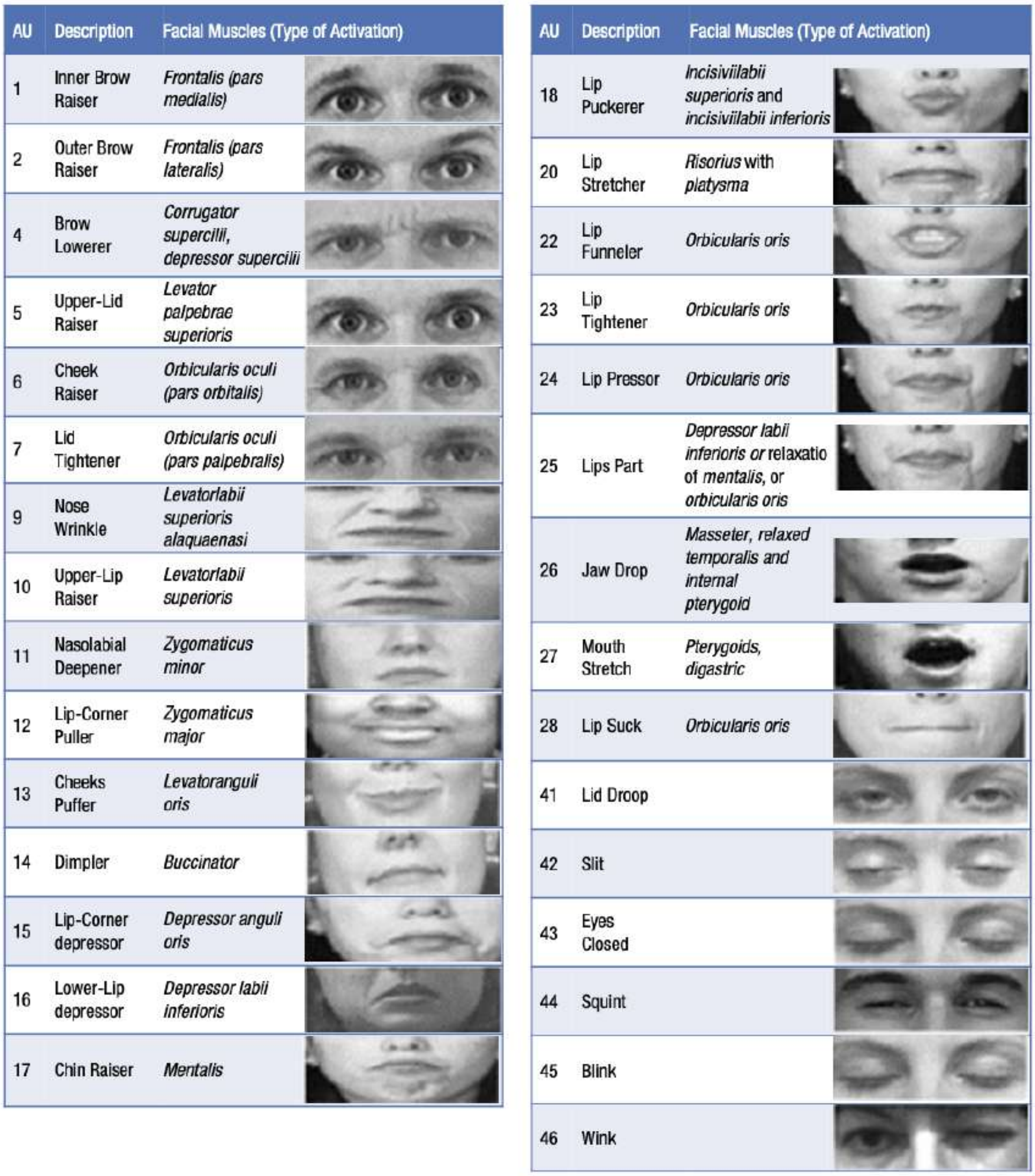
In the Ekman's Neurocultural Theory, a prominent role is assigned to facial expressions. There were originally six iconic expressions, each produced by a basic emotion: happiness, sadness, anger, fear, surprise, and disgust. Facial expressions of emotion are biological and universal: they have phylogenetic origins and are hard-wired in the brain.

Under such circumstances, Ekman promoted a method for measuring facial movements (and thus the internal emotional state of the expresser) with human coders, the Facial Action Coding System, or FACS (Ekman and Rosenberg, 1997). FACS is a systematic approach to describe what a face looks like when facial muscle movements have occurred. Human coders train for many weeks to reliably identify specific movements called action units (AUs). Each AU is hypothesized to correspond to the contraction of a distinct facial muscle or a distinct grouping of muscles that is visible as a specific facial movement. For example, lowering of the inner corners of the brows (activation of the *corrugator supercilii*, *depressor glabellae*, and *depressor supercilii*) corresponds to AU 4. Figure 5.14 displays the original FACS codes for the configurations of the facial movements that have been proposed.

A set of FACS can be used to univocally code a prescribed (facial expression of) emotion (emotion FACS, EMFACS, Figure 5.15) among the prototypic (facial expression of) emotions: anger, disgust, fear, happiness, sadness, and surprise, respectively.

Scientists agree that facial movements convey a range of information and are important for social communication, emotional or otherwise. People do sometimes smile when happy, frown when sad, scowl when angry, and so on, as proposed by BET, more than what would be expected by chance. Yet how people communicate anger, disgust, fear, happiness, sadness, and surprise varies substantially across cultures, situations, and even across people within a single situation. Furthermore, similar configurations of facial movements variably express instances of more than one emotion category. In fact, a given configuration of facial movements, such as a scowl, often communicates something other than an emotional state.

In brief, instances of an emotion category are signaled with a distinctive configuration of facial movements that has enough *reliability* and *specificity* to serve as a



AU	Description	Facial Muscles (Type of Activation)
1	Inner Brow Raiser	*Frontalis (pars medialis)*
2	Outer Brow Raiser	*Frontalis (pars lateralis)*
4	Brow Lowerer	*Corrugator supercilii, depressor supercilii*
5	Upper-Lid Raiser	*Levator palpebrae superioris*
6	Cheek Raiser	*Orbicularis oculi (pars orbitalis)*
7	Lid Tightener	*Orbicularis oculi (pars palpebralis)*
9	Nose Wrinkle	*Levatorlabii superioris alaquaenasi*
10	Upper-Lip Raiser	*Levatorlabii superioris*
11	Nasolabial Deepener	*Zygomaticus minor*
12	Lip-Corner Puller	*Zygomaticus major*
13	Cheeks Puffer	*Levatoranguli oris*
14	Dimpler	*Buccinator*
15	Lip-Corner depressor	*Depressor anguli oris*
16	Lower-Lip depressor	*Depressor labii inferioris*
17	Chin Raiser	*Mentalis*

AU	Description	Facial Muscles (Type of Activation)
18	Lip Puckerer	*Incisiviilabii superioris* and *incisiviilabii inferioris*
20	Lip Stretcher	*Risorius* with *platysma*
22	Lip Funneler	*Orbicularis oris*
23	Lip Tightener	*Orbicularis oris*
24	Lip Pressor	*Orbicularis oris*
25	Lips Part	*Depressor labii inferioris or* relaxatio of *mentalis*, or *orbicularis oris*
26	Jaw Drop	*Masseter, relaxed temporalis and internal pterygoid*
27	Mouth Stretch	*Pterygoids, digastric*
28	Lip Suck	*Orbicularis oris*
41	Lid Droop	
42	Slit	
43	Eyes Closed	
44	Squint	
45	Blink	
46	Wink	

Figure 5.14: *The Facial Action Coding System. AU = action unit. Adapted from Barrett et al. (2019)*

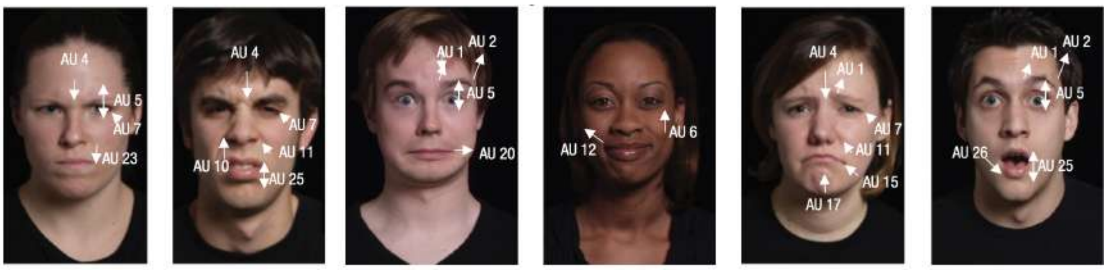


Figure 5.15: *Emotion FACS (EMFACS). From left to right: the proposed expression for anger corresponds to a prescribed EMFACS code for* ***anger*** *(described as AUs 4, 5, 7, and 23);* ***disgust*** *(described as AU 10);* ***fear*** *(AUs 1, 2, and 5 or 5 and 20);* ***happiness*** *(AUs 6 and 12);* ***sadness*** *(AUs 1, 4, 11, and 15 or 1, 4, 15, and 17);* ***surprise*** *(AUs 1, 2, 5, and 26). Adapted from Barrett et al. (2019)*

diagnostic marker of those emotion states.

In a detailed and technical review of these claims, Barrett et al. (2019) have shown that this Ekman's approach suffers from:

- limited reliability: instances of the same emotion category are neither reliably expressed through nor perceived from a common set of facial movements;

- lack of specificity: there is no unique mapping between a configuration of facial movements and instances of an emotion category;

- limited generalizability: the effects of context and culture have not been sufficiently documented and accounted for;

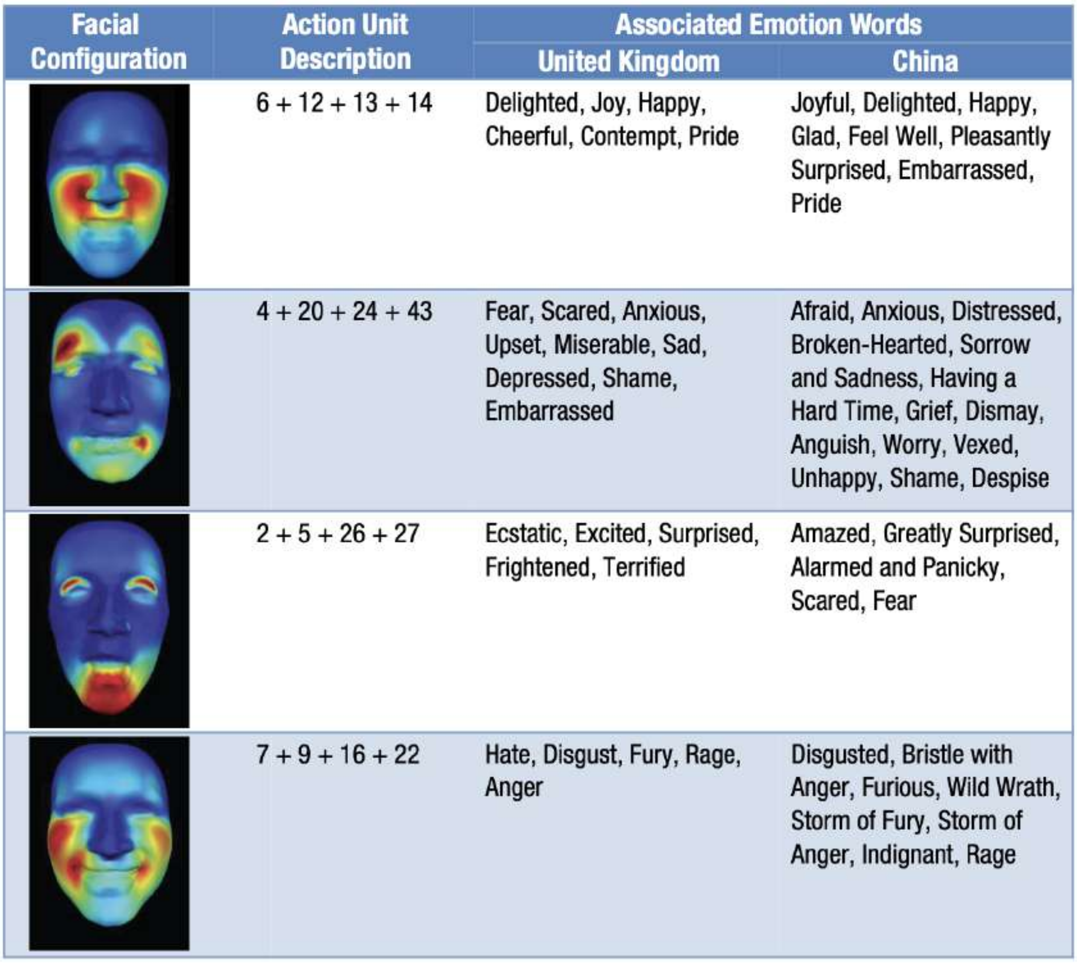
- questionable validity: whether an observed variable actually measures what is claimed (for example, whether a facial movement reliably expresses an emotion - convergent validity - and specifically that emotion - discriminative validity) where the presence of the emotional instance can be verified by objective means.

Barrett et al. (2019) also propose a cautious list of recommendations when adopting a FACS based approach for emotion detection/recognition.

Numerous critiques of BET have appeared since its inception. To addresse them, by and large, Ekman and other BET theorists have "weakened" in varying degrees some of their original assumptions. This extensions aim at accounting for more nuanced and complex processes involved in emotion recognition, in the structure of how people perceive emotional expression and finally, in weighing contextual influences upon emotion recognition. One example of the FACS encoding how facial configurations might relate to different emotion words (categories) depending on culture is provided in Figure 5.16.

A recent review of such advancements has been given by Keltner et al. (2019).

Notwithstanding, Ekman's approach is nowadays judged to be more and more controversial, at the light of new empirical evidence in emotion neuroscience. For in-depth reviews, see Fridlund and Russell (2021); Crivelli and Fridlund (2019); Barrett and Satpute (2019); Leys (2017).



Facial Configuration	Action Unit Description	Associated Emotion Words	
		United Kingdom	China
	6 + 12 + 13 + 14	Delighted, Joy, Happy, Cheerful, Contempt, Pride	Joyful, Delighted, Happy, Glad, Feel Well, Pleasantly Surprised, Embarrassed, Pride
	4 + 20 + 24 + 43	Fear, Scared, Anxious, Upset, Miserable, Sad, Depressed, Shame, Embarrassed	Afraid, Anxious, Distressed, Broken-Hearted, Sorrow and Sadness, Having a Hard Time, Grief, Dismay, Anguish, Worry, Vexed, Unhappy, Shame, Despise
	2 + 5 + 26 + 27	Ecstatic, Excited, Surprised, Frightened, Terrified	Amazed, Greatly Surprised, Alarmed and Panicky, Scared, Fear
	7 + 9 + 16 + 22	Hate, Disgust, Fury, Rage, Anger	Disgusted, Bristle with Anger, Furious, Wild Wrath, Storm of Fury, Storm of Anger, Indignant, Rage

Figure 5.16: *Culturally common facial configurations and related emotion words/phrases in UK end China. Red coloring indicates stronger AU presence and blue indicates weaker AU presence. Some words and phrases that refer to emotion categories in Chinese are not considered emotion categories in English. Adapted from Barrett et al. (2019)*

5.3.2 The Appraisal Theory view

Appraisal theory, as we know it today, is usually attributed to Arnold (1960), who made an early and influential statement of the cognitive approach to emotion. She proposed that people implicitly appraise or evaluate everything they encounter, and that such evaluations occur immediately and automatically.

Appraisal theories of emotion propose that emotions or emotional components are caused and differentiated by an appraisal of the stimulus as mis/matching with goals and expectations, as easy/difficult to control, and as caused by others, themselves or impersonal circumstances. Specifically, the appraisal is a process in which values are determined for a number of appraisal factors such as goal relevance, goal in/congruence, un/expectedness, control, and agency.

Using a surface taxonomy, appraisal theories can be divided into two flavors based on what they try to explain (Moors, 2020). A first flavor of appraisal theories tries to explain specific emotions; these set out to explain specific emotions as they figure in natural language, such as anger, fear, sadness, and happiness.

A second flavor of appraisal theories tries to explain certain striking features or components of emotions, such as their intense, overwhelming, nature, that they have positive or negative valence, and/or their embodied aspects. For instance, features to be considered might be specific action tendencies (e.g., tendencies to flee, fight, and give in), specific somatic response patterns, specific facial expressions, and/or specific feelings.

As to the causal explanation of emotion, appraisal theories, in contrast do the classic view, emphasize that there are hardly any one-to-one relations between features of stimuli and features of emotions. One stimulus can produce different emotions in different individuals or on different occasions.

For what concerns mechanistic explanation, theories of appraisal split the process from stimulus to emotion into two steps: one step in which a stimulus is processed by appraisal and another step in which the output of the appraisal process is translated into a specific emotion (in first-flavor appraisal theories) or in specific values of the other emotional components (in second-flavor appraisal theories).

Regarding the first step, there are no restrictions on the operations that may be involved in appraisal, as long as the output of these operations are representations of values on the proposed appraisal factors. For what concerns the second step, appraisal theories of the first flavor set out to explain specific emotions. They propose that the appraisal pattern resulting from the appraisal process is integrated in a summary appraisal value (called a core relational theme by Lazarus, 1991) and that this summary value determines the specific emotion that is at stake. This, in turn, determines the values of the output components. Appraisal theories of the second flavor, by contrast, set out to explain the values of the output components, without linking them to specific emotions. They propose that each appraisal value has a separate influence on the values of the output components and together these values form the emotion (Scherer, 2009). Each appraisal output has an influence on the action tendency, which mobilizes somatic responses that prepare the organism for overt action. Aspects of all these components seep into consciousness where their integrated sum makes up the content of the feeling component. In this scenario, the organism at no point has to determine the specific emotion that is at stake; instead, emotions are considered as emergent phenomena.

This for the general aspects. In the following we largely draw on Barrett's taxonomy (Barrett, 2016) in order to discuss appraisal theories in relation to other approaches.

Causal appraisal theories

According to causal appraisal approaches, a cognitive appraisal (Arnold, 1960; Lazarus, 1991), or a suite of appraisals (Roseman, 1984; Scherer, 2009; Smith and Ellsworth, 1985) makes meaning of the stimulus situation. This in turn triggers the emotion.

As in BET, many early causal appraisal approaches assumed that the resulting emotion triggers a set of consistent and specific physiological changes, facial muscle movements, behavior, feelings, and so on (e.g., when a person appraises that a dark alley is uncertain, unpleasant, and that she lacks control of the situation, she might experience fear which in turn generates an increased heart rate, sweating, widened eyes, and the tendency to run away)

Emotions are natural kinds by virtue of homology: all instances of the same category (e.g., anger) emerge from the same causal mechanism. By hypothesizing that the appraisal is an intervening mechanism between the stimulus and emotion, causal appraisal models made it easier to accommodate evidence of variability in physiological, facial, and behavioral patterns into the natural kinds framework: such variability could occur because different people appraise the same stimulus in a different way, and thus experience different emotions.

Prima facie, causal appraisal models appear similar to psychological constructionist approaches, especially since both can accommodate greater variability in emotional responding than basic emotion approaches. Yet they differ from constructionist approaches in two important ways. First, causal appraisal approaches view appraisals as a specific mechanism that is itself distinct from the emotion. Second, causal appraisal models assume that emotions include distinct steps: appraisals evaluate the stimulus situation (Ellsworth and Scherer, 2003), which then causes the emotion, which causes associated bodily changes.

Constitutive appraisal theories

Recent constitutive appraisal approaches have moved away from making strong causal hypotheses about the role of appraisals in emotion (e.g., Clore and Ortony, 2008; Moors, 2013; Scherer, 2001). Like causal appraisal approaches (and unlike psychological constructionism), constitutive appraisal models still assume that emotion categories refer to distinct states with specific functional importance. Yet unlike causal appraisal approaches (and like psychological constructionism), constitutive appraisal approaches highlight the informational content of appraisals rather than viewing them as mechanisms of the emotion, per se. Some constitutive appraisal models (e.g., Clore and Ortony, 2008) are thus quite similar to psychological constructionist approaches because they view appraisals as descriptions of what it is like to experience an emotion (for discussions, see Gross and Feldman Barrett, 2011; Lindquist, 2013; Lindquist et al., 2012; Barrett, 2016).

One clear example is the OCC model (the model proposed by Ortony, Clore and Collins, 1988), one of the several appraisal theories that arose in the 1980s. This psychological model is relevant beyond its theoretical merits also because it became popular among computer scientists building systems that reason about emotions or incorporate emotions in artificial characters.

According to OCC (Clore and Ortony, 2013): (a) emotions are more readily distinguished by the situations they signify than by patterns of bodily responses; (b) emotions emerge from, rather than cause, emotional thoughts, feelings, and expressions; (c) the impact of emotions is constrained by the nature of the situations they represent; (d) appraisals are psychological aspects of situations that distinguish one emotion from another, rather than triggers that elicit emotions; (e) analyses of the affective lexicon indicate that emotion words refer to internal mental states focused on affect; (f) the modularity of emotion, long sought in biology and behavior, exists as mental schemas for interpreting human experience in story, song, drama, and conversation.

Consistent with a constructivist approach, the OCC model posits that emotions are emergent conditions reflecting multiple modalities of affective reactions to psychologically important situations. The model distinguishes 22 emotion types differentiated by the psycholog- ically significant situations they represent. It distinguishes emotions involving a focus on events from those focused on actions and those focused on objects. Emotions concerned with outcomes of events are distinguished by such factors as whether they concern one's own (e.g., sad) or another's outcomes (e.g., pity), and whether they involve prospective outcomes (e.g., fear) or known outcomes (e.g., grief). Among emotions focused on prospective outcomes, some concern whether such prospects have been realized (e.g., satisfaction, fears confirmed) or not (e.g., disappointment, relief). But not all emotions are about the outcomes of events. Some concern the agency of actions. These emotions involve appraisals of actions as praiseworthy (e.g., pride) or blameworthy (e.g., shame). Within this focus, it matters whether a praiseworthy or blameworthy action is one's own (e.g., pride, shame) or another's (e.g., admiration, reproach).

The OCC argument is not that emotions are situations, but rather that emotions are embodied, enacted, and experienced representations of situations. Specific emotions surely do involve patterns of physiology, neurology, experience, expression, motivation, and so on. But the variation in these responses within a particular kind of emotion may be too great to discriminate among emotions on such bases (Clore and Ortony, 2013).

Central to OCC is the notion of (mental) *schemas*. People all have accessible, stereotypic scenarios of anger, fear, jealousy, and other emotions. These stereotypic scenarios can bring order to what people have to say. They provide ready-made frames for everyday experiences, and help interpret the present, remember the past, and anticipate the future. These schemas are not emotions, of course, but cartoon versions of emotions that provide categories for interpreting and communicating the essential aspects of important situations to self and others in a compelling form (Clore and Ortony, 2013).

5.3.3 The constructionist view

The constructionist view of emotion includes social construction theories (e.g., Averill, 1980; Mesquita et al., 2016), psychological construction theories (e.g., (Barrett et al., 2015; Russell, 2003)), and descriptive appraisal theories (e.g., Ortony and Clore, 2015), as well as the theory of constructed emotion that integrates social construction and psychological construction, as well as neuroconstructive and rational constructionist perspectives (Barrett and Satpute, 2019; Barrett, 2017c; Atzil et al., 2018).

Psychological Construction Models

In these models, emotions are not special mental states, unique in form, function, and cause from other mental states such as cognition and perception. This is because emotions are not caused by dedicated mechanisms. Instead, all mental states are seen as emerging from an ongoing, continually modified constructive process that involves more basic ingredients that are not specific to emotion (see Table 5.1). Psychological construction models treat emotions as folk categories, where each category is associated with a range of measurable outcomes. By some psychological construction accounts, emotions (like all mental states) are the emergent products of psychological ingredients — they are more than the sum of their parts — making these views continuous with descriptive appraisal accounts found to the very right of the yellow zone.

Social Construction Models

The right-most band (Figure 1, in blue) is occupied by social construction models. Here, emotions are viewed as social artifacts or culturally-prescribed performances that are constituted by sociocultural factors, and constrained by participant roles as well as by the social context (see Table 5.1). Some social construction models (particularly in psychology) treat social configurations as triggers for basic emotional responses, much as early appraisal models conceived of appraisals as cognitive triggers of basic emotions. However, other models in this zone view emotions as socio-cultural products that are prescribed by the social world and constructed by people, rather than by nature. Emotions are performances of culture, rather than internal mental states. Whether a socially constructed event is seen as an emotion (as opposed to some other kind of psychological event) depends on the network of social consequences it produces. To the extent that cognitive processes are involved as transmitters of cultural expectations and constraints, they are seen as learned, rather than given by nature (in contrast to some appraisal views), so that such cognitions vary from culture to culture. Both the mental and the behavioral components of emotion are thought to co-evolve as a function of local social meanings, and are considered primarily for their social function.

The different approaches to answer James' original question can be graphically summarized as in Figure 5.17. The Figure also highlights the relation between the different approaches and language concepts.

5.3.4 Conclusive remarks on CAT in the context of emotion theories

CAT has been originally proposed as a theory of constructed emotion relying on the predictive brain hypothesis. In its development it has been extended so to integrate social construction and psychological construction, as well as neuroconstructive and rational constructionist perspectives (Atzil et al., 2018; Barrett and Satpute, 2019). In the remainder of this thesis, this extended CAT will be referred to as CAT *tout court*, for simplicity.

We have straightforwardly introduced CAT drawing on the rational or theoretical model perspective, namely in the framework of Bayesian theories. Motivation for this

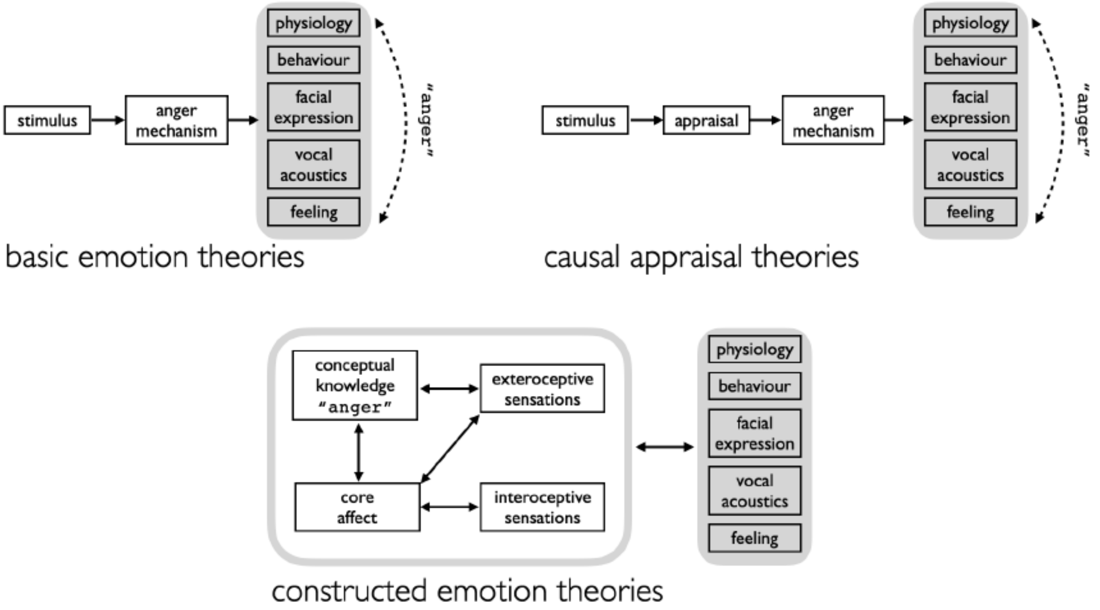


Figure 5.17: *Theories of emotion and language. In basic emotion theories emotion is triggered by a stimulus, each emotion (e.g., anger) having its own innate mechanisms. Causal appraisal models hypothesize that the cognitive appraisal (or a suite of appraisals) of the stimulus situation triggers a discrete emotion. In this regard, they are similar to basic emotion approaches, and are natural kinds approaches. Psychological constructionist approaches hypothesize that core affective changes (caused in part by interactions with the stimulus situation) are made meaningful using representations of prior experiences that are tied to the context. In a BET view, linguistic concepts are at most invoked after an emotion has formed and are purely used for communicating emotions to others. Causal appraisal models hypothesize that a cognitive appraisal intervenes between the stimulus and emotion, but this is not typically thought to be a linguistic process* per se. *By contrast, in a constructionist view, linguistic concepts help make meaning of ambiguous body states in light of the present context. Linguistic concepts are thus constitutive of the emotion, helping to create the experience in the first place. Adapted from Lindquist et al. (2015b); Lindquist (2013)*

choice relies on the central role played in the theory by category/concepts as generative tools and the possibility of exploiting approaches that have been previously set for modelling categorical perception of objects. Indeed, there is a close connection between categorical perception of objects and that of affect along ontogeny development (Hoemann et al., 2020b) as it has been summarized in Figure 5.1.

What we have assumed here is the following central hypothesis:

> All brains are faced with an inverse inference problem, that is Bayesian inference: ambiguous, noisy sense data continually arrives from inside the animal's body (the result of allostasis) and from the surrounding environment (the animal's niche). The brain does not have access to the causes of the sense data so it must infer them. So, a brain constructs inferences — hypotheses about the causes of sensations — by remembering past events that are similar to present conditions.
>
> The brain solves the inference problem by continually constructing *ad hoc* concepts to make sense of the harsh discordant mixture of signals arriving from its sensory organs. By "concept" we denote a mental representation of a category, a group of events or objects that are similar in some way.

It is worth noting that in most recent versions of CAT, the rational analysis framework is often equated with the predictive coding approach. However, in our perspective, as discussed at the beginning of this Chapter, predictive coding is but one of the viable alternatives apt to provide a solution to the complexities entailed by Bayesian inference (thus at the implementation model level), rather than a theoretical framework *per se*.

The aspects of CAT, involved by its extension to cope with subjects' social interactions (Hoemann et al., 2020b), are particular appealing for us because of the prominent role played by words and language (cfr., Figure 5.17).

Words set a powerful context for shaping mental inferences because they are a special type of sensory input that is inextricably linked to concepts and categories. Simply perceiving a word involves remembering related concept knowledge. Conceptual knowledge is a context that categorizes incoming sensory inputs and makes them meaningful, thereby influencing how facial configurations and other sensory signals are understood and acted upon.

As we have previously remarked, language and emotions are tightly intertwined: language and words have a constitutive role in emotion and the communicative act always conveys emotions, either unintentionally or through shared intentionality; communicating through language can even modulate emotion regulatory processes.

Further, in line with social constructionism, CAT suggests that emotion categories are a product of social reality and are culturally relative. Similarly, emotion concepts develop through contextualized social interactions in which language plays a significant part. In this way, CAT is consistent with perspectives that highlight the inherent intersubjectivity of emotional development. However, CAT extends beyond these perspectives to emphasize the role of the body and its anticipated energy needs. Social constructionism holds that emotion concepts are inherently about the relationship between social interactants. In comparison, the CAT account anchors the construction of emotion concepts (like all concepts) in the service of efficient physiological regulation. Humans interactively establish and reinforce emotion categories to co-regulate each other, but this is always in support of keeping bodily systems in balance (Atzil et al., 2018; Barrett and Satpute, 2019; Hoemann et al., 2020b).

5.4 What is a computational model of emotion? State of the art

Trying to exhaustively review the contributions to the field of computational models of affect, in order to systematically compare the different approaches is a mind-blowing endeavour and out of the scope of this thesis. We urge the reader to refer to some in-depth reviews (Gratch, 2021; Ma and Yarosh, 2021; Zhao et al., 2021; Schuller et al., 2021; Susanto et al., 2021; Ong et al., 2019b; Wang et al., 2020; Richardson, 2020; Schuller and Schuller, 2018; Hortensius et al., 2018; D'Mello et al., 2018; Poria et al., 2017; D'Mello and Kory, 2015; Ojha et al., 2021; Dimitrievska and Ackovska, 2020; Sheridan, 2020; Cavallo et al., 2018; Reisenzein et al., 2013).

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

We shall try to delineate at least some trends, by considering models developed within the main research programs involving the computation of affect and to relate them to psychological theories we have discussed so far.

That being so, we arm the reader with a preliminary map outlined in Figure 5.18.

Such map shows at a glance that there are different research programs that contend for the computation of affect, which can be roughly identified as machine learning-based models, robotic models, cognitive artificial intelligence (AI) based models. It is shown how several research areas have contributed, in different vein and substance, to such programs, psychological and neurobiological theories providing the necessary underpinning, at least in our view. Other areas that have fostered the flourish of methods are machine-learning and artificial intelligence (AI), either in their classic knowledge-based or the probabilistic, uncertainty-based accounts.

5.4.1 The affective computing / machine learning way

Affective Computing (AC) is the interdisciplinary field of study concerned with recognizing, understanding, simulating and stimulating affective states in the design of computational systems. Since the coining of the term by Picard (1997), affective computing has emerged as a cohesive and increasingly impactful discipline spanning computer science, psychology, neuroscience, philosophy, art and industry. Technology giants such as Apple, Amazon, Google and Facebook, as well as hundreds of smaller companies are deploying affective computing methods to predict or influence consumer behavior (Gratch, 2021; Schuller et al., 2021).

Picard's definition of the field was quite broad:

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

In such perspective artificial agents might have the ability to

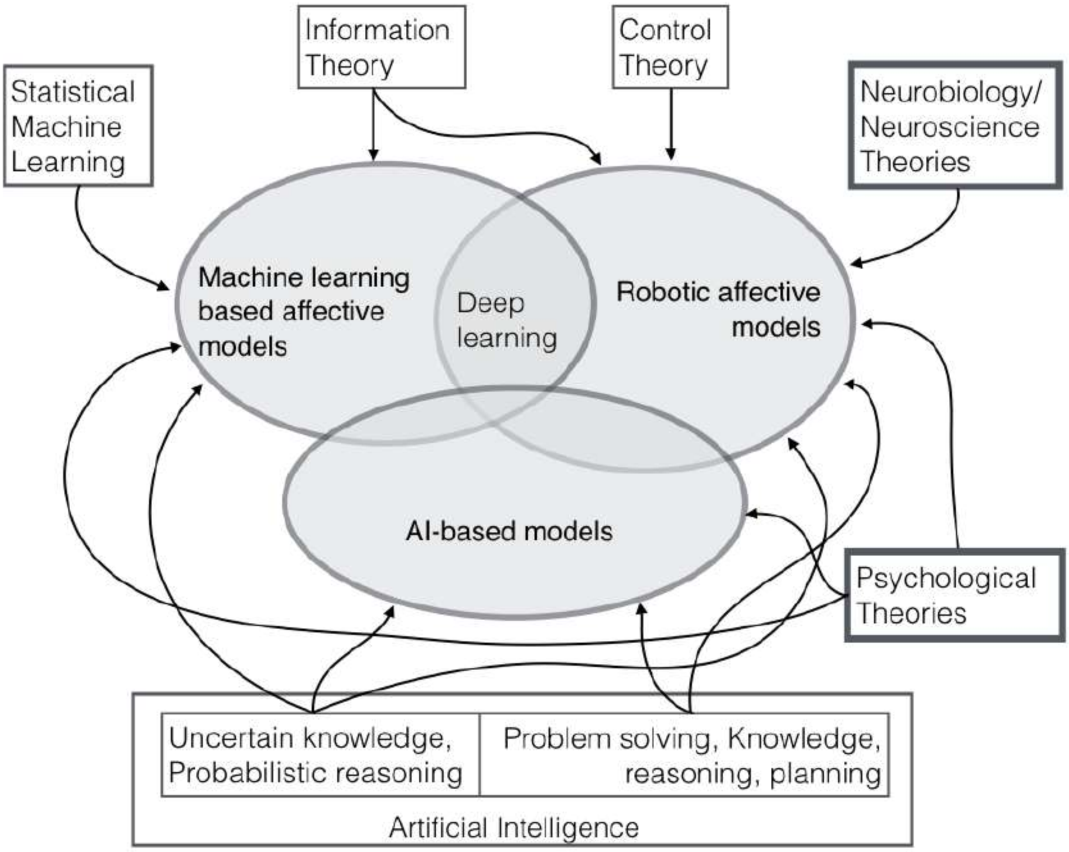

Information Theory
Control Theory
Statistical Machine Learning
Neurobiology/ Neuroscience Theories
Machine learning based affective models
Deep learning
Robotic affective models
AI-based models
Psychological Theories
Uncertain knowledge, Probabilistic reasoning
Problem solving, Knowledge, reasoning, planning
Artificial Intelligence

Figure 5.18: *A map of the main research programs in which the field of computational modelling of affect has been developed: machine learning-based models, cognitive robotics' models, cognitive artificial intelligence (AI) based models. Boxes and related arrows represent independent research areas and theories that have mainly contributed to such field. Adapted from Cuculo (2018)*

1. recognise/detect emotion,
2. express emotion,
3. “have emotions”,

the latter point being the hard stuff.

However, as witnessed by a recent review Ma and Yarosh (2021) a large number of papers (75%) that have been published in the field deals withe the detection problem (cfr., Figure 5.19, top panel).

Considering modalities, facial expressions, vocal expressions and others play a major role (circa 50%, cfr., Figure 5.19, centre panel), but also physiological signals are considered (20%), and more recently, multi-modal analysis (D’Mello and Kory, 2015; Schuller et al., 2021). Text and music (20%) are relevant too, the former fostered by recent developments in computational language processing, markedly sentiment analysis (Susanto et al., 2021; Wang et al., 2020).

As to the psychological models adopted, basically 50% of the works adopt variants of BET, and address the recognition of Ekman’s six basic emotions moving more recently to consider a higher number. Within the field this is usually referred as the categorical or discrete approach to emotion recognition. Other works rely on the Russell’s core affect, valence/arousal (V/A) space, which is commonly named the continuous or dimensional approach to emotion recognition. It is worth remarking, that the choice of the V/A representation is most of the times instrumental, being such representation a rather impoverished version of the original, at best, if not wrongly interpreted. Basically, the V/A affect space it is conflated *tout court* with discrete categorical emotion space, so that each emotion (e.g., anger) is assumed to be represented as a point in such two-dimensional continuous space. This merging of two completely different levels (affect and emotion), is clearly inconsistent with the constructionist view where core affect was conceived

The main problem in this field, beyond the original Picard’s statement of its research program, is that, by and large, the detection problem is solved through the classic pattern recognition pipeline

$$\text{signal} \rightarrow \text{feature extraction} \rightarrow \text{classification / regression}$$

where, more recently, supervised deep learning techniques applied to very large datasets (notably, in facial expression recognition, much like in NLP, have fused the feature extraction → classification steps, thus avoiding the need of hand-crafted features .

This is well described by D’Mello et al. (2018). The goal of the AC approach is to computationally model the link between signals and affect/emotion inference, which requires solving two main challenges. The first is to obtain abstractions (called descriptors or features) from raw signals recorded by sensors. The second main challenge is to produce affect estimates from the descriptors. The most common approach uses techniques from a subfield of machine learning. The term computational model in most cases boils down to computing the output of a supervised classifier. Based on the supervised learning method, the computational model can take on many forms, such as an equation, a set of rules, a decision tree, a forest of decision trees, and a neural network (D’Mello et al., 2018). Interestingly enough, D’Mello and Kory (2015) have provided a meta-analysis of multi-modal AC which is unique in this research field.

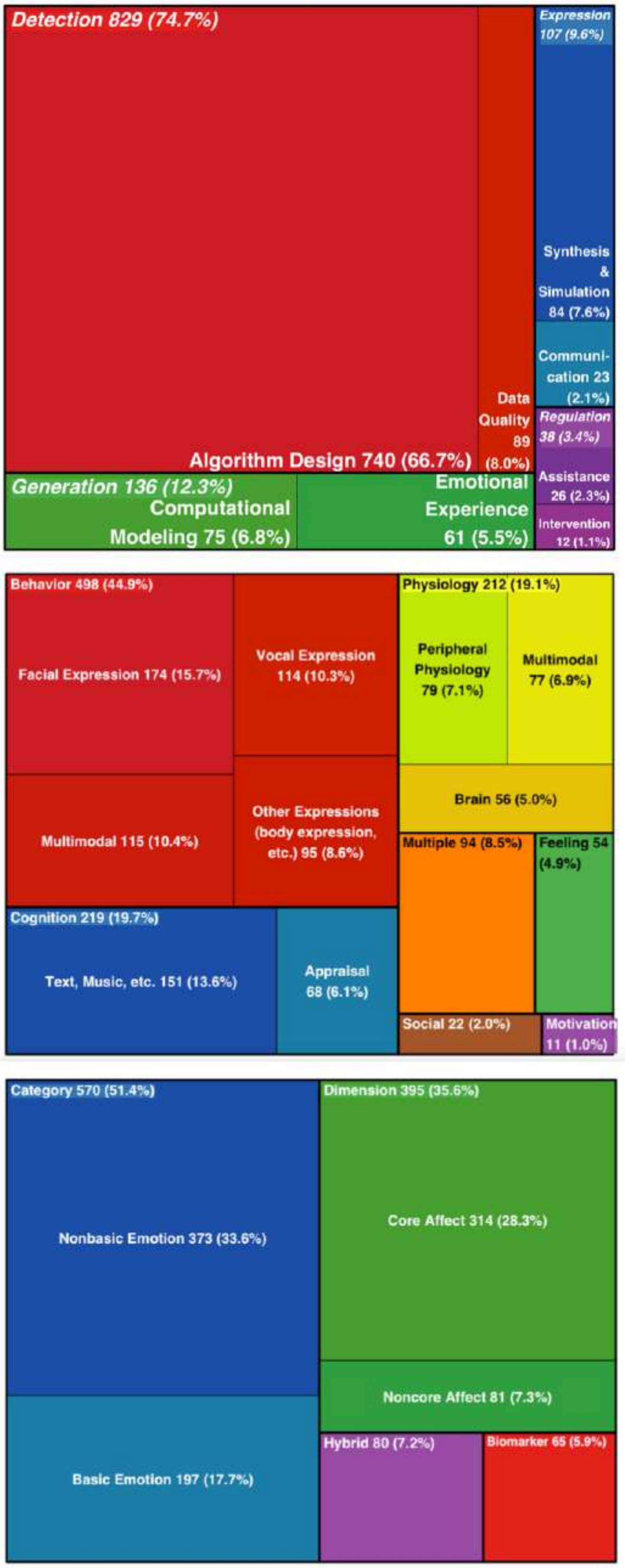


Figure 5.19: *Distribution of papers over topics (top panel), modalities (centre), psychological models (bottom). Adapted from Ma and Yarosh (2021)*

5.4.2 The cognitive Artificial Intelligence way

The aim of AI-oriented models of emotion is to the enhance human-computer interaction, by controlling the behaviour of virtual agents to motivate and establish empathy and bonding. A lucid and in-depth review of the relationships between computational modelling of emotion and the general goals of AI, when these are restricted to the domain of emotions, has been provided by Reisenzein et al. (2013). More recent and up to date reviews are Gratch (2021); Ojha et al. (2021); Ong et al. (2019b).

In this field, the goals of computational modeling of emotion largely correspond to the general goals of AI (in particular to classic AI), when these are restricted to the domain of emotions: (i) achieve a better theoretical understanding of emotions in natural and artificial agents by creating computational models of them (theoretical goal); (ii) to enrich the architecture of artificial agents with emotion mechanisms similar to those of humans, and thus with the capacity to "have" emotions (applied goal, Reisenzein et al., 2013).

The two goals are intertwined: endowing artificial agents with truly human-like emotion mechanisms presupposes reasonably faithful computational models of these mechanisms; conversely, by synthesizing mental processes, including emotions, in artificial agents is suitable way to attain a deep theoretical understanding of them. The common ground is the functionalist view of mental states of classic AI: understanding the capacity of artificial agents to have internal states that are functionally equivalent or at least similar to emotions in humans, i.e., that play causal roles in the agent architecture that mimic those played by emotions in humans. The causal effects of emotions that can be modeled in artificial agents include self-awareness of emotions when they occur, albeit no claim is made that artificial agents, at least those that currently exist, are conscious of their emotions in the sense of having qualitative phenomenal experiences. In this perspective, this research program is akin to the original affective computing research program as stated by Picard (1997).

Such goals have been pursued following two main paths (Reisenzein et al., 2013): 1) formalizing emotion theories in implementation-independent formal languages (set theory, agent logics, e.g., Broekens et al., 2008); 2) modeling emotions using general cognitive architectures (such as Soar and ACT-R), general agent architectures (such as the Belief-Desire-Intentions, BDI, architecture) or general-purpose affective agent architectures (e.g., the EMA - EMotion and Adaption - architecture, Marsella and Gratch, 2009, or the FAtiMA architecture, Dias et al., 2014).

Appraisal and componential theories have been largely influential in this research program. Computational models of emotion had early success exemplified by Scherer's GENESE expert system. GENESE was built on a knowledge base that mapped appraisals to different emotions as predicted by Scherer's heuristics and theory (Scherer, 1993). However, despite its value as a tool to test emotion theories, the system was too limited for most real-world applications. It is too difficult, if not impossible, to construct the significant knowledge base needed for complex realistic situations. More recently, the OCC model by Ortony et al. (1990) has been widely adopted.

As to the formal treatment of emotion theories, fresh attempts have been proposed to bridge BDI and the OCC model of emotions by using probabilistic logic (e.g., Gluz and Jaques, 2017).

The BDI (Belief-Desire-Intention) model is a well known reasoning model in artificial intelligence research. It provides a high-level abstraction of human reasoning, allowing reasoning process modeling using only three mental states: belief, desire and intention, which represent, respectively, the informational, motivational and deliberative states of an intelligent agent. These agents are generally characterized by affective states such as emotions, mood or personality but sometimes also by affective capacities such as empathy or emotional regulation. (Sánchez-López and Cerezo, 2019)

As to architectures (at the implementation model level, in our perspective), a large number has been proposed over years (Ojha et al., 2021). A snapshot of their logic can be captured by considering the example of F FAtiMA (Fearnot AffecTIve Mind Architecture,Dias et al., 2014). This is an agent architecture with planning capabilities designed to use emotions and personality to influence the agent's behaviour. FAtiMA processing cycle is based on the following steps 1. perception of events; 2. on this basis, memory is updated and 3. appraisal processes are triggered; 4. an affective state is generated; 5. actions are executed based on goal-based planning (a BDI-style reasoning process); and 6. actions are executed. For a detailed up-to-date review of these architectures based on the appraisal theory of emotion, and related critical issues, see Ojha et al. (2021).

Clearly, although symbolic/cognitive architecture approaches are capable of solving a variety of AI tasks, they are limited with respect to learning from exploration and feedback in unstructured tasks. More recently, the effort of bridging the cognitive AI approach to machine-learning based affective computing - and, in perspective, deep learning techniques (Schuller and Schuller, 2018), has gained currency (e.g., in the reinforcement learning setting, see Moerland et al., 2018, for a detailed review and discussion)

5.4.3 The cognitive roboticist's way

The research landscape of cognitive robotics, markedly social robotics (Sheridan, 2020), is more nuanced with respect to the cognitive AI and ML-based affective computing fields, and a plurality of methods and approaches have been adopted by virtue of the very fact that roboticists are confronted with hard problems: actual bodies, real environments, computational time constraints (Yan et al., 2021; Dimitrievska and Ackovska, 2020; Sheridan, 2020; Hortensius et al., 2018; Cavallo et al., 2018).

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

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

Wiese et al. (2017) argue that the likelihood of humanoid robots being perceived as social companions can be increased by designing them in a way that they are perceived as intentional agents that activate areas in the human brain involved in social-cognitive processing, and provide an in-depth and wide review of how neuroscientific methods can contribute to make robots appear more social. Robots that are supposed to act as social interaction partners in the future need to fit in human-attuned environments by emulating human form and cognition. Indeed, psychological research has shown that anthropomorphism, and specifically mind perception, are highly automatic processes that activate social areas in the human brain (Wiese et al., 2017).

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

First attempts to build socially competent robots can be traced back to the work done at MIT Brook's robotics lab, e.g., Cog (Brooks et al., 1999) and Kismet (Breazeal and Scassellati, 1999; Breazeal, 2003). Kismet is still a paradigmatic example of the problems to be solved in this field and of the consequent methodological pluralism therein adopted; thus, it is worth being presented to some detail.

With Kismet, Breazeal and Scassellati (1999) studied how an expressive robot elicited appropriate social responses in humans by displaying attention and turn-taking mechanisms. They also identified some of the requirements of the visual system of such robots (Breazeal et al., 2001) as for example the advantages of foveated vision, eye contact, and a number of sensorimotor control loops (e.g., avoid and seek objects and people). Cognitive/social visual behaviour grounded in a motivation system which consisted of drives and emotions. The robot's drives represent the basic needs of the robot: to interact with people (the social drive); to be stimulated by toys and other objects (the stimulation drive); to rest (the fatigue drive). For each drive, there is a desired operation point, and an acceptable bound of operations around that point (the homeostatic regime). Unattended drives drift toward an under-stimulated regime. Excessive stimulation (too many stimuli or stimuli moving too quickly) push a drive toward an over-stimulated regime. When the intensity level of the drive leaves the homeostatic regime, the robot becomes motivated to act in ways that will restore the drives to the homeostatic regime.

The robot's emotions, in turn, were a result of its affective state. The affective state of the robot was represented as a point along three dimensions: arousal (i.e. high, neutral, or low), valence (i.e. positive, neutral, or negative), and stance (i.e. open, neutral, or closed). This core affective state is thus based on Russell's core affect concept (Russell, 2003). Operatively, the affective state is computed by summing contributions from the drives and behaviours. Percepts may also indirectly contribute to the affective state through the releasing mechanisms. Each releasing mechanism has an associated somatic marker processes, which assigns arousal, valence and stance tags to each releasing mechanism (a technique inspired by the Damasio's somatic marker hypothesis, SMH, in emotion neuroscience, Bechara et al., 2000; Bechara and Damasio, 2005). At the same time, this continuous state was partitioned into a discrete set of emotion regions, which roughly correspond to Ekman's discrete emotions (Ekman, 1993).

Since then, several empathic robots that consider the internal state of others for their own expressions have been proposed (Yan et al., 2021; Dimitrievska and Ackovska, 2020; Sheridan, 2020; Hortensius et al., 2018; Cavallo et al., 2018). If one glimpse over these reviews, it is readily apparent that roboticists have drawn on either BET and appraisal models, at the theoretical level, and exploited a wealth of cognitive AI / ML affective computing approaches, moving more recently to deep learning, at the implementation theory level. Also, there has been a turn currently towards psychological constructivism theories, mostly related to concept and abstract concept learning and the modelling of communication acts in dyadic interactions for learning (Cangelosi and Stramandinoli, 2018).

Indeed, there is a tradition of robotic research that utilised neuroscience studies as a starting point (Kawato, 1999; Scassellati, 2002; Demiris et al., 2014).

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

The key concept of shared sensorimotor representations, dating back to Liberman and Mattingly (1985), guided a variety of implementations utilising, for example, recurrent neural networks (Tani et al., 2004) or various other machine-learning methods that learn direct-inverse models from examples (but see, for a review, Oztop et al., 2006, 2013). Among these attempts to implement a mirror neuron system into artificial agents, some models were more neuroscientifically accurate than others. Inspired by the mirror neuron systems, Lim and Okuno (2014) proposed multimodal emotional intelligence (MEI), which utilises an integrated architecture to recognise the emotional states of others and generate its own emotional facial expressions.

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

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

Kim et al. (2013) have rectified their MEI/SIR original proposal using deep neural networks that learn to extract features for emotional categorisation from audio-visual signals. In their system, deep belief networks (DBNs) comprising restricted Boltzmann machines (RBMs, see Goodfellow et al., 2016) were used as unsupervised learning mechanisms. The RBM can abstract input signals and reconstruct the signals therefrom. In experiments, their model extracted emotion specific features from general ones, which are not always important for the classification of emotion.

More recently, Horii et al. (2016) proposes a model that can estimate human emotion and generate its own emotional expressions to imitate the human expressions based on the estimation of his/her emotion in human-robot interaction. The model overcomes two issues confronting the previous emotional model: constructing an emotional representation of multimodal signals for estimation and generation for emotion instead of using heuristic features, and actualising mental simulation to infer the emotion of others from their ambiguous multimodal signals. In the same vein of Kim et al. (2013), they employed RBMs to address these two issues as they are able to abstract input signals and recall the signals there from. The abstraction capability of RBMs was exploited to overcome the first limitation by reducing the dimensions of multimodal signals and associating the multimodal signals. The model also carries out mental simulation by exploiting the ability to generate sensorimotor signals. The mental simulation mechanism enables the model to estimate the emotional states of others from partial multi- modal expressions based on its own experiences. Related to this proposal, Boccignone et al. (2018a) have presented a theoretical Bayesian model of multimodal affect enactment, at the core affect level, triggered by facial expressions displayed in the course of an expresser/observer interaction. The goal of the model is to allow the observer to reach a core affect state similar to that of the expresser. Indeed, such condition is suitable to ground subsequent cognitive processing for affect understanding Adolphs (2002). The model accounts for mirroring, simulation-based mechanisms that are likely to be at the heart of face-based emotion understanding Adolphs (2002); Wood et al. (2016), and, more generally of affective interactions. Further, it can address a flexible, multimodal and embodied representation of perceived facial actions while lending itself well to the task of learning from few as possible examples (in the experiments reported here just one expresser example, much like as in a mother/infant interaction). The core implementation model was set from a predictive processing perspective and based on deep Gaussian processes architecture (Boccignone et al., 2018a).

Indeed, there is in cognitive robotics a growing interest in the predictive processing framework and its associated schemes, such as predictive coding, active inference, perceptual inference, and free-energy principle (for an in-depth review, see Ciria et al., 2021). Yet, to the best of our knowledge, there are no principled approaches addressing the bridge between communication acts and conceptual acts,integrating emotions and affect from a constructivism standpoint.

CHAPTER 6

The Model

In this chapter, taking stock of the discussion presented in previous chapters, we present the model that unifies in a probabilistic framework what we have named the communication and the conceptual acts.

To such end, following the methodology discussed in Chapter 3, we first devise the constraints at the neurobiological and the psychological description levels. As to the former, we will present in Section 6.2 a synthesis drawing on results of a vast and controversial literature. Many details are left in the Appendices. Then, a bridge to the psychological level will be provided in Section 6.4 in order to provide the necessary psychological infrastructure that underpins the theoretical model discussed in Section 6.5.

6.1 Model Overview

At the core level, we can straightforwardly frame our problem of an agent involved in some kind of interaction as an agent-in-context model Koban et al. (2021). Self-in-context models endow events with personal meaning and allow predictive control over behaviour and peripheral physiology, including autonomic, neuroendocrine and immune function. In brief (cfr., Fig. 6.1), over time, the agent is influenced at the conceptual level by the social and environmental context (social norms, relationships, cultural beliefs, neighbourhood characteristics, etc.). The agent is also shaped by beliefs, memories and learning. At the perceptual level the agent takes into account, both exteroceptive sensations from the world and interoceptive sensations from the body. Accordingly, the agent regulates his/her body's visceral physiology and behavioural outflow. To provide a principled framework for self-in-context models we shall resort to recent developments in computational neuroscience concerning the view of the brain as a predictive machine Zhang et al. (2019); Chanes and Barrett (2016); Barrett and Simmons (2015); Hutchinson and Barrett (2019), which will be discussed in the following.

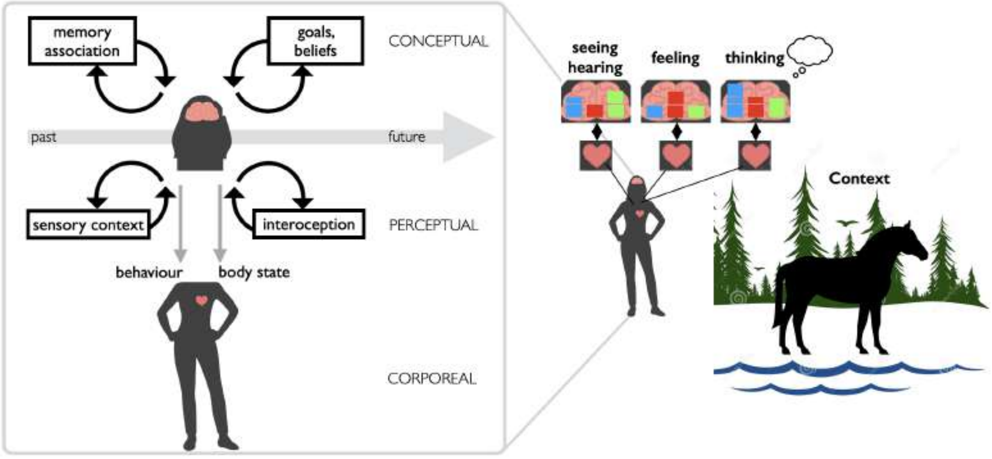


Figure 6.1: *Self-in-context model of the agent. Over time, the agent exhibits a dynamics which involves coordination at different levels: conceptual, perceptual, and body's behavioural/-physiological responses (see also Fig. 5.12). Adapted from Koban et al. (2021)*

In a nutshell, Figure 6.1 draws on Figure 5.12 (right side) and proposes a scheme (left side) where mental states, comprised of different combinations of psychological primitives (represented in red, green, and blue), are the result of the generative/predictive dynamics of brain and body coordination. Based on the combination and relative weighting of psychological primitives in a given instance, mental states might be recognized as seeing/hearing or thinking or feeling. We refer to the model presented in Figure 6.1 as the core model.

The core model is then subsequently extended so that the context might include at least another agent. This extended agent-in-context model, outlined in Figure 6.2 provides the infrastructure needed to account for all the elements involved by the communication act, first of all language capabilities and the minimal ability to gauge another agent's intentions and behaviour.

6.2 Structural Constraints: The Neurobiological Level (core)

As discussed in Chapter 3, the theoretical model, which casts our problem at the behavioural level, can be structurally constrained from the "bottom-up" by resorting to the underlying neural architecture. Obviously, even limiting to the last decade, there is a tremendously vast (and often controversial, Barrett and Satpute, 2019) amount of work concerning the neural underpinnings of language, emotions, and social interactions. It is out of the scope of this thesis a complete discussion.

At the neurobiological level the most suitable grain of analysis, to ground psychological primitives is that of functional brain areas, markedly, area networks (Figure 6.3), as usually specified by functional Magnetic Resonance Imaging (fMRI) studies (see Appendix D for a quick tour).

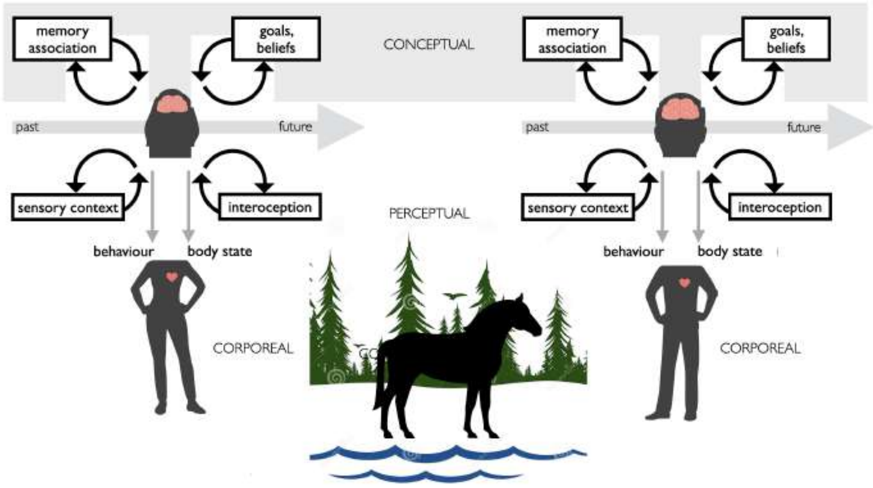


Figure 6.2: *Extending the self-in-context model (Fig.6.2). Two interacting agents exhibit, over time, a dynamics which involves both within-agent and between-agent activities that coordinate ate the different levels conceptual, perceptual, and body's behavioural/physiological responses*

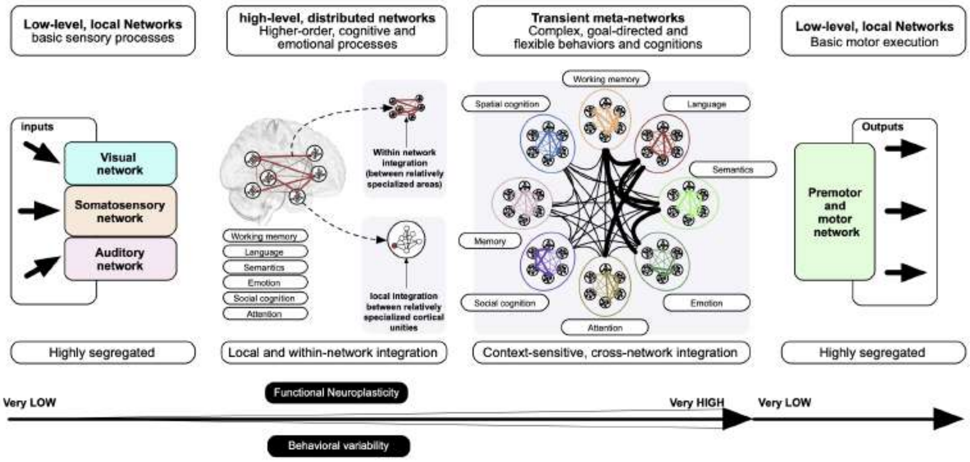

Low-level, local Networks
basic sensory processes
high-level, distributed networks
Higher-order, cognitive and emotional processes
Transient meta-networks
Complex, goal-directed and flexible behaviors and cognitions
Low-level, local Networks
Basic motor execution
Inputs
Visual network
Somatosensory network
Auditory network
Within network integration (between relatively specialized areas)
Working memory
Language
Semantics
Emotion
Social cognition
Attention
local integration between relatively specialized cortical unities
Working memory
Spatial cognition
Language
Semantics
Memory
Social cognition
Emotion
Attention
Outputs
Premotor and motor network
Highly segregated
Local and within-network integration
Context-sensitive, cross-network integration
Highly segregated
Functional Neuroplasticity
Very LOW
Very HIGH
Very LOW
Behavioral variability

Figure 6.3: *The meta-networking approach of complex cerebral functions. Basic sensory inputs are processed by anatomically highly segregated and local networks. Higher-order cognitive and emotional functions, or mental states, are rather sustained by cortical networks that are widely distributed at the whole-brain level. The information processed by the different cortical unities forming a relatively specialized cortical area is integrated locally (local integration); then, the information from each cortical area forming the specialized network is integrated globally (within-network integration). In the context of highly complex, goal-directed, and flexible cognitions/behaviors, the information coming from different specialized networks is integrated in a context-sensitive manner (i.e., as a function of current cognitive demands). This specific pattern of between-network integration corresponds to a functional meta-network transiently generated to reach a complex goal or to produce a flexible behavior. Adapted from Herbet and Duffau (2020)*

As we are about to see, the novel perspective of the predictive brain, is the most suitable to account for the overall framework on language and emotions that we have outlined in previous chapters. Further details can be found in Appendix E.

The basic assumptions of such perspective can be summarised as follows (Barrett, 2020, 2017b).

1. *The brain is not for thinking (or loving).* Its most important function is neither rationality, nor emotion, nor imagination, but rather to control the agent's body for managing allostasis, by predicting energy needs before they arise so that the agent can efficiently act and survive.

2. *There is one and only brain.* It has not phylogenetically evolved by adding layers as blueprints of specific functions (survival/reptilian brain → limbic/emotional brain → human neocortical/rational brain). Rather, brains of all species, follow a common "manufacturing plan", albeit developing at different size, organization and complexity.

3. *The brain is a network.* Basically, a network of 128 billion neurons, by and large, connected as a single, massive flexible structure, resulting in trillions of activity patterns, where connections become stronger or weaker depending on what is happening in the world and in the agent's body throughout life. No single neuron or area is the locus of a single psychological function (vision, touch, reasoning, memory, etc.). See Figure 6.3.

4. *The brain develops by wiring itself to its world.* Ontogenetically, infant's genes carrying neural wiring instructions, are guided and regulated by the surrounding physical and social environments that help in tuning and pruning neural connection in order to manage the body budget (cfr., Hyp. 1,3). In such endeavour, a brain becomes optimized for the particular environment in which it develops.

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

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

7. *One brain makes more than one kind of mind.* Agents come into the world with a basic brain plan that can be wired in a variety of ways (Hyp. 4). Beyond the individual, micro-wiring is tuned and pruned by social groups and culture.

8. *Brains can create reality*. Boundaries between social and physical realities is porous (e.g., studies showing that people judge wine as tasting better when expensive). Brains do not just select information from the environment, but by creating categories add new functions to the world. These are communicated and shared with other brains and weaved into the world to become part of the social environment, which, in turn, will help to wire novel brains (cfr., Hyp. 4,6)

These assumptions are difficult to reconcile with the classical view of "faculty psychology"[1], where specific mental functions are assumed to be instantiated in a given brain area, (or network of areas). For instance, limbic regions, such as portions of the cingulate cortex, orbitofrontal cortex, entorhinal cortex, and anterior insula (together with other subcortical regions, such as the amygdala), are considered to "host" emotions (see Figure 6.4); in the most extreme version, some of these regions might be the organs of specific emotions: the amygdala for fear, the insula for disgust, and so on.

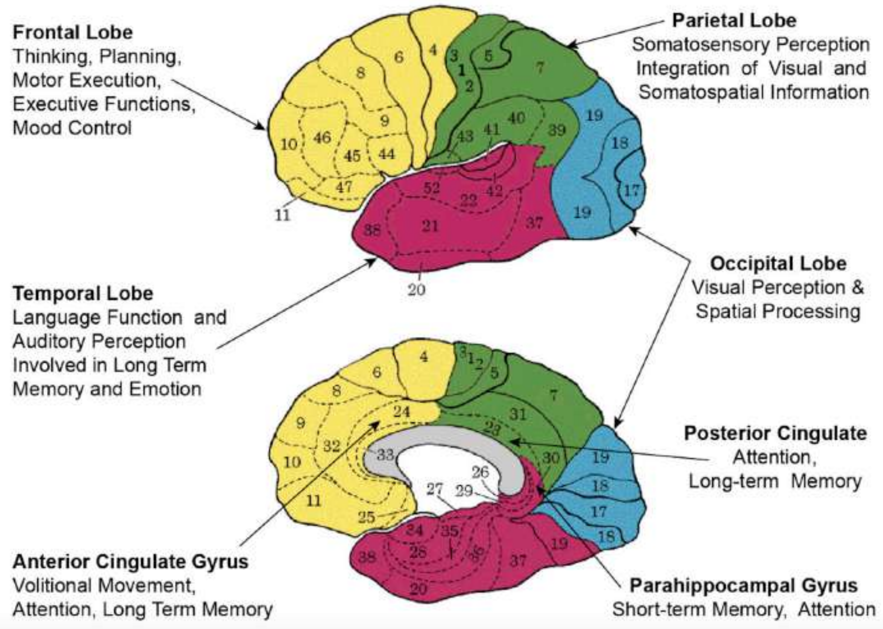

Frontal Lobe
Thinking, Planning, Motor Execution, Executive Functions, Mood Control
Parietal Lobe
Somatosensory Perception Integration of Visual and Somatospatial Information
Occipital Lobe
Visual Perception & Spatial Processing
Temporal Lobe
Language Function and Auditory Perception Involved in Long Term Memory and Emotion
Posterior Cingulate
Attention, Long-term Memory
Anterior Cingulate Gyrus
Volitional Movement, Attention, Long Term Memory
Parahippocampal Gyrus
Short-term Memory, Attention

Figure 6.4: *The classical view of the organization of the human cortex. For more than a century, neuroscientists have studied the cerebral cortex by delineating individual cortical areas and mapping their function. Numbers identify the Brodmann areas: these were originally defined and numbered by the German anatomist Korbinian Brodmann based on the cytoarchitectural organization of neurons he observed in the cerebral cortex*

However, a different perspective can be taken. An elegant account for grounding Hypotheses 1, 2, 3 and 5 is provided by the Structural Model (SM, see Appendix E for details), a model of cortical systematic variation (García-Cabezas et al., 2019; Barbas and García-Cabezas, 2016).

[1] In faculty psychology, the mind is thought to consist of functionally encapsulated mental faculties or "mental organs" (akin to the organs of the body), each with a specific and distinct physical cause. In modern neuroscience, this view manifests as the hypothesis that a specific faculty corresponds consistently and specifically to increased activity in a given brain area, network of areas, or even in specific neurons.

The SM describes how agranular cortical areas, regardless of their placement on the cortical mantle, modulate granular cortices through feedback connections and, in the opposite direction, granular cortices project into agranular cortices via feedforward pathways. This way the functional architecture of the brain is derived in terms of cortical types (agranular, dysgranular, eulaminate, etc) rather than spatial cortical placement of specific areas (each associated to a specific function, as suggested by the classical view, cfr. Figure 6.4).

Chanes and Barrett (2016) argued that feedback and feedforward communication flows can be conceived in terms of prediction and prediction-error signalling, respectively (see Figure E.3, Appendix E for details). In a crude summary, predictions flow from cortical regions with less laminar differentiation to regions with increasing laminar differentiation (e.g., from limbic cortices to motor, interoceptive, and primary somatosensory, auditory, and visual cortices); prediction-errors are obtained in the opposite, feedforward direction (Hutchinson and Barrett, 2019; Chanes and Barrett, 2016). Limbic cortices, plus the hippocampal areas, are the source of prediction signals, driving action and perception in an inferential way concerned with energetics, not just during episodes of emotion, but during all mental events (Hyp. 1). According to the SM, these have anatomical features that place them at the top of a predictive architecture and provide the initial representations of prediction signals that propagate throughout the cortex (Hutchinson and Barrett, 2019; Chanes and Barrett, 2016). The limbic ensemble, via a series of connections to the hypothalamus and throughout the brain stem, is also responsible for regulating the body's global energy budget via control of the autonomic nervous system, the neuroendocrine and neuroimmune systems, and the other systems of the body's internal milieu. Such ensemble is thought to regulate the body by anticipating its needs and attempting to meet those needs before they arise, namely the process of allostasis (Schulkin and Sterling, 2019). Efficient energy regulation and metabolism are at the core of the brain's internal model.

The overall result of the approach, in a nutshell, is that the brain hosts an internal model of the world from the perspective of its body's physiological needs. Over time, predictions/simulations are generated and issued from the top of the hierarchy to lower levels. At any level, comparisons between prediction signals and ascending sensory input results in prediction error that is available to update the brain's internal model so to provide efficient body regulation.

Due to the complexity of the system, the functional architecture of such predictive machine (see Figure 6.5) is best described in terms of intrinsic networks, each including areas with varying degree of laminar differentiation. To recap the analysis detailed in Appendix E in terms of the levels presented in Figure 6.1:

- **Conceptual level**. The default mode network (DMN), frontoparietal central executive network (CEN) and the limbic network (LN) cooperating with the salience network (SN) account for conceptualization based on context and background knowledge mostly retained in DMN/parahippocampal areas in the medio-temporal lobe (MTL). In particular, the DMN conceptualizes perceptual input for the body and from exteroceptive sensory systems based on past experience. The DMN also initiates simulations and represents part of their pattern; its multimodal sensorimotor summaries become more detailed and particularized as they cascade out to primary sensory and motor regions. The CEN represents multimodal prior expectations (goal states) and the associated top-down prediction can amplify or suppress activity in other cortical systems based on current goals. It sculpts and maintains simulations for longer than the several hundred milliseconds required to process imminent prediction errors. It may also have a role in managing sensory prediction errors, by applying attention to select those body movements that will generate the expected sensory input (on the basis of subcortical cerebellar and striatal prediction errors). These movements then generate the sensory inputs that reduce prediction error and confirm an existing prediction.

- **Perceptual level**. The somatic interoceptive system represents and regulates current somatic / visceral states; the exteroceptive system represents and controls perceptions of the external world flowing from main peripheral perceptual networks (visual, auditory, etc.). Signals forwarded bottom-up from the body, are taken into account by the somatomotor network (SMN) (which includes primary and secondary somatosensory regions within parietal cortex and posterior insula, and premotor/motor cortex regions, among others). The SN (the anterior insula being the most relevant area) sends predictions that adjust the internal model to the conditions of the sensory periphery. The SN anticipates which prediction errors are likely to be allostatically relevant and therefore worth the metabolic cost of encoding and consolidation, and then modulates the gain on those errors accordingly.

- **Corporeal level**. This level accounts for the behavioural/physiological actions of the body (thus, both external, oriented towards the surrounding world, and internal). These are mediated by a complex variety of subcortical and peripheral control and sensing nuclei and ganglia. These involve the skeletal-motor system, the autonomic nervous system, the neuroendocrine and neuroimmune systems, and general reward system. Of particular relevance is the role of the amygdala signalling uncertainty about the predicted sensory input (via the basolateral complex) that help to adjust physiological functions in support of allostasis. The arousal signals that are associated with increases in amygdala activity can be considered as learning signals, at the core of what is usually named automatic emotion attention. However, it is worth recalling that, in the framework of the predictive brain, information flowing from the amygdala to the cortex is not emotional *per se*. The amygdala signalling activity (in either direction) has the general role of assigning higher weight to prediction-error signals estimated to have higher reliability in the current context.

The resulting neurobiological functional architecture that substantiates the self-in-context model presented in Figure 6.1 (left side) is shown in Figure 6.5.

6.2.1 What about emotion?

It is clear from the overall picture we have so far outlined, that it makes no sense questing for the neurobiological basis of emotion, in terms of identifying a defined set of brain areas or a circumscribed network suitable to provide the emotion “fingerprint” in the brain.

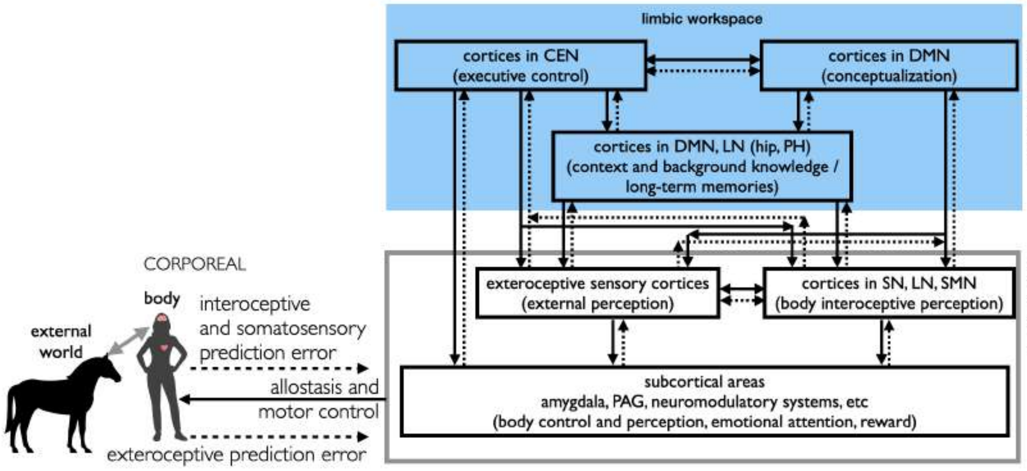

limbic workspace
cortices in CEN (executive control)
cortices in DMN (conceptualization)
CONCEPTUAL
cortices in DMN, LN (hip, PH) (context and background knowledge / long-term memories)
CORPOREAL
body
external world
interoceptive and somatosensory prediction error
allostasis and motor control
exteroceptive prediction error
exteroceptive sensory cortices (external perception)
cortices in SN, LN, SMN (body interoceptive perception)
PERCEPTUAL
subcortical areas
amygdala, PAG, neuromodulatory systems, etc
(body control and perception, emotional attention, reward)

Figure 6.5: *A neurobiological view of the predictive/generative architecture of the agent-in-context, from the top level of the limbic ensemble (enacting the conceptual level) to the bottom level of body's behaviour and internal states (corporeal level). Solid arrows denote (top-down/feedback) prediction signals; dashed arrows, (bottom-up/feedforward) prediction error signals. The first two levels map (cortical components of) the main intrinsic networks recruited at each level: DMN, default mode network; CEN, frontoparietal central executive networ; LN, limbic network; SN, salience network; SMN, somatosensory network; hip, hippocampus; PH, parahippocampal formation (at the intersection of DMN and LN). For sake of clarity, motor networks, dorsal and ventral attention networks, and perceptual networks (auditory, visual, etc) have been omitted*

Surprisingly, the core architecture devised in Figure 6.5 will suffice (apart from the linguistic process that we shall discuss later on) to account for emotion too.

As previously outlined in Figure 5.12, emotion is just one possible state of mind among other states that can be experienced (cognition, perception, sense of self, etc) and result from the different combinations and relative weighting of psychological primitive processes (such as categorization, core affect and so on) in a given instance of time (Hyp. 1: brain is not for thinking or loving).

On one side we have sensations, from the external world and from the inner body; on the other side, we have a categorization process that drives predictions (see Figure 5.10). The core business of the brain is to make predictions in order to give sensations meaning in the service of allostatis. Emotion like fear or happiness might be recognized as such depending on the categorization at a certain time and on the weighting/relevance assigned to the core affect state (valence and arousal), which is in turn based on interoception and exteroception. Emotion is just the name for psychological moments in which the focus is on understanding what the internal sensations from the body represent.

In Section 5.2 we have introduced and defined the psychological primitive named as "core affect". In modern psychological usage, "affect" refers to the mental counterpart of internal bodily representations associated with emotions, actions that involve some degree of motivation, intensity, and even personality dispositions. In the science of emotion, "affect" is a general term that has come to mean anything emotional. A cautious term, it allows reference to something's effect or someones internal state without specifying exactly what kind of an effect or state it is. 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 for instance presented in Figure 6.6, 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 mod- alities, including the anterior insula that represents interoceptive sensations.

- *Visceromotor network*: it is part of a functional circuit that guides autonomic, endocrine, and behavioral responses to an object. It includes the medial portions of the OFC (extending into what is sometimes called the vmPFC), as well as subgenual and pregenual areas of the ACC, with robust reciprocal connections to all limbic areas (including many nuclei within the amygdala, and the ventral striatum), as well as to the hypothalamus, midbrain, brainstem, and spinal cord areas that are involved in internal-state regulation. These areas modulate changes in the viscera associated with the autonomic nervous system (including tissues and organs made of smooth muscle, such as the heart and lungs) and neuroendocrine

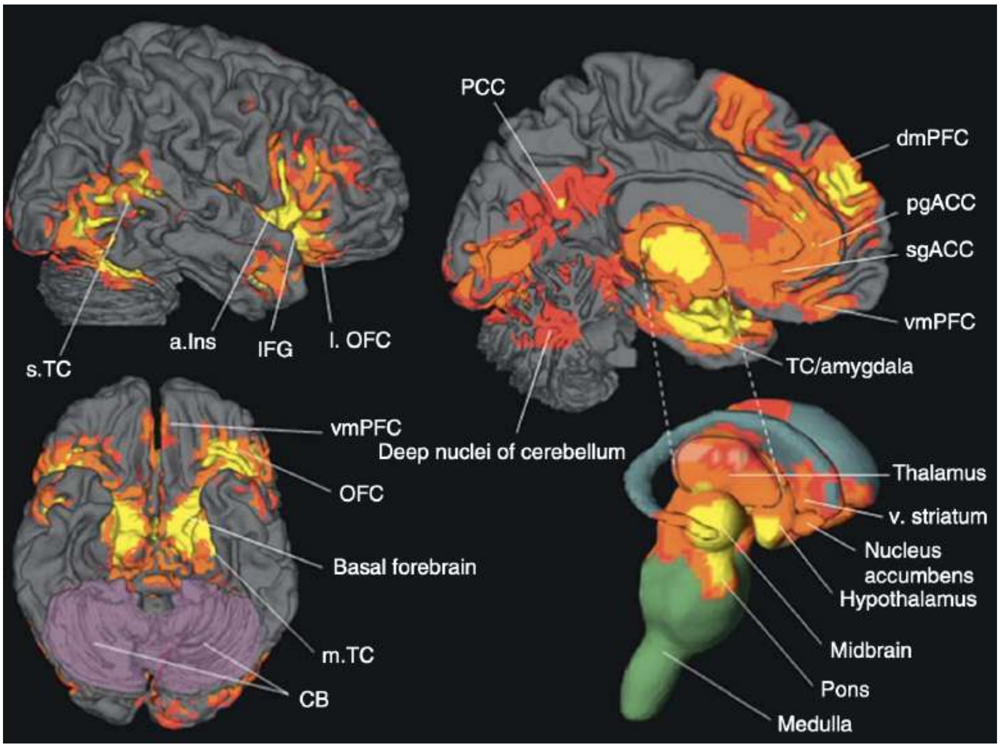


Figure 6.6: *Neural reference space for core affect. 165 neuroimaging studies of emotion (58 using PET and 107 using fMRI) summarized in a multilevel meta-analysis to produce the observed neural reference space for emotion. These areas include (from top left, clockwise) anterior insula (aIns), lateral OFC (lOFC), pregenual cingulate cortex (pgACC), subgenual cingulate cortex (sgACC), ventral medial prefrontal cortex (vmPFC), temporal cortex/amygdala (TC/Amygdala), thalamus, ventral striatrum (v Striatum), nucleus accumbens, hypothalamus, midbrain, pons, medulla, OFC, and basal forebrain. Other areas shown in this figure (e.g., inferior frontal gyrus (IFG), superior temporal cortex (sTC), dorsal medial prefrontal cortex (dmPFC), posterior cingulate cortex (PCC), medial temporal cortex (mTC), and cerebellum (CB)) relate to other psychological processes involved with emotion perception and experience. From Barrett and Bliss-Moreau (2009)*

> 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 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 which 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 categorized as that of feeling an emotion.

In a Wundtian sense, affect is a feeling state that is a fundamental ingredient of the human mind, a psychological primitive. Affect and sensation are two sides of the same mental coin (Barrett and Bliss-Moreau, 2009). As such the core affect provides a source of attention in the human brain (where attention is defined as anything that increases or decreases the firing of a neuron). Affect does not reveal what in the world has changed, where the change is, or what to do about it. Rather, it is just a quick and dirty sixth sense that something has happened. That particular something may be outside the body and require a rapid and energetically costly response (Barrett and Quigley, 2021).

This implies that core affect has an important role to play in normal perceptual functioning, including consciousness. When sensory information from the world sufficiently influences a person's internal bodily state, the processing of that information is prioritized so that the resulting object is more easily seen and remembered. In brief, core affect is a fundamental feature of conscious experience.

More generally, by scrutinizing areas depicted in Figure 6.6 and by taking stock of what we have outlined so far concerning the structure of the brain and the role played by intrinsic networks, we can find such neural reference space as defined by the interplay of the two main networks accounting for allostasis and interoception: the salience network (SN) and the default mode network (DMN) The DMN and SN are concurrently regulating and representing the internal milieu, while they are routinely engaged in a wide range of tasks spanning cognitive, perceptual and emotion domains, all of which involve value-based decision-making and action. Tract tracing and functional connectivity studies suggest that the DMN and salience network share hubs in the thalamus, hypothalamus, amygdala, and ventral striatum; areas that themselves control visceral activation. This suggest (Kleckner et al., 2017) that whatever other psychological functions the DMN and the SN are performing during any given brain state, they are simultaneously maintaining or attempting to restore allostasis and are integrating sensory representations of the internal milieu with the rest of the brain. In other terms, the default mode and salience networks create a highly connected functional ensemble for integrating information across the brain, with interoceptive and allostatic information at its core, even though it may not be apparent much of the time.

On the other hand, considering the top-level of mental abstraction, DMN may be involved in emotion by supporting conceptualization. Conceptualization requires generation of an internal prediction about the meaning of internal and external sensations and behaviors given the present context. To such end it must extend across levels of abstraction, spanning across sensory features, multimodal sensory information, and more abstract levels (e.g., dimensions of meaning). It has been proposed by Satpute and Lindquist (2019) that the DMN plays a role in representing discrete emotions because it abstracts across instances with heterogeneous features. The neuroanatomical properties of the DMN support its role in abstraction and recent findings findings are consistent with the idea that the DMN drives autonomic and visceromotor activity to be situationally appropriate with respect to the more abstract adaptive themes associated with a given emotion category (Satpute and Lindquist, 2019). Increased connectivity occurs between DMN and salience nodes during discrete emotions and may reflect information processing as DMN uses conceptualizations (i.e., draws on prior experience and knowledge) to drive somatovisceral activity in a given context.

6.3 Structural Constraints: The Neurobiological Level (extended)

In order to fully address the hypotheses stated ad the beginning of this Section, we need to extend our core architecture to account for the capability of the brain to grow by wiring to the world and function as a "social brain" interacting with other brains (hyp. 4, 6, 7, and 8). In other terms, the self-in-context model must be extended so that the context might include at least another brain (Figure 6.2).

As we have discussed in Section 4.2.4, we need a neurobiological infrastructure of shared intentionality, which, either phylogenetecally and ontogenetically (Figures 4.7 and 5.1), has developed from gestures to language (in humans). The bare essentials of the neurobiology of language, in general, and of semantics and pragmatics are presented in Appendix F.

To recap the key points:

1. The language network plausibly includes a functionally specialized "core" (brain regions that coactivate with each other during language processing, also named the perisylvian system) and a domain-general "periphery", namely a set of brain regions that may coactivate with the language core regions at some times but with other specialized systems at other times, depending on task demands Hertrich et al. (2020);

2. The extended language network (Figure F.6) is suitable to underpin the brain semantic network; in particular, semantic processing might be accounted for through the cooperation of three networks (Xu et al., 2017): (1) the perisyl-

vian "language-supported system" (partially overlapping with the core language network), (2) the "multimodal experiential system" also addressed as the DMN, integrating experience-based knowledge across multiple modalities, and (3) the left-dominant frontoparietal CEN as a semantic control system;

3. Agents interacting through language couple through their DMN activity; the DMN - that integrates over time incoming external information with intrinsic information (long-term memories, LTMs, conditional responses, beliefs) and exhibits a tight connection with the theory of mind (ToM) networks, which is in turn synchronized with the language network during language comprehension (Hertrich et al., 2020; Xu et al., 2017) - provides a space for negotiating a shared neural code necessary for establishing shared meanings, shared communication tools and narratives (Yeshurun et al., 2021) at the conceptual level (cfr., Figure 6.2).

Points 1 and 2 can be schematized as the extended neurobiological architecture of Figure 6.7.

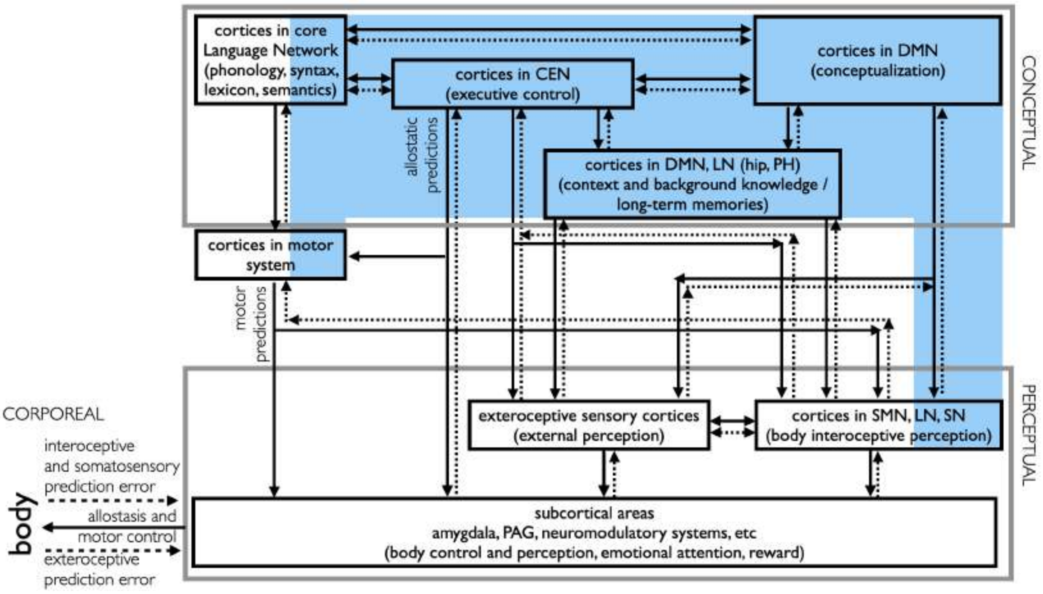

cortices in core Language Network (phonology, syntax, lexicon, semantics)
cortices in CEN (executive control)
cortices in DMN (conceptualization)
CONCEPTUAL
allostatic predictions
cortices in DMN, LN (hip, PH) (context and background knowledge / long-term memories)
cortices in motor system
motor predictions
CORPOREAL
interoceptive and somatosensory prediction error
allostasis and motor control
exteroceptive prediction error
body
exteroceptive sensory cortices (external perception)
cortices in SMN, LN, SN (body interoceptive perception)
PERCEPTUAL
subcortical areas
amygdala, PAG, neuromodulatory systems, etc
(body control and perception, emotional attention, reward)

Figure 6.7: *Extending the infrasctructure of Fig. 6.5. Core language area/hubs interact with main intrinsic networks DMN, CE, SN and the motor system in order to frame phonological/synctactical capabilities within agent's semantics and pragmatics at the conceptual level. Now, brain's conceptualization processes can rely on the language neural reference space (lexicon and semantics) for handling and updating available categories and for shaping novel ones*

By taking considering point 3), the functional architecture outlined in Figure 6.7 can be easily exploited to address the case of two interacting agents as in Figure 6.8.

Eventually, this modelling step fully provides the neural infrastructure necessary to account for Hypotheses 6,7 and 8.

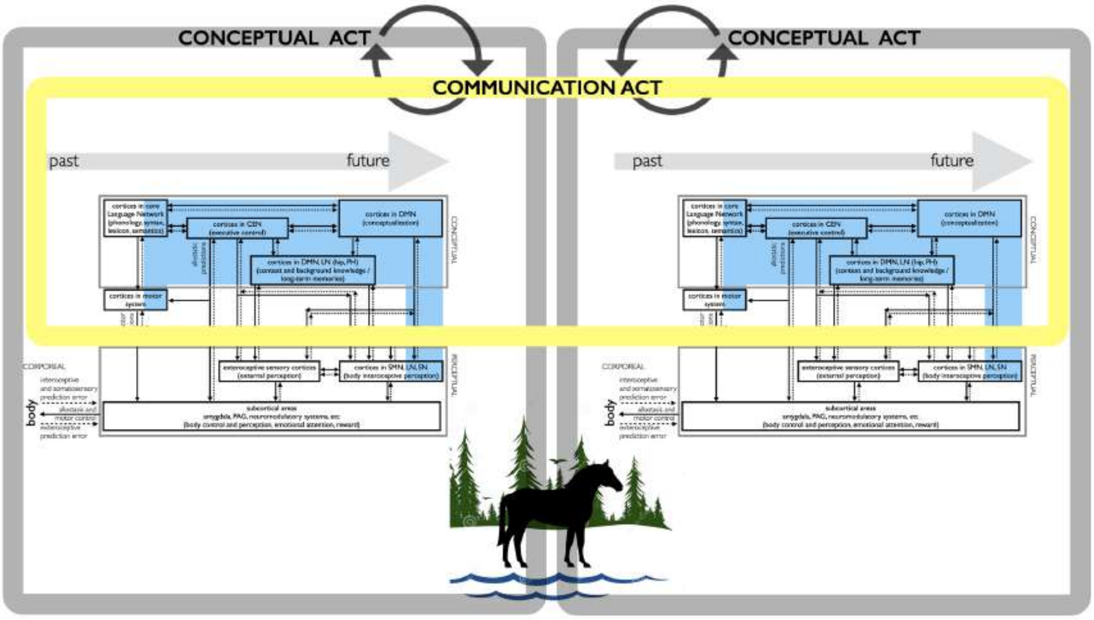

CONCEPTUAL ACT
CONCEPTUAL ACT
COMMUNICATION ACT
past
future
cortices in core Language Network (phonology, syntax, lexicon, semantics)
cortices in CEN (executive control)
cortices in DMN (conceptualization)
allostatic predictions
cortices in DMN, LN (hip, PH) (context and background knowledge / long-term memories)
CONCEPTUAL
cortices in motor system
CORPOREAL
exteroceptive sensory cortices (external perception)
cortices in SMN, LN, SN (body interoceptive perception)
PERCEPTUAL
interoceptive and somatosensory prediction error
allostasis and motor control
exteroceptive prediction error
body
subcortical areas amygdala, PAG, neuromodulatory systems, etc. (body control and perception, emotional attention, reward)
past
future
cortices in core Language Network (phonology, syntax, lexicon, semantics)
cortices in CEN (executive control)
cortices in DMN (conceptualization)
allostatic predictions
cortices in DMN, LN (hip, PH) (context and background knowledge / long-term memories)
CONCEPTUAL
cortices in motor system
CORPOREAL
exteroceptive sensory cortices (external perception)
cortices in SMN, LN, SN (body interoceptive perception)
PERCEPTUAL
interoceptive and somatosensory prediction error
allostasis and motor control
exteroceptive prediction error
body
subcortical areas amygdala, PAG, neuromodulatory systems, etc. (body control and perception, emotional attention, reward)

Figure 6.8: *Neurobiological infrastructure of two interacting agents-in-context. Agents' brains synchronize (via the DMN, which integrates over time incoming external information with intrinsic information). Each agent performs a conceptual act, based on interoceptive and exteroceptive (visual and auditory) sensation, and undertaken actions (external and internal) in order to regulate her body budget. The communication act is one such action. Spoken and heard utterances are such that the communication act between agents regulates agent's conceptual acts and vice versa. This way, agents regulate one another's body budgets through their interactions, create communicate and share categories, even add new functions to the world (Hyp. 6,7,8)*

6.4 Structural Constraints: The Psychological Level

The setting of our problem at the psychological level entails the difficult and controversial issue of providing a mapping between the brain and the mind. As James (1890) put it: "A science of the relations of mind and brain must show how the elementary ingredients of the former correspond to the elementary functions of the latter".

In our case, as sketched in Figure 6.9 we need to bridge the neurobiological infrastructure defined in Figure 6.8 with the psychological structure of the agent-in-context model, to address how agents mental states are originated (Figure 5.12).

The constructionist view we have embraced in this thesis offers suitable means fur such mapping (Barrett, 2009).

On the one hand, we have made clear that psychological faculties such as emotion cognition, perception, sense of self are not natural kinds, that is they cannot be ontologically reduced (one-to-one mapped) to the activity of circumscribed brain regions. Such complex, discrete psychological moments are experienced by the agent (and "recognized" by an external observer) as the result of the continuous fundamental brain dynamics: making predictions in the service of allostasis by combining three sources of information: sensory stimulation from the world outside the skin (exteroceptive sensations), sensory signals captured from within the body (interoceptive sensations), and prior experience that the brain makes available by the reactivation and reinhibition of sensory and motor neurons.

On the other hand, we have seen so far that the brain contains a set of intrinsic networks that can be understood as performing domain-general operations. These operations serve as the functional architecture for how mental events and behaviors are constructed.

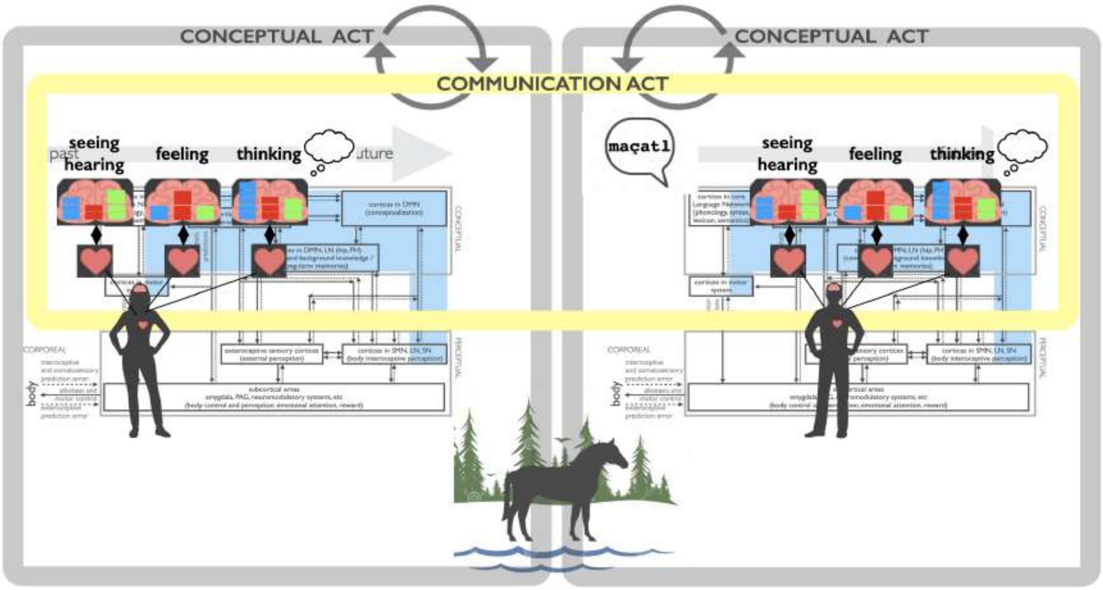


Figure 6.9: *Moving to the psychological level: the problem of brain/mind mapping*

Under such circumstances, the following hypotheses can be assumed:

1. the mind is realized by the continual interplay of more basic primitives that can

be described in psychological terms;

2. all mental states (however categorized: emotion, perception, cognition, etc.) can be mapped to these basic psychological primitives;

3. basic psychological primitives are functional abstractions of fundamental networks in the brain (or, more precisely, functional motifs of such networks[2]);

4. depending on the combination and the relative weighting (e.g., via the focus of attention) of psychological primitives in a given instance, mental states can be categorized as seeing or thinking or feeling

As represented in Figure 6.10, complex psychological constructs, like emotion, cognition and social cognition, emerge as a weighted mixture of a number of primitives, such as executive function, motor movements, conceptualization, and so on. In other terms, complex constructs are actually better understood as arising from a smaller set of common computational building blocks, with prediction-related processing at the core. For instance, interoception, as a result of somatovisceral integration, which is usually experienced as a low-dimensional form of valence and arousal core affect, might be better thought of as properties of consciousness, rather than properties of emotional episodes per se. In this perspective, all psychological events exist in affective space: thus, all words have affective connotations and even putatively neutral objects are experienced with subtle affective features.

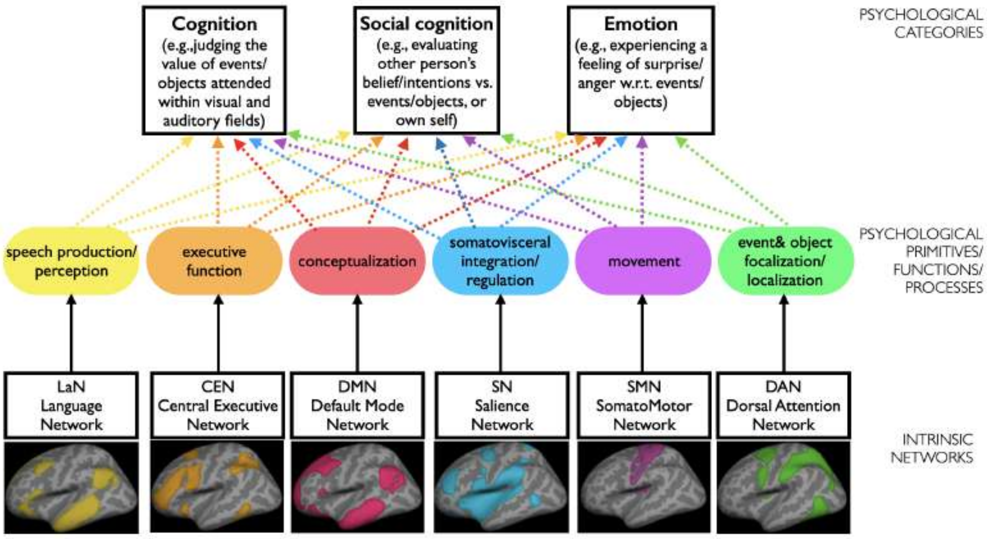

Cognition
(e.g.,judging the value of events/ objects attended within visual and auditory fields)
Social cognition
(e.g., evaluating other person's belief/intentions vs. events/objects, or own self)
Emotion
(e.g., experiencing a feeling of surprise/ anger w.r.t. events/ objects)
PSYCHOLOGICAL CATEGORIES
speech production/ perception
executive function
conceptualization
somatovisceral integration/ regulation
movement
event& object focalization/ localization
PSYCHOLOGICAL PRIMITIVES/ FUNCTIONS/ PROCESSES
LaN Language Network
CEN Central Executive Network
DMN Default Mode Network
SN Salience Network
SMN SomatoMotor Network
DAN Dorsal Attention Network
INTRINSIC NETWORKS

Figure 6.10: *Bridging levels: from intrinsic networks up to complex psychological categories. For clarity's sake, some intrinsic networks have been omitted (e.g., visual network, limbic network, etc.)*

[2] A *structural motif* consists of a set of brain areas and pathways that can potentially engage in different patterns of interactions depending on their degree of activation, the surrounding neural context or the behavioral state of the organism. *Functional motifs* represent elementary processing modes of a network and refer to specific combinations of nodes and connections contained within structural motifs that may be selectively recruited or activated in the course of neural information processing (Sporns et al., 2004)

Based on this assumption, and constrained by the hierarchy of neural components previously schematized in Figure 6.7, we can outline, at the psychological level, the infrastructure presented in Figure 6.11. The infrastructure specifies both the different representations involved and the primary processes that generate/act upon them, the latter abstracting main functional motives of intrinsic networks (Figure 6.10)

The infrastructure is suitable to support the general hypothesis that complex mental phenomena are constructed from more basic components or ingredients in accord with the psychological constructionist view. Perceptions, cognitions, emotions, memories, etc. are superordinate categories, not mechanisms, that describe and organize mental phenomena. The identified processes (colored ovals) are coordinated in the unifying predictive/generative dynamics of the mind unfolding in time in order to give meaning to external and internal perceptions and to program appropriate actions for self-regulation.

As a consequence, performing, for instance, a complex action such as a speech act will involve the intertwining of many primitives such as the conceptualization of exteroceptive and interoceptive sensations, with respect to a context while focusing on a specific event/object within the context, executive control to pursue the goal behind speaker's own communicative intentions framed by the evaluation of listener expectation/intentions (theory of mind), motor control of speech production and accompaning gestures/body behavior.

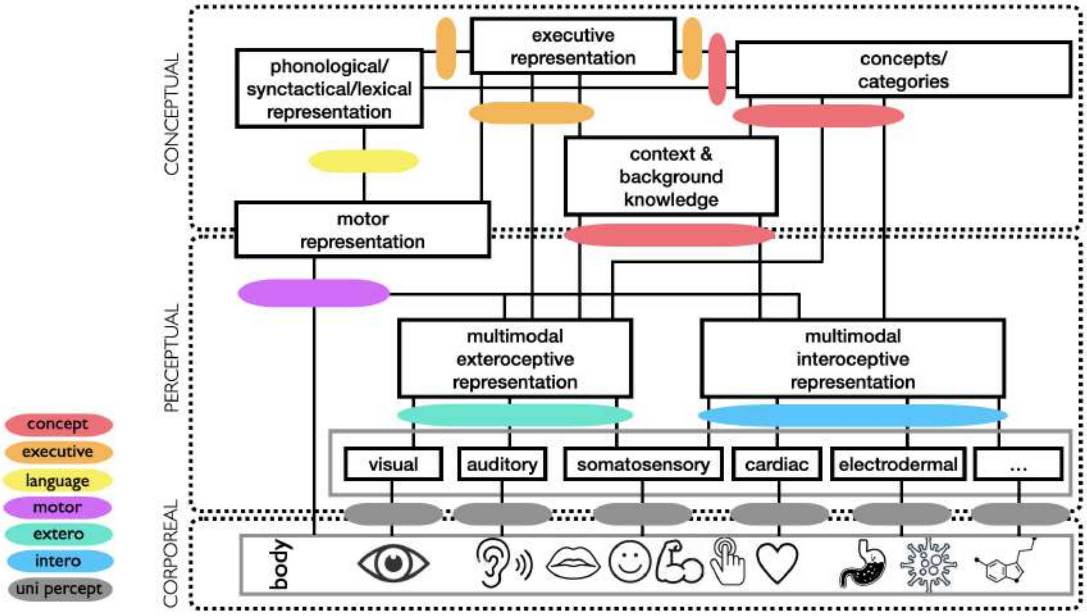


Figure 6.11: *The psychological infrastructure. The main representations (boxes) and the processes (ovals, see legend on the left) involving such representations at the different levels (conceptual, perceptual, corporeal). Processes include the "primitives" introduced in Figure 6.10 and the intrinsic unimodal perceptual networks (grey ovals: visual network, auditory network, etc.) Attention/focusing primitives and representations have been omitted for simplicity.*

Meanwhile, beyond naturally underpinning the communication act, language plays a constitutive role in the conceptual act. Words and the concepts they refer to, when viewed as situated conceptualizations, contribute to create emotion and memory (here, as a mental state, simply defined as subjectively remembering something that happened). Indeed, narratives are considered in psychology to generate mental models in people's minds (rather than represent a set of logical propositions). Narratives create expectations that guide how incoming information is processed. A situated concept, as activated by a word (for instance the emotion of anger through the word "anger"), also creates a brain state (composed of expectations) for how to process future information. This general idea extends to the use of both rich narratives and simple, singular words. Crucially, these expectations are not simply posthumous, after an event occurred. Rather, they can infiltrate consciousness itself. Language, therefore, plays a constitutive role for categories in general (emotion, memory, the self etc.) by shaping which expectations, and deviations from expectations (or prediction errors), occur during the creation of mental phenomena. It activates different top–down predictions, that are generally experienced as semantic or episodic memory, about bottom–up information, either ongoing processing, or incoming sensory input both from the body and the world.

Eventually, to sum up, we have all the ingredients to support both the communication and perceptual acts as summarised in Figure 6.12, and all necessary constraints to devise the theoretical model.

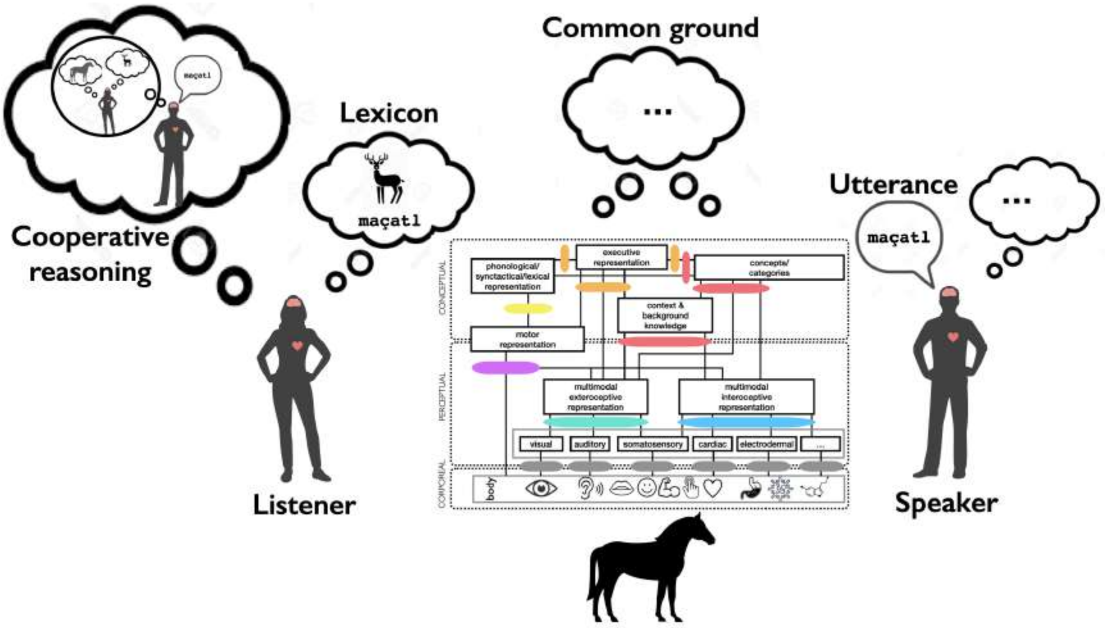


Figure 6.12: *The common psychological infrastructure of the speaker and the listener supporting both the communication and perceptual acts.*

6.5 Theoretical Model

At the most general level, we are interested in reasoning about the state of an agent (a speaker or a listener) as it evolves over time, in terms of a system state whose value at time t is a snapshot of its relevant attributes, hidden or observed, at that time.

Due to the stochastic nature of the system at hand, we represent the agent-in-context state $\mathcal{S}_t$ at time t as a collection of RVs $\mathcal{S}_t$ and we denote by $\mathcal{S}_{t_1:t_2}$ the random process $\{\mathcal{S}_t : t \in [t_1, t_2]\}$ indexed over the subset of reals $[t_1, t_2]$.

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.

We then introduce two simplifying assumptions.

Our first simplification is to discretise the timeline into a set of time slices, that are measurements 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 to $\mathcal{S}_0, \mathcal{S}_1, \cdots$, where $\mathcal{S}_t$ are the ground random variables that represent the system state at time $t \cdot \Delta$. Without loss of generality, this assumption simplifies our problem from representing distributions over a continuum of RVs to representing distributions over countably many RVs, sampled at discrete intervals.

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

$$P(\mathcal{S}_{0:T}) = P(\mathcal{S}_0) \prod_{t=0}^{T-1} P(\mathcal{S}_{t+1} \mid \mathcal{S}_{0:t}) \tag{6.1}$$

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

The second simplification entails the Markovianity of the process. A dynamic system over the variable $\mathcal{S}$ satisfies the Markov assumption if, $\forall t \geq 0$,

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

The Markov assumption allows us to define a more compact representation of the distribution in Eq. 6.1:

$$P(\mathcal{S}_{0:T}) = P(\mathcal{S}_0) \prod_{t=0}^{T-1} P(\mathcal{S}_{t+1} \mid \mathcal{S}_t). \tag{6.2}$$

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

Next, we need to specify the state of the agent $\mathcal{S}$.

Use the following time-indexed collections or *ensembles* of RVs.

- $\mathcal{C}_t$: the ensemble of RVs denoting the conceptual states;
- $\mathcal{E}_t$: the ensemble of RVs denoting the exteroceptive states;
- $\mathcal{I}_t$: the ensemble of RVs denoting the interoceptive states;
- $\mathcal{A}_t$: the ensemble of RVs indicating the states supporting both motor and allostatic actions and planning

- $\mathcal{X}_t^i$: the ensemble of RVs standing for the low-level states of action directed at the regulation of the internal milieu;
- $\mathcal{X}_t^e$ the ensemble of RVs representing the low-level states of actions directed towards the external world;
- $\mathcal{O}_t$: the ensemble of RVs standing for the possible outcomes of low-level body "sensors" capturing signals either from the external world (visual, auditory, etc.) or from the internal milieu body (heart activity, respiratory activity, etc); in principle one should distinguish between exteroceptive outcomes $\mathcal{O}_t^e$ and interoceptive ones $\mathcal{O}_t^i$; from a computational modelling standpoint the former are in general observable, whilst the observability of the latter depend on the experimental setting (the suitability/availability of physiological sensing devices, for instance)

Then, the state of the agent at time t is fully specified as

$$\mathcal{S}_t = \left\langle \mathcal{O}_t, \mathcal{I}_t, \mathcal{E}_t, \mathcal{X}_t^e, \mathcal{X}_t^i, \mathcal{A}_t, \mathcal{C}_t \right\rangle$$

We assume that only outcomes from the external world and the body milieu are observable (perceivable), whilst the other ensembles are hidden or latent representations. Thus,

$$\mathcal{S}_t^{hidden} = \left\langle \mathcal{I}_t \mathcal{E}_t \mathcal{X}_t^i, \mathcal{X}_t^e, \mathcal{A}_t, \mathcal{C}_t \right\rangle$$

$$\mathcal{S}_t^{obs} = \langle \mathcal{O}_t \rangle$$

Under such basic setup, the probabilistic model of the agent state (and its dynamics) is captured by the joint distribution $P(\mathcal{S}_{0:T}^{hidden}, \mathcal{S}_{0:T}^{obs})$.

The observable/hidden distinction allows to further refine the Markovian assumption over the state variables through the following conditional independence (CI) assumptions.

The latent state variables evolve in a Markovian way

$$(\mathcal{S}_{t+1}^{hidden} \perp \mathcal{S}_{0:t-1}^{hidden} \mid \mathcal{X}_t^{hidden}). \tag{6.3}$$

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

$$(\mathcal{S}_{t+1}^{obs} \perp \mathcal{S}_{0:t-1}^{hidden}, \mathcal{S}_{t+1:\infty}^{hidden}) \mid \mathcal{S}_t^{hidden}) \tag{6.4}$$

These assumption account for the dynamics of the agent as a state-observation model. In a state-observation model, we view the system as evolving naturally on its own, with observations of it occurring in a separate process (Koller and Friedman, 2009).

Thus, we can write Eq.6.2 as

$$P(\mathcal{S}_{0:T}) = P(\mathcal{S}_0) \prod_{t=0}^{T-1} P(\mathcal{S}_{t+1}^{obs} \mid \mathcal{S}_{t+1}^{hidden}) P(\mathcal{S}_{t+1}^{hidden} \mid \mathcal{S}_t^{hidden}). \tag{6.5}$$

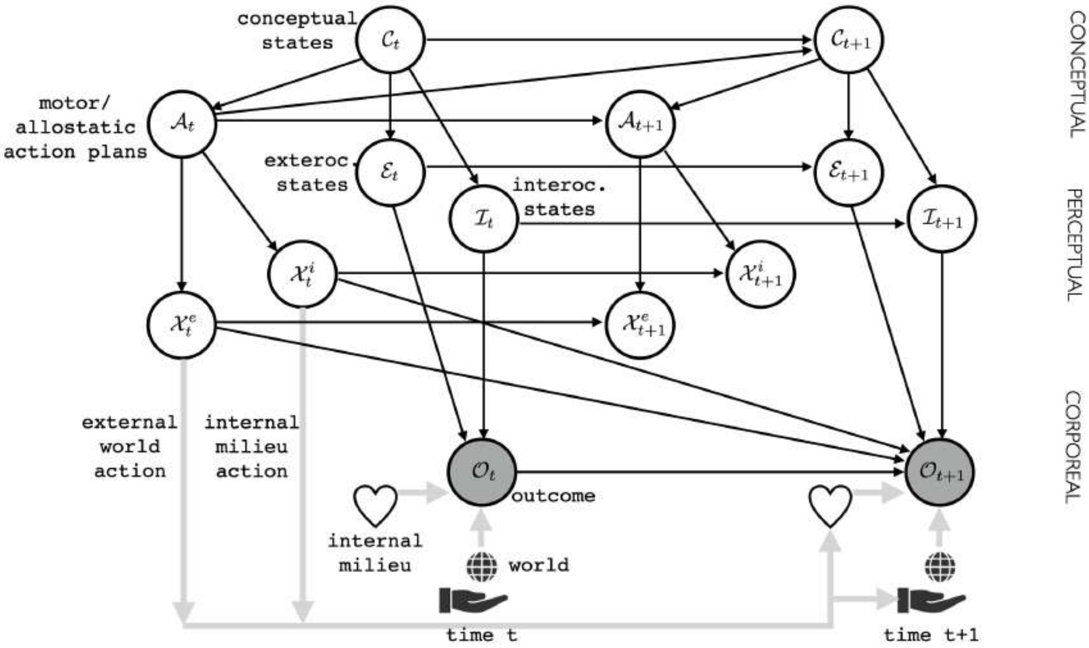


Figure 6.13: *An overview of the theoretical model*

We can further simplify Eq. 6.5 by individuating other CI conditions. This is straightforward by defining a probabilistic graphical model $\mathcal{GM}$ based on the functional/structural constraints given at the psychological and neurobiogical levels. The PGM is presented in Fig. 6.13.

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

Thus, the following factorisation of Eq.6.5 formally holds:

$$
\begin{aligned}
P(\mathcal{S}_{0:T}^{hidden}, \mathcal{S}_{0:T}^{obs}) = P(\mathcal{S}_0) \prod_{t=0}^{T-1} & P(\mathcal{O}_{t+1} \mid \mathcal{O}_t, \mathcal{X}_t^i, \mathcal{X}_t^e, \mathcal{I}_{t+1}, \mathcal{E}_{t+1}) \\
& P(\mathcal{X}_{t+1}^e \mid \mathcal{A}_{t+1}, \mathcal{X}_t^e) \\
& P(\mathcal{X}_{t+1}^i \mid \mathcal{A}_{t+1}, \mathcal{X}_t^i) \\
& P(\mathcal{A}_{t+1} \mid \mathcal{A}_t, \mathcal{C}_{t+1}) \\
& P(\mathcal{E}_{t+1} \mid \mathcal{E}_t, \mathcal{C}_{t+1}) \\
& P(\mathcal{I}_{t+1} \mid \mathcal{I}_t, \mathcal{C}_{t+1}) \\
& P(\mathcal{C}_{t+1} \mid \mathcal{C}_t, \mathcal{A}_t)
\end{aligned}
\tag{6.6}
$$

[3] Denoting X_S the the set of RVs forming the subgraph $S \subseteq \mathcal{GM}$, then we can wite $X_A \perp X_B \mid X_C$ if A is independent of B given C in the graph $\mathcal{GM}$. Let $\mathcal{IM}(\mathcal{GM})$ be the set of all such CI statements encoded by the graph. We say that $\mathcal{GM}$ is an I-map (independence map) for P, or that P is Markov wrt $\mathcal{GM}$, iff $\mathcal{IM}(\mathcal{GM}) \subseteq \mathcal{IM}(P)$, where $\mathcal{IM}(P)$ is the set of all CI statements that hold for distribution P. In other words, the graph is an I-map if it does not make any assertions of CI that are not true of the distribution. This allows us to use the graph as a safe proxy for P when reasoning about P's CI properties.

Equation 6.6 fully specifies, under the simplifying assumptions previously introduced, the inferential processess that define the infrastructure model.

- *Conceptual level.* The distribution $P(\mathcal{C}_{t+1} \mid \mathcal{C}_t, \mathcal{A}_t)$ captures the conceptual ensemble update at time $t+1$, $\mathcal{C}_{t+1}$, given its previous state $\mathcal{C}_t$ and action plan state $\mathcal{A}_t$.

  In turn, the current action plan $\mathcal{A}_{t+1}$, depends on its previous state $\mathcal{A}_t$, and current conceptual state $\mathcal{C}_{t+1}$

- *Perceptual level.* Based on the current conceptual state, $\mathcal{C}_{t+1}$, both interoceptive and exteroceptive predictions are generated, through distributions $P(\mathcal{I}_{t+1} \mid \mathcal{I}_t, \mathcal{C}_{t+1})$ and $P(\mathcal{E}_{t+1} \mid \mathcal{E}_t, \mathcal{C}_{t+1})$, respectively. Such distributions also capture the dependency (dynamics) of current states $\mathcal{I}_{t+1}$ and $\mathcal{E}_{t+1}$ from their previous ones, $\mathcal{I}_t$ and $\mathcal{E}_t$.

  In succession, $\mathcal{I}_{t+1}$ and $\mathcal{E}_{t+1}$ contribute to shape the prediction/inference of likely perceptual outcomes $\mathcal{O}_{t+1}$. The latter is formalized via the distribution

  $$P(\mathcal{O}_{t+1} \mid \mathcal{O}_t, \mathcal{X}^i_t, \mathcal{X}^e_t, \mathcal{I}_{t+1}, \mathcal{E}_{t+1})$$

  Clearly, outcome prediction/generation also is conditioned upon: 1) actual actions previously performed at time t, at the motor/allostatic level, $\mathcal{X}^i_t, \mathcal{X}^e_t$ (for instance, the shifting of gaze, as occurring in visual attention processes, focuses a circumscribed region of the viewed scene on retinal receptors, thus inducing a shift $\mathcal{O}_t \rightarrow \mathcal{O}_{t+1}$, depending on visual RVs, say $\mathcal{V}_t \subset \mathcal{O}_t$, included in the outcome ensemble $\mathcal{O}_t$; 2) the base outcome dynamics $\mathcal{O}_t \rightarrow \mathcal{O}_{t+1}$. The latter, at the most general level, is intended to capture the external world and internal milieu dynamics.

- *Corporeal level.* The current action plan $\mathcal{A}_{t+1}$ is exploited to generate appropriate motor actions $\mathcal{X}^e_t$), via $P(\mathcal{X}^e_{t+1} \mid \mathcal{A}_{t+1}, \mathcal{X}^i_t)$, towards the external world (e.g., by uttering a word, moving a limb, making a gaze shift), or actions intended for the body internal milieu, $\mathcal{X}^i_{t+1}$ (e.g., moderating the heart rate), via $P(\mathcal{X}^i_{t+1} \mid \mathcal{A}_{t+1}, \mathcal{X}^i_t)$ in the service of allostatic regulation.

  Meanwhile, the body perceptual apparatus provide necessary inputs concerning the state of the external world and the state of the internal milieu in order to shape current perceptual outcomes $\mathcal{O}_{t+1}$. For biological agents, this entails the signals collected from either visual, auditory, touch, etc. receptors together with internal body/visceral receptors (e.g., baroceptors, etc.). When an artificial agent is considered, then their number depends on the measurement modalities available (cameras, physiological, motion capture sensors, for example)

As discussed above the Eq. 6.6 and the PGM shown in Figure 6.13. We notice however, that we have so far considered e 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 operationalize 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, but 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 modeling, 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 modeling 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. However, in the Simulations chapter, we will exploit the Pyro PPL to provide concrete proofs of the concepts outlined here.

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 $\mathcal{S}_t$ and related state distribution (prior); current observed outcome $\mathcal{O}_{t+1}$
    and its distribution (evidence)
Output: Agent's state $\mathcal{S}_{t+1}$ and updated state distribution (posterior)
  Conceptual sampling:
  $\mathcal{C}_{t+1} \sim P(\mathcal{C}_{t+1} \mid \mathcal{C}_t, \mathcal{A}_t)$
  $\mathcal{A}_{t+1} \sim P(\mathcal{A}_{t+1} \mid \mathcal{A}_t, \mathcal{C}_{t+1})$                          ▷ action plan sampling
  Perceptual sampling:
  $\mathcal{I}_{t+1} \;\; P(\mathcal{I}_{t+1} \mid \mathcal{I}_t, \mathcal{C}_{t+1})$                          ▷ interoceptive sampling
  $\mathcal{E}_{t+1} \sim P(\mathcal{E}_{t+1} \mid \mathcal{E}_t, \mathcal{C}_{t+1})$                          ▷ exteroceptive sampling
  Corporeal sampling and observation:
  $\mathcal{X}^i_{t+1} \sim P(\mathcal{X}^i_{t+1} \mid \mathcal{A}_{t+1}, \mathcal{X}^i_t)$                     ▷ internal motor sampling
  $\mathcal{X}^e_{t+1} \sim P(\mathcal{X}^e_{t+1} \sim \mathcal{A}_{t+1}, \mathcal{X}^e_t)$                     ▷ external motor sampling
  OBSERVE$(P(\mathcal{O}_{t+1} \mid \mathcal{O}_t, \mathcal{X}^i_t, \mathcal{X}^e_t, \mathcal{I}_{t+1}, \mathcal{E}_{t+1}) : \mathcal{O}_{t+1})$   ▷ sensing
```

In brief we can consider such model specification as computing an approximate posterior distribution over latent state ensembles. In particular, the last step relies on the observed outcome $\mathcal{O}_{t+1}$, which in turn depends on the states of the latent ensembles $\mathcal{S}^{hidden}_{t+1}$. More precisely, since the one-step procedure returns $\mathcal{S}^{hidden}_{t+1}$, the result of this "program" is the posterior marginal distribution over $\mathcal{S}^{hidden}_{t+1}$.

Two cogent issues are worth a remark.

- First, a model of an agent defined as above, might even operate as "detached" from the world by purely relying, at the observation stage, on the predictive dynamics $\mathcal{O}_t \rightarrow \mathcal{O}_{t+1}$. This is important, for instance, when accounting for simulation-based, embodied theories of affect (thus, CAT), grounding in "as if" simulation ability (e.g., Boccignone et al., 2018b; Horii et al., 2016).

- Second, since PP can explicitly represent conditioning as part of a model, it enables us to describe reasoning about others' reasoning using nested conditioning. Much of human reasoning is about the beliefs, desires, and intentions of other people: PP can be used to formalize these inferences in a way that captures the flexibility and inherent uncertainty of reasoning about other agents. If reasoning can be viewed as probabilistic inference, then reasoning about others' reasoning boils down to inference about inference; however, if inference is not itself represented as a probabilistic model we cannot formulate inference about inference in probabilistic termsStuhlmüller and Goodman, 2014. In brief, as Stuhlmüller and Goodman (2014) have argued, PP offers a powerful opportunity for modeling theory of mind, and, consequently, the very nature of the communication act based on recursive reasoning, the RSA theory being one notable example.

To make a step further in this direction we need to detail to some extent the agent's architecture outlined in Fig. 6.13, markedly at the conceptual level. At the same time we will embrace some simplifying assumption for parts of the model that are less relevant for our research problem: modelling the wiring between words and emotions. For instance, since here and in the case studies presented in the next Chapter, we are not specifically dealing with an agent moving in the real world, grasping objects, etc., (which would be important for modelling concept learning in robot/human interaction), actual motor behaviour will be simplified to the point. In the same way, details of how allostatic physiological control is performed will be overlooked.

Specifying the conceptual level To expand on the conceptual level, we briefly introduce some definitions usually exploited in classic AI (Russell and Norvig, 2022), substantially in the BDI (Belief-Desire-Intention) modelling research field (Sánchez-López and Cerezo, 2019; Reisenzein et al., 2013).

At the conceptual level we can broadly consider three kinds of mental states: beliefs, desires and intentions, which represent the informational, motivational and deliberative states of an agent, respectively. More precisely:

- *Beliefs:* generally represent environment/context characteristics, which are updated accordingly after the perception of each action. They can be seen as the informative component of the system.

  In our case, belief states are RVs $B_t \in \mathcal{B}_t$ encoding possibly related concept states (which might be simply represented through a list of concepts, up to a more complex representational structure such as those involving Relational Probability Models, Russell and Norvig, 2022). A belief state is usually defined as a representation of the set of possible states of the world (in our case, both external and internal worlds). Thus, (degrees of) beliefs can be accounted for by probability distributions over such states, (Russell and Norvig, 2022)).

- *Desires:* are generally defined as the motivational state of the system (Russell and Norvig, 2022). They have information about the objectives to be accomplished, i.e., what priorities or payoffs are associated with the various current objectives. Desires represent situations that an agent wants to achieve. The fact that an agent has a desire does not mean that the agent will satisfy it. The agent carries out a deliberative process in which the agent confronts its desires and beliefs and chooses a set of desires that can be satisfied.

  In the case of pragmatic communication, the intention can be identified as that of conveying meaning to another agent.

  - *Goals:* A goal, denoted via the RV $G_t \in \mathcal{G}_t$ is a desire that the agent chooses for active pursuit. Goals should be consistent. Desires can be contradictory to each other, but the goals cannot. Thus, an agent can desire P and $\neg P$ at the same time, but only one of the two can become a goal.

- *Intentions:* is a goal that is chosen to be executed by a plan, namely the current action plan state $A_t \in \mathcal{A}_t$. They capture the deliberative component of the system. This choice is made because the agent believes it can satisfy the goal (it is not rational for an agent to carry out something that it does not believe it can achieve). Plans are procedures that depend on a set of conditions for being applicable. Intentions are persistent and represent the currently chosen course of action. An agent will not give up on its intentions - they will persist, until the agent believes it has successfully achieved them, it believes it cannot achieve them or the purpose of the intention is no longer present.

  In the case of a communication act an action/intention can boil down to the actual planning suitable to support the conveyance of meaning by uttering a word or a sentence or providing a non behavioral signal, such as a facial expression, a body posture or a gesture.

For the purpose of modelling concepts and pragmatic communication, in addition, we need to embed language within the conceptual level. To such end, first, a lexicon $\mathcal{L}$ should be defined, namely a list of allowable words. An associated dictionary can provide a suitable prior on word semantics.

Including words at the conceptual level is mandatory in our case, not only for communication purposes, but also to defining categories a key, constitutive notion in CAT, as we have largely discussed. Any concept itself is a mental representation of a category that is labelled through a word. Emotion categories themselves have no particular citizenship in this respect, and an emotional word, e.g. “happy” is constitutive of that emotion.

Accordingly, following Atzil et al. (2018), a concept state C can be defined as $C \in \mathcal{L} \times \mathcal{E} \times \mathcal{I}$, that is a state within the joint state-space of lexicon, exteroceptive and interoceptive states. A concept is thus a probability distribution over a concept state $C = c$, e.g. $P(w, e, i)$, with $e \in \mathcal{E}, i \in \mathcal{I}$ Such definition is the baseline. At a higher level the concept could be defined recursively in terms of another concept. A concrete example of interest is when generalizing the interoceptive sensation to an abstract categorical emotion, say happiness, indexed for short by the corresponding emotion word $P(w, e, \text{"}\texttt{happy}\text{"})$.

Second, we take into account a language model $\mathcal{LM}$, $P(w_{t+1} \mid w_{t<t+1}, \mathcal{LM})$, which can be generically defined as a probability distribution of any string (characters, words, sentences). Most often, an $\mathcal{LM}$ is used to predict the next word in text given the previous words and is often used as a building block for more complex tasks.

In principle, handling language calls for grammars and parsing. However, from a more practical stance, current implementation models available for $\mathcal{LM}$ that embed words in contextual representations (e.g., of a word in a sentence) are likely to implicitly learn latent representations that capture the same basic ideas as grammars and even shallow semantics representations (cfr. Section 4.1.4).

Eventually, a state at the conceptual level can be defined as the ensemble

$$(\mathcal{B}_t, \mathcal{C}_t, \mathcal{G}_t, \mathcal{A}_t, \mathcal{LM}_t, \mathcal{L}_t)$$

(for generality, we have time indexed the lexicon $\mathcal{L}_t$ to encompass cases of dynamical lexicon learning or restriction)

Specifying the perceptual level The perceptual level basically relies on exteroceptive and interoceptive ensembles, $\mathcal{E}_t$ and $\mathcal{I}_t$, respectively.

We denote RVs $Y^e_{n,t}, n = 1, 2, \cdots$ the unimodal sensing variables or features (e.g., vision), while $Y^i_{m,t}, m = 1, 2, \cdots$ are unimodal physiological variables (e.g heart rate, skin conductance).

At a higher level Z^e_t and Z^i_t stand for the multimodal exteroceptive and interoceptive representations, respectively, that integrate unimodal features.

While not strictly necessary, but to keep with classical CAT theory of affect (and other emotion theories), we introduce as a summary of the overall physiological state of the body the core affect state space $F_t = V \times A$. This is described in terms of the continuous stochastic variables of valence V_t and arousal A_t. Core affect is a suitable tool, not only for summarisation purposes, but also to control the "granularity" of the perceived affect state.

Eventually, the exteroceptive and exteroceptive ensembles are defined as:

$$\mathcal{E}_t = (\{Y^e_{n,t}\}, Z^e_t)$$

$$\mathcal{I}_t = (\{Y^i_{n,t}\}, Z^i_t, F_t)$$

Specifying the corporeal level $\mathcal{O}_t$ represent the outcome signals from the body. In the case of exteroception these will be a set of stochastic signals from external perceptual and somatosensory systems, $\{O^e_{p,t}\}, p = 1, 2\cdots$ (e.g., visual, auditory, touch, etc.). Signals from the body internal milieu will be denoted $\{O^i_{q,t}\}, q = 1, 2\cdots$

Thus,

$$\mathcal{O}_t = (\{O^e_{p,t}\}, \{O^i_{q,t}\})$$

At the corporeal level, we also consider RVs $X^e_t = \{X^e_{r,t}\}$, represent a set of externally executed actions directed towards the world (limb movements, eye movements, gestures, postures expressions); as we have seen, their execution can change the perception of the world at time $t + 1$.

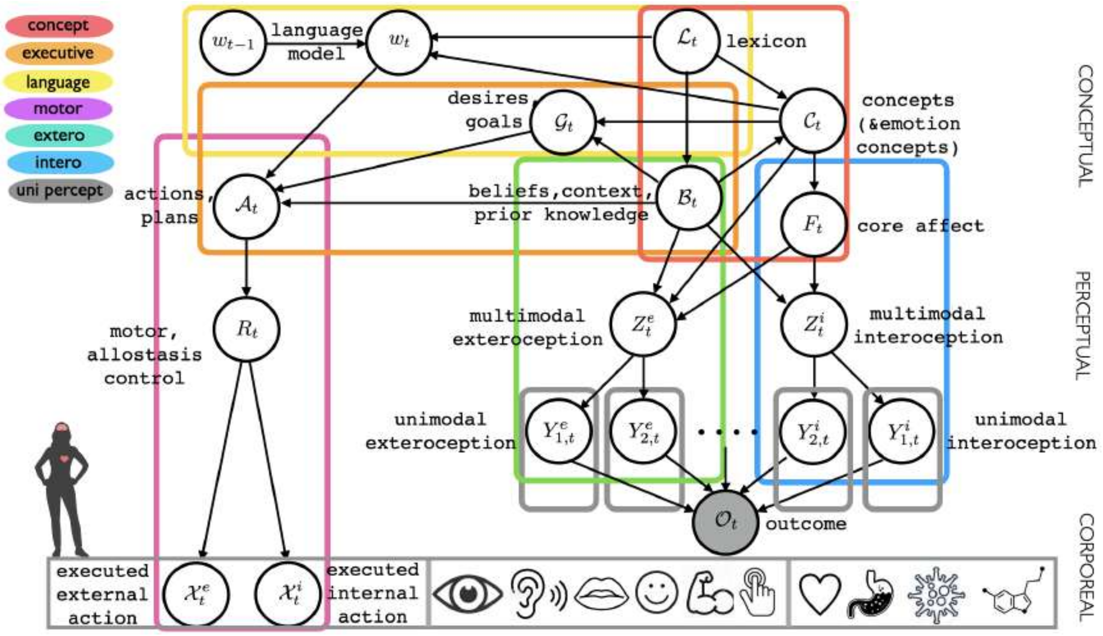


Figure 6.14: *The infrastructure PGM of the agent-in-context. The PGM further details the structure of the theoretical model of the agent presented in Fig. 6.13. For sake of clarity only the $\mathcal{GM}$ slice at time t is shown, the only exception being the generation of a word sequence according to some language model. Coloured boxes functionally map the large-scale network processes that were captured in the scheme of the psychological infrastructure presented in Fig. 6.11 as psychological "primitives"*

RVs $X_t^i = \{X_{s,t}^e\}$ are a set of internally executed actions, such as those allowing allostasis, and they will not be further considered, though interesting work is currently addressing this level in robotics (e.g., Khan and Cañamero, 2018).

For completeness sake, one should take into account an intermediate level between a planned action A_t and its terminal realization X_t. Specially when dealing with artificial agents, robots in particular, a motor program or sequence (motor parameters) R_t should be devised to implement the plan and guide X_t. One elegant example has been provided by Metta et al. (2006) for what concerns the probabilistic modelling of mirror neurons. A similar approach was exploited by Boccignone et al. (2018b) for simulation-based (mimicry) affect enactment.

For instance, when the agent has the intention to utter a word, following the conceptualization stage (in which the intention to create speech links a desired concept to the particular spoken words to be expressed) an action plan is setup. This, in the simplest case is represented by a string of characters. Accordingly, appropriate phonological information is enacted by interacting with the core language network (actually,this formulation stage should involve grammatical encoding, morpho-phonological encoding, and phonetic encoding); subsequently, supplementary motor areas in the premotor cortex are activated to coordinate the appropriate motor plan. The latter further involves motor areas, that trigger subcortical regions (e.g, the cerebellum) for the control of articulation, which is the execution of the articulatory score by the lungs, glottis, larynx, tongue, lips, jaw and other parts of the vocal apparatus resulting in the final vocalization of the word (sound wave) in the external world (Tatham and Morton, 2006). A similar scheme is involved in writing a word, but, obviously, the final motor coordination involves either limb/hand movements and oculomotor shifts (Coen-Cagli et al., 2009; Cagli et al., 2008).

We can thus assume that the ensemble $(\mathcal{A}_t, \mathcal{R}_t, \mathcal{X}_t)$ by and large abstractly accounts for the motor and allostatic regulation systems.

The PGM related to the specifications introduced above is outlined in Fig. 6.14. Conditional independencies are derived drawing from the base PGM presented in Fig. 6.11. Under such conditions the general PPL model (Algorithm 1) can be written as in Algorithm 2.

Modelling the communication act Having defined the conceptual act of the agent, now we need to formally define the communication act. To such end we resort to the RSA framework (Goodman and Frank, 2016) that we have briefly introduced in Section 4.2.5.

In the discussion that follows, for the sake of clarity, we will omit time dependencies $t \to t+1$, considering only the situation at time t. This will suffice to set up the communication component of the model grounded in the conceptual act component.

RSA deals with a pragmatic communication occurring between two agents, a speaker S and a listener L. The task of the listener L is to estimate the probability of a particular intended message m given the heard, observed utterance u by the speaker.

Assume that at time t, given the outcome $\mathcal{O}$ and prior beliefs concerning the current context, an ensemble of inferred beliefs and supporting concepts $\mathcal{B}, \mathcal{C}$

The latter according to the model presented in Algorithm 2 can involve lexicon $\mathcal{L}$ and a language model $\mathcal{LM}$ to be established as summarized by the following sampling

Algorithm 2 Full simulation-based one-step dynamics

```
Input: Agent's state $\mathcal{S}_t$ and related state distribution (prior); current observed outcome $\mathcal{O}_{t+1}$
    and its distribution (evidence), a lexicon $\mathcal{L}$, a language model parametrized by $\mathcal{LM}$
Output: Agent's state $\mathcal{S}_{t+1}$ and updated state distribution (conditional posterior)
    // Conceptual sampling:
    $\mathcal{B}_{t+1} \sim P(\mathcal{B}_{t+1} \mid \mathcal{B}_t, \mathcal{A}_t)$    ▷ belief update sampling
    $\mathcal{L}_{t+1} \sim P(\mathcal{L}_{t+1} \mid \mathcal{B}_{t+1}, \mathcal{L}_t)$    ▷ current lexicon sampling
    $\mathcal{C}_{t+1} \sim P(\mathcal{C}_{t+1} \mid \mathcal{C}_t, \mathcal{B}_{t+1}, \mathcal{L}_{t+1})$    ▷ concept sampling
    $w_{t+1} \sim P(w_{t+1} \mid w_{t+1}, \mathcal{L}_{t+1}, \mathcal{C}_{t+1}, \mathcal{LM})$    ▷ word/sentence sampling
    $\mathcal{G}_{t+1} \sim P(\mathcal{G}_{t+1} \mid \mathcal{C}_{t+1}, \mathcal{B}_{t+1})$    ▷ goal sampling
    $\mathcal{A}_{t+1} \sim P(\mathcal{A}_{t+1} \mid \mathcal{A}_t, w_{t+1}, \mathcal{G}_{t+1}, \mathcal{B}_{t+1})$    ▷ action plan sampling
    // Perceptual sampling:
    $F_{t+1} \sim P(F_{t+1} \mid F_t, \mathcal{C}_{t+1})$    ▷ core affect sampling
    $Z^i_{t+1} \sim P(Z^i_{t+1} \mid Z^i_t, F_{t+1}, \mathcal{B}_{t+1})$    ▷ multimod. interoceptive
    $Y^i_{n,t+1} \sim P(Y^i_{n,t+1} \mid Y^i_{n,t}, Z^i_t, F_{t+1})\ n = 1, 2, \cdots$    ▷ unimod. interoceptive
    $Z^e_{t+1} \sim P(Z^e_{t+1} \mid Z^e_t, F_{t+1}, \mathcal{B}_{t+1})$    ▷ multimod. exteroceptive
    $Y^e_{m,t+1} \sim P(Y^e_{m,t+1} \mid Y^e_{m,t}, R^e_{m,t}, Z^e_t, F_{t+1})\ m = 1, 2, \cdots$    ▷ unimod. exteroceptive
    // Corporeal sampling and observation:
    $R^i_{t+1} \sim P(R^i_{t+1} \mid \mathcal{A}_{t+1}, R^i_t)$    ▷ internal motor sampling
    $R^e_{t+1} \sim P(R^e_{t+1} \mid \mathcal{A}_{t+1}, R^e_t)$    ▷ external motor sampling
    $X^i_{t+1} \sim P(\mathcal{X}^i_{t+1} \mid R^i_{t+1}, X^i_t)$    ▷ internal motor exec.
    $X^e_{t+1} \sim P(\mathcal{X}^e_{t+1} \mid R^e_{t+1}, X^e_t)$    ▷ external motor exec.
    OBSERVE$(P(\mathcal{O}_{t+1} \mid \mathcal{O}_t, X^i_t, X^e_t, Y^i_{t+1}, Y^e_{t+1}) \leftarrow \mathcal{O}_{t+1})$    ▷ internal/external sensing
```

steps:

$$\mathcal{B}_{t+1} \sim P(\mathcal{B}_{t+1} \mid \mathcal{B}_t, \mathcal{A}_t) \tag{6.7}$$

$$\mathcal{L}_{t+1} \sim P(\mathcal{L}_{t+1} \mid \mathcal{B}_{t+1}, \mathcal{L}_t) \tag{6.8}$$

$$\mathcal{C}_{t+1} \sim P(\mathcal{C}_{t+1} \mid \mathcal{C}_t, \mathcal{B}_{t+1}, \mathcal{L}_{t+1}) \tag{6.9}$$

$$w_{t+1} \sim P(w_{t+1} \mid w_{t+1}, \mathcal{L}_{t+1}, \mathcal{C}_{t+1}, \mathcal{LM}) \tag{6.10}$$

More precisely, at the conceptual level the agent has available a state-space of concepts that in its simplest form can be denoted $\mathcal{L} \times \mathcal{Z}^e \times \mathcal{F}$, in simple terms a concept is represented by some multimodal exteroceptive representation together with affective value and a word from available lexicon $\mathcal{L}$ indexing a category. $\mathcal{B}_t$ be the current set of beliefs over the conceptualized states of the world the agent senses through available observations $\mathcal{O}_t$, thus defining a state-space $\mathcal{S}$. Define the overall state space $\mathcal{M} = \mathcal{S} \times \mathcal{L} \times \mathcal{Z}^e \times \mathcal{F}$

Then, the speaker's desire of conveying a meaning m represents the speaker's current goal $g_t \in \mathcal{G}_t$, which can be defined as the projection

$$g_t : \mathcal{M} \to M_X \tag{6.11}$$

where $M_X \subseteq \mathcal{M}$ is the meaningful subspace relevant for the speaker. Thus, $m_t \in M_X$.

This contextualizes goal sampling

$$\mathcal{G}_{t+1} \sim P(\mathcal{G}_{t+1} \mid \mathcal{C}_{t+1}, \mathcal{B}_{t+1}). \tag{6.12}$$

Uttering a word/sentence is supported by the simulation model via the sampling cascade

$$\mathcal{A}^e_{t+1} \sim P(\mathcal{A}^e_{t+1} \mid \mathcal{A}_t, w_{t+1}, \mathcal{G}_{t+1}, \mathcal{B}_{t+1}) \tag{6.13}$$

$$R^e_{t+1} \sim P(R^e_{t+1} \mid \mathcal{A}_{t+1}, R^e_t) \tag{6.14}$$

$$X^e_{t+1} \sim P(\mathcal{X}^e_{t+1} \mid R^e_{t+1}, X^e_t) \tag{6.15}$$

In a straight information-based communication act, the speaker's goal pursues restricts to the communication of the (semantic) literal meaning m, where $M_X = \mathcal{L}$, via the utterance $u_t \in \mathcal{UT}$ of the word $w_t \in \mathcal{L}$ (or a sentence in the most complex case, but which in principle can be conceived as a sequence $w_t = w_{\tau_1}, w_{\tau_2}, \cdots, w_{\tau_r}$, generated by some language model $\mathcal{LM}$). The uttering action $a^u_t \in \mathcal{UT} \subset \mathcal{A}_t$ boils down to the mapping,

$$a^u_t : \mathcal{L} \to \mathcal{UT}, \tag{6.16}$$

thus, $u_t = a^u_t(w_t)$. Note that according to the model, such generative mapping entails all the previously described motor planning steps necessary to operationalize $w_t \to u_t$, to generate the actual vocalization u_t. Clearly, when the action is a full sentence then $\mathcal{UT} \subset \mathcal{L} \times \mathcal{LM}$

Eventually, this will be sensed (heard) by the listener as the outcome $O(u_t) \in \mathcal{O}_t$. Assume in the following, for notational simplicity and without loss of generality, a perfect sensing $u_t = O(u_t)$.

However, in the case of a pragmatic communication act the intended meaning of the speaker goes beyond the literal meaning (the truth value according to some established semantics) of a word or sentence. In the classic RSA framework the joint communication between the listener and the speaker is accounted for via the joint distribution over meaning and utterance $P(m_t, u_t \mid g_t, \mathcal{C}_t)$ given the current goal and conceptual state, g_t and $\mathcal{C}_t$, respectively.

The speaker's perspective can be generally represented through the generative factorization

$$P(m_t, u_t \mid g_t, \mathcal{C}_t) = P_S(u_t, \mid m_t, \mathcal{C}_t) P(m_t \mid g_t, \mathcal{C}_t) \tag{6.17}$$

where $\mathcal{C}_t$ is, as defined before (cfr. Algorithm 1), a representation of the general conceptual level, in this case of the speaker.

In other terms, speaker's communication act is that of uttering u_t conditionally on goals (intended meaning m) and conceptual/belief states established on the rationally observed state of the world, which amounts to the *forward* step $m_t \to u_t$,

$$m_t \sim P(m_t \mid g_t, \mathcal{C}_t) \tag{6.18}$$

$$u_t \sim P_S(u_t, \mid m_t, \mathcal{C}_t) \tag{6.19}$$

In the baseline RSA, the speaker S chooses the utterance by maximizing his own utility $U_S(u; m)$:

$$P_S(u_t, \mid m_t, \mathcal{C}_t) \propto \exp \alpha U_S(u; m) \tag{6.20}$$

The utility of an utterance, in turn, depends on how much epistemic certainty it provides to the listener:

$$U_S(u; m) = \log P_L(m_t \mid u_t, \mathcal{C}_t) \tag{6.21}$$

The scalar value α can be interpreted as an indicator of how rational the speaker is in choosing utterances (i.e., how strongly speakers prefer the higher utility option). The speakers' 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.

On the other side, the listener's perspective is an inferential one; namely, to recover the meaning from the utterance. This can be written as the *backward* step $m_t \leftarrow u_t$, so that the joint probabilities of meaning and utterance factorize as

$$P(m_t, u_t \mid g_t, \mathcal{C}_t) = P_L(m_t \mid u_t, g_t, \mathcal{C}_t) P(u_t \mid g_t, \mathcal{C}_t) \tag{6.22}$$

Eqs. 6.17 and 6.22 can be used to join the two perspectives

$$P_L(m_t \mid u_t, g_t, \mathcal{C}_t) P(u_t \mid g_t, \mathcal{C}_t) = P_S(u_t, \mid m_t, \mathcal{C}_t) P(m_t \mid g_t, \mathcal{C}_t),$$

in order to allow the listener to capture the most likely meaning:

$$P_L(m_t \mid u_t, g_t, \mathcal{C}_t) = \frac{P_S(u_t \mid m_t, \mathcal{C}_t) P(m_t \mid g_t, \mathcal{C}_t)}{P(u_t \mid g_t, \mathcal{C}_t)}. \tag{6.23}$$

A first remark on Eqs. 6.23 concerns the fact that such inferential activity sets the listener as a Bayesian listener whose inference on meaning is performed through Bayes rule (Eq. 6.23), briefly $P_L(m_t \mid u_t, g_t, \mathcal{C}_t) \approx P_S(u_t \mid m_t, \mathcal{C}_t) P(m_t \mid g_t, \mathcal{C}_t)$.

Yet, the listener is an agent in context, whose forward model is set by Algorithm 1. Thus, listener's inference entailed by Eq. 6.23 is achieved, according to the PPL model, by listener's sampling through the forward step defined in Eq. 1 under common ground (priors). This means that the listener unfolds her action plan by adopting the sampling step $u_t \sim P_S(u_t \mid m_t, \mathcal{C}_t)$. In other words, the listener takes the intentional stance of putting herself in the speaker's shoes: the listener commits herself to consider how the speaker generates, via P_S the heard utterance u_t. As a result, in a nutshell, the generative story of the listener's L reasoning relies on the *simulation* of the speaker S.

A second remark is that the unfolding of the communication act follows a recursive structure. In fact, by inspecting Eqs. 6.20 and 6.21, it is apparent that the speaker utters u_t having in mind a model of the listener. In the RSA framework, to avoid infinite recursion, the base is provided by the so called literal listener Lit (or L_0) who interprets utterances in accordance with their literal semantics

$$P_{Lit}(m_t \mid u_t, \mathcal{C}_t) \propto \delta_{[[u_t]](m_t)} P(m_t \mid \mathcal{C}_t), \tag{6.24}$$

Here, $[[u]]$ is a semantic denotation for each sentence, concerning whether or not the utterance is true of a given message. $P(m_t \mid \mathcal{C}_t)$ 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.

It is through this recursive reference back to a listener, that the model captures the interdependence of speaker and listener in communicative interactions. The combination of these two terms — speaker likelihood and prior — to form the listener's belief represents the outcome of a social-cognitive inference about the likely intended meaning of an utterance in context.

This for what concerns the baseline RSA agents.

Cogently, our speaker-in-context, beyond the intentional utterance u_t intended to convey meaning m_t is likely to generate other executed external actions (beyond the internal allostatic ones), either intentional or unintentional, conditionally on goals and conceptual/belief states that ground in the currently observed state of the world, but perceived through the "filtering" of interoceptive/exteroceptive states.

This means that, beyond the uttering action, can rely on Eqs. 6.37,6.38 and 6.39, for the sampling of other external actions.

For example, the speaker might want to communicate an affective feeling jointly with the word, which might or might not be congruous with the literal meaning. Further, even if speaker's intention is that to convey bare literal meaning, his nonverbal behaviour, which is observed by the listener, provides a "context" to frame the literal meaning.

Actually, together with the observed speaker's utterance u_t and the environmental outcomes $\mathcal{O}_t^{env}$ (e.g., defining the context of the scene where the communication act unfolds, which is in common with the speaker), also considered the classic RSA framework, the listener has available at least one other outcome from the sensed world, the speaker non verbal behavior $\mathcal{O}_t^{S_{NV}}$ (e.g., facial expression, gesture, posture, prosody).

Assume that, to keep simple the discussion, together with the uttering action $u_t = a_t^u(w_t)$, the speaker performs a facial expression action $x_t^f = a_t^{NV}(\mathcal{C}_t)$ as a result of speaker's overall conceptual state $\mathcal{C}_t$. Here, we might distinguish between two cases that are relevant from the standpoint of emotion theories (Crivelli and Fridlund, 2019; Fridlund and Russell, 2021).

1. The facial expression action x_t^f is the outcome of an intended, "voluntary" social signalling of the speaker. In this case, x_t^f is *au pair* with u_t. It is a nonverbal cue constrained by the meaning m the speaker wants to convey together with the verbal cue (for instance, to enforce or to bias its literal meaning). Thus, speaker's forward sampling is, in principle, a bare multi-cue extension of Eq. :

$$m_t \sim P(m_t \mid g_t, \mathcal{C}_t) \tag{6.25}$$

$$u_t, x_t^f \sim P_S(u_t, x_t^f \mid m_t, \mathcal{C}_t) \tag{6.26}$$

2. The facial expression action x_t^f is the unintended, albeit to some extent compulsory outcome of speaker's current conceptualization, which is in turn constructed, as seen before, also based on the speaker's affective/interoceptive state (in the extreme case of BET, x_t^f would be recognised as a facial expression of emotion). We can formally express this case by considering two options. The first is, by referring to Fig. 6.14 is to exploit the direct conditioning from (a subset of) the current belief states $\mathcal{B}_t$, that might link memories of past contextual situation bearing a high interoceptive/affective value, and available actions $\mathcal{A}_t$ leading to a higher probability of sampling an "affective " external action $x_t^f = a_t^{NV}(\mathcal{B}_t)$, thus by-passing, in this case, the goal sampling stage.

   The other option is to adopt a finer distinction between kinds of goals. In the psychological/neurobiological literature it is often the case that a general distinction is made between exogenous (originating from outside the observer's organism, e.g., the instruction to perform a task) and endogenous (internal) goals. However, a finer distinction can be made according to the different types of reward that is expected in pursuing a certain goal as discussed in-depth by Berridge and Robinson (2003). High level mental states such as "cognition" "motivation", and "emotion" involve explicit and implicit psychological component processes. Explicit processes are consciously experienced (e.g. explicit desire, expectation or pleasure), whereas implicit psychological processes are unconscious in the sense that they can operate at a level not always directly accessible to conscious experience. Examples of the latter are implicit incentive salience (unconscious "wanting"), habits, and implicite "liking" reactions. For instance, one might distinguish between a conscious pleasure (liking), tied to explicit hedonic feelings, and core hedonic "liking" involving affective reactions, e.g., a non voluntary facial expression, and implicit affect (Berridge and Robinson, 2003). Additional psychological and neural processes of cognitive awareness can sometimes transform the products of implicit processes into explicit representation, but explicit awareness is not necessary for implicit processes (and beliefs) to powerfully influence behavior (Berridge and Robinson, 2003). Achieving reward, either explicit or implicit, is at the basis of allostatic regulation of the body budget, and reward is in generally achieved by setting some goal. Under such circumstances, we might generally distinguish between explicit and implicit goals $\mathcal{G} = (\mathcal{G}^{exp}, \mathcal{G}^{imp})$. Then

$$x_t^f \sim P_S(x_t^f \mid g_t^{imp}, a_t^{NV}, \mathcal{C}_t).$$

Further, one might associate a biological, implicit functional meaning m_t^{imp} to the executed action. Eventually, the speaker's sampling writes:

$$m_t^{exp} \sim P(m_t^{exp} \mid g_t^{exp}, \mathcal{C}_t) \tag{6.27}$$

$$m_t^{imp} \sim P(m_t^{imp} \mid g_t^{imp}, \mathcal{C}_t) \tag{6.28}$$

$$u_t \sim P_S(u_t \mid m_t^{exp}, \mathcal{C}_t) \tag{6.29}$$

$$x_t^f \sim P_S(x_t^f \mid m_t^{imp}, \mathcal{C}_t). \tag{6.30}$$

What is interesting here is that the listener has now available two sources of outcome u_t and x_t^f (again under the simplifying notation $x_t^f = \mathcal{O}_t^{NV}(x_t^f)$), beyond the environmental outcomes $\mathcal{O}_t^{env}$ to infer meaning m

$$P_L(m_t \mid u_t, x_t^f, g_t, \mathcal{C}_t) = \frac{P_S(u_t, x_t^f \mid m_t, \mathcal{C}_t) P(m_t \mid g_t, \mathcal{C}_t)}{P(u_t, x_t^f \mid g_t, \mathcal{C}_t)}, \tag{6.31}$$

Clearly, the speaker might have chosen the bare informative option m_t as in case 1), i.e., sampling via Eq. 6.26, or acted as in case 2), i.e., via 6.30, where m_t represents the pair (m_t^{exp}, m_t^{imp}). In the most natural settings, the listener is likely to be confronted with cases in which, for instance, the speaker had the intention to act according to Eq. 6.26 by displaying a constructed facial expression (e.g., a non Duchenne smile, adopting BET's jargon, or simply a fake smile), while other sensed non verbal actions (prosody or postures) appear to contrast the speaker's intended meaning.

Clearly, Eq.6.31 provides a simplified view of the complete conceptual act underlying such inferential step. As represented in Algorithm 2, such inference is actually to be performed through the predictive sampling, based on current beliefs and concept, of core affect together with exteroceptive and interoceptive states down to the perceived outcomes $\mathcal{O}_{t+1}$

$$F_{t+1} \sim P(F_{t+1} \mid F_t, \mathcal{C}_{t+1}) \tag{6.32}$$

$$Z^i_{t+1} \sim P(Z^i_{t+1} \mid Z^i_t, F_{t+1}, \mathcal{B}_{t+1}) \tag{6.33}$$

$$Y^i_{n,t+1} \sim P(Y^i_{n,t+1} \mid Y^i_{n,t}, Z^i_t, F_{t+1}) \; n = 1, 2, \cdots \tag{6.34}$$

$$Z^e_{t+1} \sim P(Z^e_{t+1} \mid Z^e_t, F_{t+1}, \mathcal{B}_{t+1}) \tag{6.35}$$

$$Y^e_{m,t+1} \sim P(Y^e_{m,t+1} \mid Y^e_{m,t}, R^e_{m,t}, Z^e_t, F_{t+1}) \; m = 1, 2, \cdots \tag{6.36}$$

A subtle but cogent point is that the listener actually has as available as actually observable outcomes the environmental outcomes $\mathcal{O}^{env}_t$ and the speaker non verbal behavior outcome $\mathcal{O}^{S_{NV}}_t$. Thus, in principle she should form her beliefs and concepts based only on the backward inference, say $P(\mathcal{B}_t, \mathcal{C}_t \mid O^{env}_t, \mathcal{O}^{S_{NV}}_t)$. However, the simulation loop provides her with the ability of generating "internal" action aiming at regulating physiological behavior,

$$\mathcal{A}^i_{t+1} \sim P(\mathcal{A}^i_{t+1} \mid \mathcal{A}_t, w_{t+1}, \mathcal{G}_{t+1}, \mathcal{B}_{t+1}) \tag{6.37}$$

$$R^i_{t+1} \sim P(R^i_{t+1} \mid \mathcal{A}^i_{t+1}, R^i_t) \tag{6.38}$$

$$X^i_{t+1} \sim P(\mathcal{X}^i_{t+1} \mid R^i_{t+1}, X^i_t). \tag{6.39}$$

Thus, when inferring belief and concepts based on the environmental $\mathcal{O}^{env}_t$ and the observed speaker's non verbal behavior outcome $\mathcal{O}^{S_{NV}}_t$, the listener can rely on her internally perceived physiological outcomes $\mathcal{O}(X^i_{t+1})$ "as if" they were generated by the speaker (who indeed shares the same infrastructure), in order to "fine tune" her own inference.

Indeed, this "as if" mode of operating is a key concept in the simulation-based, embodied approaches to affect understanding (e.g., for grounding empathy) and more generally to shape a theory of the mind of others. This crucial capability will be addressed in the simulations Chapter.

Eventually, note that other and more subtle variations of the behaviour of the model can be achieved by designing appropriate utility functions (Goodman and Frank, 2016). Indeed the notion of the speaker's utility (what is rewarding for a speaker) is central to the RSA approach. The basic RSA model captures the speaker's need to be informative to a listener (Eq. 6.21). Different utilities lead to different kinds of speaker, which in turn lead to different interpretations by the pragmatic listener. Several utility refinements and their combinations have been considered.

Utterance cost can be introduced to capture a tendency of speakers to be parsimonious we can simply add a cost term:

$$U_S(u; m) = \log P_L(m_t \mid u_t, \mathcal{C}_t) + cost(u) \tag{6.40}$$

The cost may reflect actual production cost (such as number of words) or proxies, such as word frequency. and is related to Grice's maxim of manner.

When the speaker does not have full knowledge of the world he should choose an utterance according to the expected utility:

$$U_S(u;k) = E_{P(m|k)}\left[U_S(u;m)\right] \tag{6.41}$$

where k summarizes the speaker's knowledge or observations.

Other social goals can be modulated by designing specific utility functions to take into account non-informational utilities, such as utility directed toward kindness, politeness, etc., (Goodman and Frank, 2016). This approach is consistent with the reward based assumptions discussed in case 2) above.

Clearly this is an abstract model which might be operationalized through the adoption of different implementation models even under suitable approximations.

6.6 Theoretical analysis of the model

Our model has been so far designed and discussed in terms of the structural/functional constraints derived at the neurobiological/psychological levels of explanation. In what follows, we investigate the theoretical implications of its structure to provide further insights. In particular its connections with the predictive processing hypothesis and the theory of stochastic processes will be taken into account.

Cogently, we made the first move by focusing on the novel perspective of the predictive brain as the most suitable to account for the overall framework on language and emotions. As a matter of fact, the neurobiological constraints we have devised are solidly grounded in such perspective. Emotions are not organized reactions to the world. They are guesses about what to do next, rooted in prior experience, and the sensory consequences of those guesses (as well as the sights, sounds, smells and other experiences of the world). Emotions may be the Bayesian filters, predictions, or active inferences (what we have referred to as embodied concepts), the representations that typically dominate as intrinsic brain activity. The psychological view of CAT largely draws from such perspective Hutchinson and Barrett (2019). Yet, the formalization of the theoretical model seems to leave on the background this fundamental aspect.

In order to shed some light on this problem, it is best to refer to the model in the more compact form, Eq. 6.5, which we now write as

$$P(\mathcal{S}^{obs}_{1:T}, \mathcal{S}^{hidden}_{1:T},) = \prod_{t=1}^{T} P(\mathcal{S}^{obs}_t \mid \mathcal{S}^{hidden}_t) P(\mathcal{S}^{hidden}_t \mid \mathcal{S}^{hidden}_{t-1}), \tag{6.42}$$

with

$$\mathcal{S}^{hidden}_t = \left\langle \mathcal{I}_t \mathcal{E}_t \mathcal{X}^i_t, \mathcal{X}^e_t, \mathcal{A}_t, \mathcal{C}_t \right\rangle$$

$$\mathcal{S}^{obs}_t = \langle \mathcal{O}_t \rangle$$

Since the variables accounting for the hidden states, as represented through the model PGM, are provided in the form of a hierarchy, we make explicit this fact by rewriting them more abstractly as follows

$$\mathcal{S}^{hidden}_t = \left\langle \mathcal{Z}^{(L)}_t, \cdots \mathcal{Z}^{(l+1)}_t \mathcal{Z}^{(l)}_t \mathcal{Z}^{(l-1)}_t \cdots \mathcal{Z}^{(1)}_t \right\rangle$$

The PGM slice at time t, considering the L levels of latent variables $\mathcal{Z}_t^{(L)}$, factorizes as

$$P(\mathcal{O}_t, \mathcal{Z}_t^{1:L}) = P(\mathcal{O}_t \mid \mathcal{Z}_t^{1:L}, \mathcal{O}_{t-1}) \prod_{l=1}^{L} P(\mathcal{Z}_t^{(l)} \mid \mathcal{Z}_t^{(l+1)}, \mathcal{Z}_{t-1}^{(l)}). \tag{6.43}$$

Thus, the joint density of the hidden and the observable ensembles writes

$$P(\mathcal{O}_{1:T}, \mathcal{Z}_{1:T}^{1:L}) = \prod_{t=1}^{T} P(\mathcal{O}_t \mid \mathcal{Z}_t^{1:L} \mathcal{O}_{t-1}) \prod_{l=1}^{L} P(\mathcal{Z}_t^{(l)} \mid \mathcal{Z}_t^{(l+1)}, \mathcal{Z}_{t-1}^{(l)}). \tag{6.44}$$

This shows that overall our model provides a form of a dynamic hierarchical autoregressive latent variable model.

The hierarchy of the structure is captured by the dependency $\mathcal{Z}_t^{(l+1)} \to \mathcal{Z}_t^{(l)}$. However, the state of the PGM node $\mathcal{Z}_t^{(l)}$ at time t is subject to two sources of information: that of its previous state, $\mathcal{Z}_{t-1}^{(l)} \to \mathcal{Z}_t^{(l)}$, and that from the upper level $\mathcal{Z}_t^{(l+1)} \to \mathcal{Z}_t^{(l)}$. The latter can be conceived as a control input, which can be made explicit, by resorting to the notation used in the control theory literature, using $\mathcal{U}_t^{(l)}$ to represent the control input of $\mathcal{Z}_t^{(l)}$.

Consider for simplicity level $l = 1$, in order to explicitly account for the outcomes $\mathcal{O}_{1:T}$; what follows can be generalized to any level l by straightforwardly considering the $\mathcal{Z}_t^{(l-1)}$ "signal" as the outcome of $\mathcal{Z}_t^{(l)}$ and the downward $\mathcal{Z}_t^{(l+1)}$ signal as the control input signal $\mathcal{U}_t^{(l)}$ to $\mathcal{Z}_t^{(l)}$. In some hierarchical structures, to mirror the structure of level $l = 1$ at any level, an intermediate outcome $\mathcal{O}_t^{(l+1)}$ of $\mathcal{Z}_t^{(l+1)}$ is exploited, which in turn becomes the control input to $\mathcal{Z}_t^{(l)}$, i.e. $\mathcal{O}_t^{(l+1)} = \mathcal{U}_t^{(l)}$ (e.g., Friston, 2008).

Thus, at level $l = 1$, by omitting the level index for notational simplicity, using controls $\mathcal{U}_t$ together with the CIs encoded in the PGM, Eq. 6.5 writes

$$P(\mathcal{O}_{1:T}, \mathcal{Z}_{1:T}, \mathcal{U}_{1:T})) = \prod_{t=1}^{T} P(\mathcal{O}_t \mid \mathcal{Z}_t \mathcal{O}_{t-1}) P(\mathcal{Z}_t \mid \mathcal{Z}_{t-1}, \mathcal{U}_t) P(\mathcal{U}_t). \tag{6.45}$$

Equation 6.45 shows that, in control theory terms, the model is working in the so-called *driven* mode, controlled by an input sequence of observed random vectors $\mathcal{U}_{1:T})$, in which case $\mathcal{O}_{1:T}$ is seen as the output sequence.

To simplify the analysis, by "freezing" as observed at each time step, the given input sequence $\mathcal{U}_{1:T})$, we can focus on modeling the distribution $P(\mathcal{O}_{1:T}, \mathcal{Z}_{1:T} \mid \mathcal{U}_{1:T}))$

$$P(\mathcal{O}_{1:T}, \mathcal{Z}_{1:T} \mid \mathcal{U}_{1:T})) = \prod_{t=1}^{T} P(\mathcal{O}_t \mid \mathcal{Z}_t \mathcal{O}_{t-1}) P(\mathcal{Z}_t \mid \mathcal{Z}_{t-1}, \mathcal{U}_t). \tag{6.46}$$

which represents an input driven autoregressive State Space Model (SSM).

Note that in control theory the driving input $\mathcal{U}_t$ can be used to generate either $\mathcal{Z}_t$, as in our case, or $\mathcal{O}_t$ or both. Also, in other autoregressive models $\mathcal{U}_t$ might in turn depend on $\mathcal{Z}_{t-1}$ and/or $\mathcal{O}_{t-1}$. This corresponds to feedback or closed-loop control in control theory, which is also strongly connected to the concept of autoregressive process, jointly found in the control theory, signal processing or time series analysis literature.

This for what concerns the representational properties of the model. The interesting issue is then to frame the inferential/learning processes given the representation.

The inference problem. We have so far discussed the inferential process in terms of the Monte Carlo approximation which straightforwardly stems from the sampling-based nature of the PPL model.

More generally, a full Bayesian analysis is computationally complex because complicated multiple integrations are involved. Except for specific cases (Gaussian distributions, etc.), involved integrals have no general analytic solution, particularly when the generative model is nonlinear; thus, some algorithmic approximation should be devised.

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

A deterministic approximate approach to the intractable Bayesian inference problem, the variational Bayesian approximation (VB), has been introduced (see Beal and Ghahramani, 2003; MacKay, 2004 for an insightful discussion). Rather than use sampling, the main idea behind variational inference is to use optimisation. In a nutshell (but see Appendix C for the mathematical details), it relies on the following steps:

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

Variational Bayes draws together variational ideas from the analysis of intractable latent variable models and from Bayesian inference. This framework facilitates analytical calculation of posterior distributions over the hidden variables, parameters and structures. For instance, they might be computed via an iterative algorithm, VBEM, a generalization of the classic EM algorithm (Beal and Ghahramani, 2003). Variational approximations rely on bound approximation, by adopting approximating posteriors. Further, if a fixed-form approximation is adopted, this choice allows one to represent the density in terms of a small number of quantities, namely its sufficient statistics.

Interestingly enough, the neural/realisation level plausibility (Marr, 1982) of each approach is currently matter of a fierce debate in the theoretical neuroscience field.

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

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

The predictive brain hypothesis is closely connected to the variational approach. Thus, by embracing the optimization perspective, we can proceed as follows. Consider the generalization of Eq. 6.46, which fully factorizes the conditional joint distribution $P(\mathcal{O}_{1:T}, \mathcal{Z}_{1:T} \mid \mathcal{U}_{1:T})$:

$$P(\mathcal{O}_{1:T}, \mathcal{Z}_{1:T} \mid \mathcal{U}_{1:T}) = \prod_{t=1}^{T} P(\mathcal{O}_t \mid \mathcal{Z}_{1:t}\mathcal{O}_{1:t-1}, \mathcal{U}_{1:t}) P(\mathcal{Z}_t \mid \mathcal{O}_{1:t-1}, \mathcal{Z}_{1:t-1}, \mathcal{U}_{1:t}). \tag{6.47}$$

The distribution of interest for solving the inference problem is the posterior distribution of state sequence $\mathcal{Z}_{1:T}$, $P_\theta(\mathcal{Z}_{1:T} \mid \mathcal{O}_{1:T}, \mathcal{U}_{1:T})$, where we have explicitly denoted the parameters θ characterizing such distribution, P_θ; note, however, that this is just a matter of notational convenience; in the Bayesian setting the θ parameters are just latent RVs much like $\mathcal{Z}_{1:T}$ and the former could be easily incorporated as a subset of the latter.

The posterior P_θ is generally intractable. Hence, in the variational setting, one defines an inference model $Q_\phi(\mathcal{Z}_{1:T} \mid \mathcal{O}_{1:T}, \mathcal{U}_{1:T})$, which is an approximation of the posterior P_θ. The index $\phi \in \mathbf{\Phi}$ explicitly refers to the set of variational parameters the must be "adjusted" to minimize the divergence between the true and the approximating distributions.

It is reasonable to assume that a good candidate for Q_ϕ would have the same structure as the exact posterior distribution P_θ. This means that variable dependencies used to factorize $P(\mathcal{O}_{1:T}, \mathcal{Z}_{1:T} \mid \mathcal{U}_{1:T})$, in virtue of the d-separation property of PGMs, can be used to simplify posterior dependencies in Q_ϕ.

Variational inference is based on the maximization of the variational free energy $\mathcal{F}(Q)$ also named, in the current deep learning literature, the evidence lower bound $ELBO(Q)$; for a detailed and formal definition see Appendix C. In the case of the model in Eq. 6.47 the free energy can be generally written as

$$\mathcal{F}(Q) = \mathbb{E}_{Q_\phi(\mathcal{Z}_{1:T}|\mathcal{O}_{1:T},\mathcal{U}_{1:T})}\left[\ln P_\theta(\mathcal{Z}_{1:T}, \mathcal{O}_{1:T} \mid \mathcal{U}_{1:T})\right] - \tag{6.48}$$

$$\mathbb{E}_{Q_\phi(\mathcal{Z}_{1:T}|\mathcal{O}_{1:T},\mathcal{U}_{1:T})}\left[\ln Q_\phi(\mathcal{Z}_{1:T} \mid \mathcal{O}_{1:T},\mathcal{U}_{1:T})\right] \tag{6.49}$$

Exploiting the factorization of Eq. 6.47 and the properties of the expectation $\mathbb{E}[\cdot]$ with respect to conditional distributions,

$$\mathbb{E}_{Q_\phi(\mathcal{Z}_{1:T}|\mathcal{O}_{1:T},\mathcal{U}_{1:T})}\left[f(\mathcal{Z}_{1:T})\right] = \tag{6.50}$$

$$\mathbb{E}_{Q_\phi(\mathcal{Z}_1|\mathcal{O}_{1:T},\mathcal{U}_{1:T})}\left[\mathbb{E}_{Q_\phi(\mathcal{Z}_2|\mathcal{Z}_1,\mathcal{O}_{1:T},\mathcal{U}_{1:T})}\left[\cdots\mathbb{E}_{Q_\phi(\mathcal{Z}_T|\mathcal{Z}_{1:T},\mathcal{O}_{1:T},\mathcal{U}_{1:T})}\left[f(\mathcal{Z}_{1:T})\right]\right]\right],$$

where $f(\mathcal{Z}_{1:T})$ is a generic function of $\mathcal{Z}_{1:T}$, Eq. 6.49 writes

$$\mathcal{F}(Q) = \sum_{t=1}^{T}\mathbb{E}_{Q_\phi(\mathcal{Z}_{1:t}|\mathcal{O}_{1:T},\mathcal{U}_{1:T})}\left[\ln P_{\theta_\mathcal{O}}(\mathcal{O}_t \mid \mathcal{O}_{1:t-1}\mathcal{Z}_{1:t},\mathcal{U}_{1:t})\right] - \tag{6.51}$$

$$\sum_{t=1}^{T}\mathbb{E}_{Q_\phi(\mathcal{Z}_{1:t-1}|\mathcal{O}_{1:T},\mathcal{U}_{1:T})}\left[KL(Q_\phi(\mathcal{Z}_t \mid \mathcal{Z}_{1:t-1},\mathcal{O}_{1:T},\mathcal{U}_{1:T})\|P_{\theta_\mathcal{Z}}(\mathcal{Z}_t \mid \mathcal{O}_{1:t-1},\mathcal{Z}_{1:t-1},\mathcal{U}_{1:t}))\right]$$

Note that, in principle, the $\mathcal{F}(Q)$ should be maximized with respect to all parameters $\phi, \theta_\mathcal{Z}, \theta_\mathcal{O}$. A general procedure would require to compute Monte Carlo estimates (i.e., empirical averages), using samples drawn from $Q_\phi(\mathcal{Z}_{1:\tau} \mid \mathcal{O}_{1:T},\mathcal{U}_{1:T})$, with $\tau \in \{1,\cdots,T\}$, an arbitrary time index. Sampling each random vector $\mathcal{Z}_{1:\tau}$ at a given time instant is straightforward, as $Q_\phi(\mathcal{Z}_{1:t} \mid \mathcal{O}_{1:T},\mathcal{U}_{1:T})$ is analytically specified by the chosen inference model (often a Gaussian distribution). Doing so, the variational free energy becomes differentiable and can then be optimized using gradient-ascent-based algorithms such as the CAVI.

In the case of the classic SSM, the model simplifies as follows:

$$P_{\theta_\mathcal{O}}(\mathcal{O}_t \mid \mathcal{O}_{1:t-1},\mathcal{Z}_{1:t},\mathcal{U}_{1:t}) = P_{\theta_\mathcal{O}}(\mathcal{O}_t \mid \mathcal{Z}_t) \tag{6.52}$$

$$P_{\theta_\mathcal{Z}}(\mathcal{Z}_t \mid \mathcal{O}_{1:t-1},\mathcal{Z}_{1:t-1},\mathcal{U}_{1:t}) = P_{\theta_\mathcal{Z}}(\mathcal{Z}_t \mid \mathcal{Z}_{t-1},\mathcal{U}_t), \tag{6.53}$$

thus,

$$P_{\theta_\mathcal{O}}(\mathcal{O}_{1:T} \mid \mathcal{Z}_{1:T}) = \prod_{t=1}^{T} P_{\theta_\mathcal{O}}(\mathcal{O}_t \mid \mathcal{Z}_t) \tag{6.54}$$

$$P_{\theta_\mathcal{Z}}(\mathcal{Z}_{1:T} \mid \mathcal{U}_{1:T}) = \prod_{t=1}^{T} P_{\theta_\mathcal{Z}}(\mathcal{Z}_t \mid \mathcal{Z}_{t-1},\mathcal{U}_t). \tag{6.55}$$

Then, the variational free energy for the SSM can be written as

$$\mathcal{F}(Q) = \sum_{t=1}^{T}\mathbb{E}_{Q_\phi(\mathcal{Z}_t|\mathcal{O}_{1:T},\mathcal{U}_{1:T})}\left[\ln P_{\theta_\mathcal{O}}(\mathcal{O}_t \mid \mathcal{Z}_t)\right] - \tag{6.56}$$

$$\sum_{t=1}^{T}\mathbb{E}_{Q_\phi(\mathcal{Z}_{t-1}|\mathcal{O}_{1:T},\mathcal{U}_{1:T})}\left[KL(Q_\phi(\mathcal{Z}_t \mid \mathcal{Z}_{t-1},\mathcal{O}_{t:T},\mathcal{U}_{t:T})\|P_{\theta_\mathcal{Z}}(\mathcal{Z}_t \mid \mathcal{Z}_{t-1},\mathcal{U}_t))\right].$$

This last representation allows us to relate the model to the predictive coding framework.

6.6.1 Connection to predictive coding

The optimization perspective on the inferential step that we have discussed above is based on the variational free energy key concept. The free energy principle (but see Appendix C) for a detailed mathematical treatment) has been set as the basis for a general theory of the necessary information-theoretic behaviours of systems which maintain a separation from their environment. A core postulate of the theory is that complex systems can be seen as performing variational Bayesian inference and minimizing the variational free energy. The free energy principle originated in, and has been extremely influential in theoretical neuroscience, having spawned a number of neurophysiologically realistic process theories, while maintaining close links with the Bayesian Brain viewpoints.

The motivation for the free-energy principle is simple but fundamental, as stated by Friston (2009) . It rests upon the fact that self- organising biological agents resist a tendency to disorder and therefore minimize the entropy of their sensory states. Minimizing entropy corresponds to suppressing surprise over time. In brief, for a well defined agent to exist it must occupy a limited repertoire of states. This means the equilibrium density of an ensemble of agents, describing the probability of finding an agent in a particular state, must have low entropy: a distribution with low entropy just means a small number of states are occupied most of the time. Because entropy is the long-term average of surprise, agents must avoid surprising states. Yet, agents cannot evaluate surprise directly. This would entail knowing all the hidden states of the world causing sensory input. However, an agent can avoid surprising exchanges with the world if it minimises its free-energy because free- energy is always bigger than surprise.

These arguments suggest that biological systems sample their environment to fulfil expectations that are generated by the model implicit in their structure. From the agent perspective, the environment is an accommodating place; fluctuations or displacements caused by environmental forces are quickly explained away by adaptive re-sampling. How do living systems preserve their order (i.e. configurational entropy), immersed in an environment that is becoming irrevocably more disordered? The premise here is that the environment unfolds in a thermodynamically structured and lawful way and biological systems embed these laws into their anatomy. The existence of environmental order is assured, at the level of probability distributions, through thermodynamics. Systems that minimise the surprise of their interactions with the environment by adaptive sampling can only do so by optimising a bound, which is a function of the system's states. This is exactly the physical (variational) free energy, or Gibbs' free energy $\mathcal{F}_G$ (Appendix C): when the free-energy is minimised, the ensemble density encoded by the system's parameters becomes an approximation to the posterior probability of the causes of its sensory input. The free-energy principle, in its simplicity, states that systems change to decrease their physical free-energy. The concept of free-energy arises in many contexts, especially physics and statistics. In thermodynamics, free-energy is a measure of the amount of work that can be extracted from a system: it is the difference between the energy and the entropy of a system (Appendix C).

The free energy approach has a clear connection to the theory of predictive coding, a neurobiologically plausible process theory. Indeed, predictive coding can be derived from the free energy principle under certain assumptions

The core of the predictive coding in the brain approach is based on two fundamental concepts. First, the idea of prediction error as the difference between the activity of the neurons in a layer and the top-down predictions from higher layers. Second, the update rule for the activities of a layer, which minimizes both the prediction errors at its own layer, as well as the layer below.

In what follows, in order to keep up with the physical view of the free energy postulated by the free energy approach to the brain, we will explicitly denote $\widetilde{\mathcal{F}}$ the variational free energy in the sense of Gibbs' free energy, thus (cfr., Eq. C.41, Appendix C)

$$\widetilde{\mathcal{F}}(\mathbf{\Phi}) = \mathcal{F}_G = -\mathcal{F}, \tag{6.57}$$

where, as previously defined, $\mathbf{\Phi}$ stands for the set of variational parameters.

Given the joint distribution $P(\mathcal{O}_{1:T}, \mathcal{Z}_{1:T}^{1:L} \mid \mathcal{U}_{1:T}^{1:L})$, consider the specific case of the SSM family that we have previously dicscussed: the linear-Gaussian state space model (LG-SSM):

$$P(\mathcal{Z}_t \mid \mathcal{Z}_{t-1}, \mathcal{U}_t) = \mathcal{N}(\mathcal{Z}_t \mid \mathbf{A}_t\mathcal{Z}_{t-1} + \mathbf{B}_t\mathcal{U}_t, \mathbf{Q}_t) \tag{6.58}$$

$$P(\mathcal{O}_t \mid \mathcal{Z}_t, \mathcal{U}_t) = \mathcal{N}(\mathcal{O}_t \mid \mathbf{C}_t\mathcal{Z}_t + \mathbf{D}_t\mathcal{U}_t, \mathbf{R}_t) \tag{6.59}$$

To simplify the notation, for the moment, we have dropped the conditioning on the inputs $\mathcal{U}_t$, and again we have assumed the control provided by the upper layer is observed (known).

For formalizing such approximation we resort to the framework of inference as optimization, as used in variational inference.

First, we rewrite the mean-field approximation (Eq. C.9, Appendix C) for the approximating posterior $Q_\phi(\mathcal{Z}_{1:T})$, which gives the fully factorized form:

$$Q_\phi(\mathcal{Z}_{1:T}) = \prod_{t=1}^{T} Q_\phi(\mathcal{Z}_t) \tag{6.60}$$

Recall that, such temporal factorization of the free energy means that the minimization at each time step is independent of the others.

Then, the variational free energy expressed by Eq. 6.56 can be compactly written in terms of the physical free energy $\widetilde{\mathcal{F}}$:

$$\widetilde{\mathcal{F}}(\mathbf{\Phi}) = \sum_{t=1}^{T} \widetilde{\mathcal{F}}_t(\mathbf{\Phi}) = \sum_{t=1}^{T} \mathbb{E}_{Q_\phi(\mathcal{Z}_t)}\left[KL(Q_\phi(\mathcal{Z}_t) \| P(\mathcal{Z}_t, \mathcal{O}_t \mid \mathcal{Z}_{t-1}))\right] \tag{6.61}$$

To develop it further, first, consider a Gaussian approximation for the variational posterior Q at time step t,

$$Q_\phi(\mathcal{Z}_t) = Q(\mathcal{Z}_t \mid \boldsymbol{\mu}_t) = \mathcal{N}(\mathcal{Z}_t \mid \boldsymbol{\mu}_t, \mathbf{\Sigma}_t(\boldsymbol{\mu}_t)), \tag{6.62}$$

where $\boldsymbol{\mu}_t = \mathbf{\Phi}$ identifies the variational parameter of interest; the latter, once computed, allows to derive the covariance matrix $\mathbf{\Sigma}_t(\boldsymbol{\mu}_t)$.

In the mean-field approximation, Eq. 6.60, we can consider the single time step; all time steps will be identical in terms of the solution method. Then,

$$\widetilde{\mathcal{F}}_t(\boldsymbol{\mu}_t) = -\mathbb{E}_{Q(\mathcal{Z}_t|\boldsymbol{\mu}_t)}\left[\ln P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1})\right] - \mathbb{H}(Q(\mathcal{Z}_t \mid \boldsymbol{\mu}_t)) \tag{6.63}$$

The entropy of a Gaussian $\mathbb{H}(Q(\mathcal{Z}_t \mid \boldsymbol{\mu}_t)$ does not depend on the variational parameter $\boldsymbol{\mu}_t$, thus it can be omitted from the variational computation, which boils down to optimizing

$$\widetilde{\mathcal{F}}_t(\boldsymbol{\mu}_t) = -\mathbb{E}_{Q(\mathcal{Z}_t|\boldsymbol{\mu}_t)}\left[\ln P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1})\right] \tag{6.64}$$

Use then the saddle-point or Laplace approximation in Eq. 6.64, that is the second order Taylor series approximation around the mode

$$\begin{aligned}
&\mathbb{E}_{Q(\mathcal{Z}_t|\boldsymbol{\mu}_t)}\left[\ln P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1})\right] \approx && (6.65)\\
&\mathbb{E}_{Q(\mathcal{Z}_t|\boldsymbol{\mu}_t)}\left[\ln P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1})\right] + && (6.66)\\
&\mathbb{E}_{Q(\mathcal{Z}_t|\boldsymbol{\mu}_t)}\left[\nabla_{\mathcal{Z}_t} P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1}) \mid_{\mathcal{Z}_t=\boldsymbol{\mu}_t} (\mathcal{Z}_t - \boldsymbol{\mu}_t)\right] + && (6.67)\\
&\mathbb{E}_{Q(\mathcal{Z}_t|\boldsymbol{\mu}_t)}\left[\nabla^2_{\mathcal{Z}_t} P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1}) \mid_{\mathcal{Z}_t=\boldsymbol{\mu}_t} (\mathcal{Z}_t - \boldsymbol{\mu}_t)^2\right] = && (6.68)\\
&\ln P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1}) + && (6.69)\\
&\nabla_{\mathcal{Z}_t} P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1}) \mid_{\mathcal{Z}_t=\boldsymbol{\mu}_t} \mathbb{E}_{Q(\mathcal{Z}_t|\boldsymbol{\mu}_t)}\left[(\mathcal{Z}_t - \boldsymbol{\mu}_t)\right] + && (6.70)\\
&\nabla^2_{\mathcal{Z}_t} P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1}) \mid_{\mathcal{Z}_t=\boldsymbol{\mu}_t} \mathbb{E}_{Q(\mathcal{Z}_t|\boldsymbol{\mu}_t)}\left[(\mathcal{Z}_t - \boldsymbol{\mu}_t)^2\right] = && (6.71)\\
&\ln P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1}) + \nabla^2_{\mathcal{Z}_t} P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1}) \mid_{\mathcal{Z}_t=\boldsymbol{\mu}_t} \boldsymbol{\Sigma}_t && (6.72)
\end{aligned}$$

where the linearity of the expectation has been used and that $\mathbb{E}_{Q(\mathcal{Z}_t|\boldsymbol{\mu}_t)}\left[(\mathcal{Z}_t - \boldsymbol{\mu}_t)\right] = 0$.

Thus, by neglecting the term independent of $\boldsymbol{\mu}_t$, the optimization problem to solve is

$$\begin{aligned}
\boldsymbol{\mu}_t^{opt} &= \arg\min_{\boldsymbol{\mu}_t} \widetilde{\mathcal{F}}_t(\boldsymbol{\mu}_t) = \arg\min_{\boldsymbol{\mu}_t} \ln P(\mathcal{Z}_t, \mathcal{O}_t \mid \boldsymbol{\mu}_{t-1}) && (6.73)\\
&= \arg\min_{\boldsymbol{\mu}_t} -\{(\mathcal{O}_t - \mathbf{C}\boldsymbol{\mu}_t))^\top \boldsymbol{\Sigma}_{\mathcal{O}}^{-1} (\mathcal{O}_t - \mathbf{C}\boldsymbol{\mu}_t)) + && (6.74)\\
&\quad (\boldsymbol{\mu}_t - \mathbf{A}\boldsymbol{\mu}_{t-1} - \mathbf{B}\mathcal{U}_{t-1})^\top \boldsymbol{\Sigma}_{\mathcal{Z}}^{-1} (\boldsymbol{\mu}_t - \mathbf{A}\boldsymbol{\mu}_{t-1} - \mathbf{B}\mathcal{U}_{t-1})\} && (6.75)
\end{aligned}$$

where the dependence of $\mathcal{Z}_t$ on the control signal has been reintroduced for completeness, but setting $\mathbf{D}_t = \mathbf{0}$ in Eq. 6.59, since we are not considering the dependence of outcomes $\mathcal{O}_t$ on control $\mathcal{U}_t$ (cfr. Eq. 6.53).

The problem posed in Eq. 6.73 can be solved by gradient descent. By using standard results on the derivative of quadratic forms[4]:

$$\nabla \widetilde{\mathcal{F}}_t(\boldsymbol{\mu}_t) = -\mathbf{C}^\top \boldsymbol{\Sigma}_{\mathcal{O}}^{-1} (\mathcal{O}_t - \boldsymbol{\mu}_t) + \boldsymbol{\Sigma}_{\mathcal{Z}}^{-1} (\boldsymbol{\mu}_t - \mathbf{A}\boldsymbol{\mu}_{t-1} - \mathbf{B}\mathcal{U}_{t-1}). \tag{6.76}$$

Equation 6.76 tells that minimizing the variational free energy $\widetilde{\mathcal{F}}_t(\boldsymbol{\mu}_t)$ entails the minimization of two errors:

[4] $\frac{\partial}{\partial \mathbf{s}}(\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s}) = -2\mathbf{A}^\top \mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s})$ and $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{s})^\top \mathbf{W}(\mathbf{x} - \mathbf{s}) = 2\mathbf{W}(\mathbf{x} - \mathbf{s})$, $\mathbf{W}$ being symmetric, Petersen et al. (2008)

- the error between the current observation and the prediction current observation

$$\mathbf{e}_{\mathcal{O}} = \mathcal{O}_t - \mathbf{C}_t \boldsymbol{\mu}_{t|t-1} \tag{6.77}$$

  weighted by precision $\boldsymbol{\Sigma}_{\mathcal{O}}^{-1}$

- the error between the current state/variational parameter prediction current observation and the past state estimate;

$$\mathbf{e}_{\mathcal{Z}} = \boldsymbol{\mu}_t - \mathbf{A}\boldsymbol{\mu}_{t-1} - \mathbf{B}\mathcal{U}_{t-1} \tag{6.78}$$

  weighted by precision $\boldsymbol{\Sigma}_{\mathcal{Z}}^{-1}$.

The gradient perfectly recapitulates the standard predictive coding scheme with precision weighted prediction errors. In fact under the LG-SSM assumption the minimization described by Eq. 6.76 is exactly the same result obtained via Kalman filtering, an optimal linear Bayesian filtering algorithm, (Eqs. B.6 and 6.59).

Kalman filtering analytically solves this objective directly, while in predictive coding the dynamics of the parameters are set to be a gradient descent on the variational free energy: this reduces to the MAP objective solved by the Kalman Filter (cfr., Eq. B.13 . To sum up, predictive coding and Kalman filtering derive their dynamics from a gradient descent on the same objective.

Clearly, this result is easy to extend to hierarchical and nonlinear models. Also in this case the procedure is to minimize the free energy using gradient descent.

Further, the procedure can be extended by optimize the free energy with respect to the model parameters.

In the case of LG-SSMs, following the same gradient calculation as in Eq. 6.76, we obtain the following results: for the dynamics matrix $\mathbf{A}$,

$$\nabla_{\mathbf{A}} \widetilde{\mathcal{F}}_t = -\boldsymbol{\Sigma}_{\mathcal{Z}} \mathbf{e}_{\mathcal{O},t} \boldsymbol{\mu}_{t-1}^{\top}; \tag{6.79}$$

for the control matrix $\mathbf{B}$,

$$\nabla_{\mathbf{B}} \widetilde{\mathcal{F}}_t = -\boldsymbol{\Sigma}_{\mathcal{Z}} \mathbf{e}_{\mathcal{O},t} \mathcal{U}_{t-1}^{\top}; \tag{6.80}$$

eventually, for the observation matrix $\mathbf{C}$,

$$\nabla_{\mathbf{C}} \widetilde{\mathcal{F}}_t = -\boldsymbol{\Sigma}_{\mathcal{Z}} \mathbf{e}_{\mathcal{O},t} \boldsymbol{\mu}_{t}^{\top}. \tag{6.81}$$

Interestingly enough, the above equations can be generalized to nonlinear models showing that predictive coding can approximate backpropagation for many kinds of models Hosseini and Maida (2020).

To sum up, the theoretical model formalizing the infrastructure that enables the conceptual and communication acts provides a generalized form of Bayesian predictive processing.

6.6.2 A stochastic process perspective

In the previous Section we have analyzed the model from a dynamic input-state-output model to account for its overall dynamics. In particular we have seen that when the model is reduced, under suitable assumptions to a hierarchical LG-SSM, its Gaussian dynamics can be interpreted in terms of the basic predictive processing hypothesis. At the different levels of the hierarchy, all state-space denoting variables actually are stochastic processes, even though, for notational simplicity and presentation convenience, we have treated them as one realisation of the process.

Let us now give a more precise shape to the stochastic view of the theoretical model and to related implications.

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

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

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

$$d\mathcal{Z}_t = f(\mathcal{Z}_t,\mathcal{U}_t)dt + D^{1/2}dW_t, \tag{6.82}$$

$$d\mathcal{O}_t = g(\mathcal{Z}_t,\mathcal{U}_t)dt + R^{1/2}dV_t, \tag{6.83}$$

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

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

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

In fact, Eqs. 6.82 and 6.83 can be easily recognised as diffusion processes, f and g being their respective drifts (Van Kampen, 2011).

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

$$\mathcal{Z}_{t+\Delta t} - \mathcal{Z}_t = f(\mathcal{Z}_t,\mathcal{U}_t)\Delta t + D^{1/2}\sqrt{\Delta t}\epsilon_{\mathcal{Z}_t}, \tag{6.84}$$

$$\mathcal{O}_{t+\Delta t} - \mathcal{O}_t = g(\mathcal{Z}_t,\mathcal{U}_t)\Delta t + R^{1/2}\sqrt{\Delta t}\epsilon_{\mathcal{O}_t}, \tag{6.85}$$

Equations 6.84 and 6.85 are known as the Euler-Maruyama approximation of Eqs 6.82 and 6.83.

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

Then, Eqs. 6.82 and 6.83 define a generalised input-output state-space system (SSM) where the states $\mathcal{Z}_t$ mediate the influence of the input on the output and endow the system with memory. The state and observation perturbations or fluctuations are provided by noise terms $\epsilon_{\mathcal{Z}},\epsilon_{\mathcal{O}}$, which can be defined via the stochastic integrals $W_t = \int_0^t \epsilon_{\mathcal{Z}_s} ds$, $V_t = \int_0^t \epsilon_{\mathcal{O}_s} ds$. In the case of W,V being Wiener processes, $\epsilon_{\mathcal{Z}},\epsilon_{\mathcal{O}}$ represent Gaussian additive noise, and have the same dimension of $\mathcal{Z},\mathcal{O}$, respectively. If errors are iid Gaussian random variables, then the specific scaling of the white noise with Δt gives rise to the nondifferentiable trajectories of sample paths characteristic for a diffusion process.

The classic input-output SSM can be recovered from Eqs. 6.82 and 6.83, under the independence assumption $(\mathcal{O}_t \perp \mathcal{O}_{t-1} \mid \mathcal{Z}_t)$:

$$d\mathcal{Z}_t = f(\mathcal{Z}_t,\mathcal{U}_t)dt + D^{1/2}dW_t, \tag{6.86}$$

$$\mathcal{O}_t = g(\mathcal{Z}_t,\mathcal{U}_t) + R^{1/2}\epsilon_{\mathcal{O}_s}, \tag{6.87}$$

Obviously, considering Eqs. 6.84 and 6.85, under the conditional independence assumption $(\mathcal{O}_{t+\Delta t} \perp \mathcal{O}_t \mid \mathcal{Z}_{t+\Delta t})$, then $\mathcal{O}_t$ only depends on $\mathcal{Z}_t,\mathcal{U}_t$ and we recover the discrete time input-output SSM. In such case, the stochastic difference equations can be easily obtained.

Equations 6.82 and 6.83 above formalise the generative process: when the dynamics unfolds, the process $\mathcal{O}$ generates a σ-algebra.

Denote $\mathcal{O}_t = \mathcal{O}_{0:t} = \{\mathcal{O}_0, \mathcal{O}_1, \cdots, \mathcal{O}_t\}$ a filtration. We can define then an innovation process $E = \{E_t, 0 \le t \le T\}$ or *prediction error*

$$E_t = \mathcal{O}_t - \int_0^t \mathbb{E}_\theta[g(\mathcal{Z}_s,\mathcal{U}_s) \mid \mathcal{O}_{0:s},\mathcal{U}_{0:s}]ds \tag{6.88}$$

which is central when we want to invert the model. Notably, this is the stochastic process counterpart of the predictive error in predictive coding.

The framework defined via Eqs. 6.82 and 6.83 ensures that the probability measures in $\mathcal{M}$ are mutually absolutely continuous. Then, the connection of the stochastic process to the Bayesian setting can be made via the Radon-Nikodym Theorem (Stuart, 2010). If we let θ_0 be the reference set of parameter and write $\mathbb{P}_{\theta_0}$ as $\mathbb{P}_0$ (prior measure, with associated prior density P_0), the Radon-Nikodym derivative of $\mathbb{P}_\theta$ with respect to $\mathbb{P}_0$ provides the complete data likelihood.

$$\frac{d\mathbb{P}_\theta}{d\mathbb{P}_0} = \frac{P_\theta(\mathcal{O}_t \mid \mathcal{U}_t)}{P_\theta(\mathcal{O}_{0:T} \mid \mathcal{U}_{0:T})} \tag{6.89}$$

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

$$\mathcal{L}(\theta \mid \mathcal{O}_{0:T}) = E_0 \left[\frac{d\mathbb{P}_\theta^{\mathcal{O}}}{d\mathbb{P}_0} \mid \mathcal{O}_{0:T}, \mathcal{U}_{0:T} \right]. \tag{6.90}$$

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

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

A key observation we have discussed in the previous section relates to the fact that the optimisation of the free energy in dynamical systems involves the optimisation of the prediction error entailed by the optimisation of the expected internal energy. In stochastic processes terms this corresponds to the innovation (Eq. 6.88) of the process,

One interesting example has been provided by Daunizeau et al. (2009), for the case of the classic input-output SSM. This as previously mentioned, can be recovered from Eqs. 6.82 and 6.83, under the independence assumption $(\mathcal{O}_t \perp \mathcal{O}_{t-1} \mid \mathcal{Z}_t)$, rewritten more compactly, with some abuse of notation:

$$\dot{\mathcal{Z}}_t = f(\mathcal{Z}_t, \mathcal{U}_t) + \epsilon_{\mathcal{Z}_t}, \tag{6.91}$$

$$\mathcal{O}_t = g(\mathcal{Z}_t, \mathcal{U}_t) + \epsilon_{\mathcal{O}_t}, \tag{6.92}$$

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

In this case, under Gaussian assumptions on the state and observation noises and under the Euler-Maruyama discretisation scheme (Eqs. 6.84, 6.85), the discrete-time variant of the state-space model, by setting unitary time step $\Delta t = 1$ yields the Gaussian likelihood and transition densities

$$\mathcal{Z}_{t+1} \sim P_{\mathcal{Z}}(\mathcal{Z}_{t+1} \mid \mathcal{Z}_t, \mathcal{U}_t, \theta_{\mathcal{Z}}) = \mathcal{N}(f(\mathcal{Z}_t, \mathcal{U}_t), D^{-1}), \tag{6.93}$$

$$\mathcal{O}_t \sim P_{\mathcal{O}}(\mathcal{O}_t \mid \mathcal{Z}_t, \mathcal{U}_t, \theta_{\mathcal{O}}) = \mathcal{N}(g(\mathcal{Z}_t, \mathcal{U}_t), R^{-1}) \tag{6.94}$$

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

Cogently, based on the stochastic process view of affect dynamics Oravecz et al. (2011) have developed a model for temporal fluctuations in the core affect state over time, with individual differences for the crucial parameters. The core of the model can be described in terms of two equations. They focus on a particular case of the system described by Eqs. 6.82, 6.83, where state equation (6.82) is assumed to be the Ornstein-Uhlenbeck (OU) process (Uhlenbeck and Ornstein, 1930).

$$d\mathcal{Z}_t = \beta(\mathcal{U} - \mathcal{Z}_t)dt + D^{1/2} dW_t, \tag{6.95}$$

$$\mathcal{O}_t = \mathcal{Z}_t + R^{1/2} \epsilon_{\mathcal{O}_s}, \tag{6.96}$$

where $\beta > 0$ and control $\mathcal{U}$ in the simplest case is assumed to be a constant, i.e. $\mathcal{U} = const$, or time-varying in the most general case. The instantaneous change in $\mathcal{Z}_t$, that is, $d\mathcal{Z}_t$, depends on how far the current state $\mathcal{Z}_t$ is from the point $\mathcal{U}$. This control parameter is called a steady state or attractor: as a straightforward example in the one dimensional case, if $\mathcal{Z}_t$ is below $\mathcal{U}$ (i.e., $\mathcal{U} - \mathcal{Z}_t < 0$), the first derivative is positive, and consequently $\mathcal{Z}_t$ will increase; the opposite holds when $\mathcal{Z}_t$ is above. The parameter β controls the magnitude of the "attraction" effect: if β is large ($\beta \gg 1$), the difference between the actual state and $\mathcal{U}$ tends to be magnified; therefore a faster change will occur in the direction of $\mathcal{U}$; with small β), the change becomes substantially slower. Based on this property, the parameter is often called the dampening force or centralising tendency. The stochastic innovation term dW_t incorporates the multiple smaller and larger impacts that the core affect system undergoes at a given moment (Fig. 6.15).

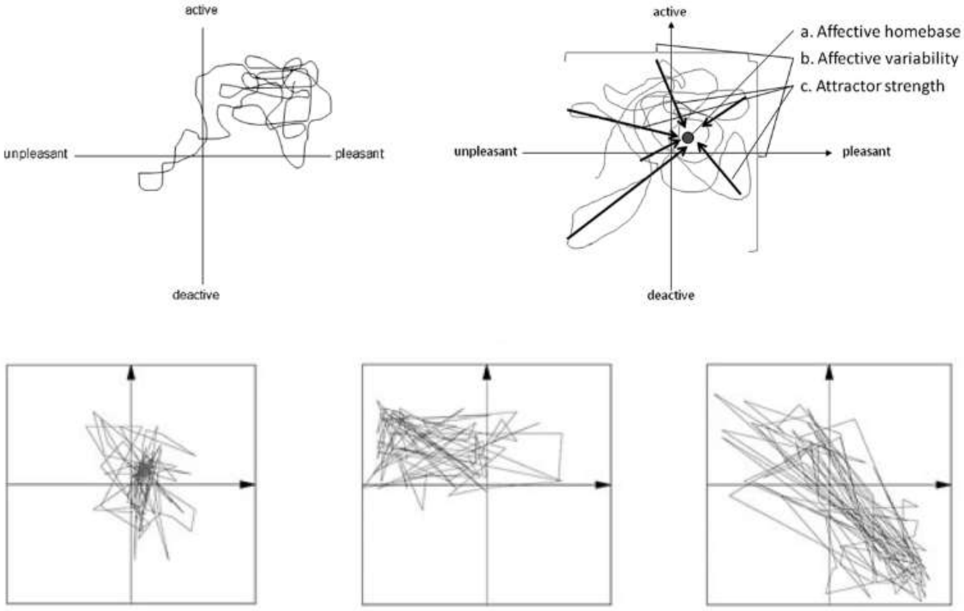


Figure 6.15: *The core affect dynamics according to Kuppens et al. (2010). The top panel shows the key features of the model, where the dynamics is an instance of an OU stochastic process; the bottom panel, shows the observed core affect trajectories of three participants involved in the experience-sampling study on people's core affective experiences. Adapted from Kuppens et al. (2010)*

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

To sum up, these three key processes - affective home base, variability, and attractor strength -are largely responsible for producing the myriad ways people can display changes and fluctuations in their core affect throughout daily life (e.g., Fig. 6.15, bottom panel, (Kuppens et al., 2010).

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

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

Indeed, the hierarchy we have stated from the beginning in this theoretical analysis defines a hierarchy of stochastic processes, which can be written generalising the input-output SSM (Eqs, 6.91, 6.92) as spanning on $l = 1, \cdots, L$ levels:

$$\begin{aligned}
\dot{\mathcal{U}}_t^{(L)} &= g(\mathcal{U}_t^{(L+1)}) + \epsilon_{\mathcal{U}_t}^{(L+1)}, \\
&\vdots \\
\dot{\mathcal{Z}}_t^{(l)} &= f(\mathcal{Z}_t^{(l)}, \mathcal{U}_t^{(l)}) + \epsilon_{\mathcal{Z}_t}^{(l)}, \\
\dot{\mathcal{U}}_t^{(l-1)} &= g(\mathcal{Z}_t^{(l)}, \mathcal{U}_t^{(l)}) + \epsilon_{\mathcal{U}_t}^{(l)}, \\
&\vdots \\
\dot{\mathcal{Z}}_t^{(1)} &= f(\mathcal{Z}_t^{(1)}, \mathcal{U}_t^{(1)}) + \epsilon_{\mathcal{Z}_t}^{(1)}, \\
\dot{\mathcal{O}}_t &= g(\mathcal{Z}_t, \mathcal{U}_t) + \epsilon_{\mathcal{O}_t}
\end{aligned} \tag{6.97}$$

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

In other terms, what we expect from an implementation of our model is that, under suitable constraints (e.g., Gaussian SSM assumption), the core affect dynamics of the stochastic process $F = \{F_t, 0 \leq t \leq T\}$ exhibits similar properties to the OU process posited and empirically measured by Oravecz et al. (2011).

6.7 A brief remark on the Implementation Model

As pointed out in Chapter 3, moving from the level of the theoretical model to the level of the implementation model hides many subtleties. If a fine grain scale of implementation is not addressed, which might be the case for affective computing and NLP purposes, there are, at least in principle, different viable solutions. Since our model is essentially a Bayesian model, the first question to consider *prima facie* is whether to pursue a straight Monte Carlo approximation or opt for optimization, variational techniques, as considered in a straight predictive processing perspective. These are the two most prominent strategies for approximating Bayesian inference: in the first case the chain is run until it has hopefully reached equilibrium and collect samples to approximate the posterior. In variational inference, as seen a flexible family of distributions is defined over the hidden variables, indexed by free parameters, then a setting of the parameters is found that is closest to the posterior, solving an optimization problem.

Yet, in current machine learning practice this the border between these to alternatives is often blurred; one clear example is provided by generative deep neural networks, such as variational autoencoders (VAE) or generative adversarial networks (GAN) and related architectures, where sampling and optimization are intertwined. Another example, as to learning, is stochastic variational inference (SVI) that has been conceived for efficiently analyzing massive data sets with complex probabilistic models Hoffman et al. (2013)

In the predictive coding perspective the general SSM model that underpins our model might be implemented in terms of (hierarchy of) dynamical VAEs (DVAE). A DVAE can be seen as a combination of a VAE and a a recurrent neural network (RNN). One such example is provided by architectures based on variational recurrent neural networks (VRNN, Chung et al., 2015); a variety of proposals addressing have been published in the ML literature (but for an in-depth and insightful analysis, see Girin et al., 2021.

As to the stochastic process perspective, the seminal work by Rezende et al. (2014) on stochastic backpropagation (gradient backpropagation through stochastic variables) was a milestone in making a connection between stochastic processes and deep neural network fields of investigation. They derived an algorithm that allows for joint optimisation of the parameters of both the generative and recognition models, such as Deep latent Gaussian models, which has fostered a growing interest for considering neural stochastic differential equations as deep latent models in the diffusion limit (e.g.,Tzen and Raginsky, 2019a,b; Kidger et al., 2021; Massaroli et al., 2020). These works analyze the diffusion limit of deep latent models, where the number of layers tends to infinity. The limiting latent object is an Ito diffusion process that solves a stochastic differential equation whose drift and diffusion coefficient are implemented by neural nets. A variational inference framework is adopted for these neural SDEs via stochastic automatic differentiation in Wiener space and where the computation of gradients is based on the theory of stochastic flows. This allows, in principle, the use of black-box SDE solvers and automatic differentiation for end-to-end inference.

Also, for what strictly concerns RSA, neural architectures have been considered for implementing neural listeners and speakers, markedly in the specific field of image captioning (e.g.,Andreas and Klein, 2016; Cohn-Gordon et al., 2018)

Clearly, all these neural-based approaches come into the game for scaling to large datasets. But, conversely, deep neural models are effective if large datasets are available. The transfer learning solution, which often comes without additional costs for research projects thanks to the availability of pretrained models, becomes rather problematic when the target domain might be highly specialized, inducing a large bias between the source and target domain. Indeed, generalizing well from limited data is a cogent issue in machine learning, however works aiming at directly learning from small datasets data are surprisingly scarce, and it is obviously beyond the scope of this thesis. These issue will be further explored along the simulations presented in Chapter 7.

CHAPTER 7

Simulations

The blossoming of modern machine learning and its widespread adoption that came about in the late 2000s relied and still rely on three structural prerequisites: the increasing availability of large quantities of high quality data, the increasing complexity of models, and the ever stronger support of both hardware and software computing power for training these models. Indeed, these conditions have allowed some remarkable advancements and applications that, though still in their infancy, are predicted to disruptively and irrevocably alter many industries for the better. However, the same conditions have also been pointed out as its bane. Gathering and cleaning data is costly, more so given that most machine learning problems are framed as supervised learning problems, requiring labeling of data, usually needing a human hand. The recent large successes of machine learning can undeniably be attributed to the advances in deep learning. Yet, deep learning models are data hungry and together with the issue of most problems being dealt with in a supervised learning framework provide two learning conditions that can hardly be reminiscent of learning conditions which a human may experience and thrive in. This in turn has rendered difficult reaching the same advancements in areas where data is more scarce, more costly or more difficult to acquire. The handicap of data availability is particularly evident in fields relating to affective computing. The datasets available are frequently of modest size and provide only a single, or few modalities of otherwise multimodal complex phenomena.

Faced with this plight, and having to deal with a theoretical model of imposing complexity considering an array of modalities, instead of developing implementations that only partially or incompletely adhere to the theoretical model, we shy away from implementation models that would require copious data. Instead, we proceed by considering models that, albeit rendered much more straightforward in certain regards by introducing appropriate approximations to the theoretical model, are nonetheless able to capture it in its totality, provide a working proof of concept, and ultimately reveal a remarkable complexity in their ability to simulate human behaviour. Namely, we develop three models that are herein presented: two models of non literal language use, hyperbole and irony, and a third model dealing with politeness, a form of social reasoning.

These conversational implicatures are typical case studies in pragmatics, irony has been known at least since Aristotle, with Grice being the first to systematically study cases in which what a speaker means differs from what the sentence used by the speaker means (Neale, 1992). The same were later to become the target of efforts in computational pragmatics. Recently, and most notably, the RSA framework has been successfully used to model hyperbole (Kao et al., 2014), irony (Kao and Goodman, 2015), and politeness (Yoon et al., 2020). Hyperbole and irony rely on communicating affect through exaggerated utterances and utterances with apparent meaning opposite in polarity to the intended meaning, respectively, while politeness involves intentions of conveying information falsely and out of care for the listener's feelings. However, the simulations of the same under the RSA model have not included a general model of core affect. Instead, the affective component was modelled in a more rudimentary manner, e.g. using table lookup (Kao et al., 2014) or experimental data (Kao and Goodman, 2015; Yoon et al., 2020), with much of the focus laid primarily on modelling pragmatics. In our case, having developed a general model of emotion as presented in the previous section, we integrate the two approaches for a more comprehensive and exhaustive treatment of these phenomena.

Eventually, a fourth simulation will be concerned with assessing the stochastic dynamics of the key component of the model, core affect. In this case, we will exploit a publicly available dataset.

The logical organization of the simulations that follows is outlined at a glance in Fig. 7.1.

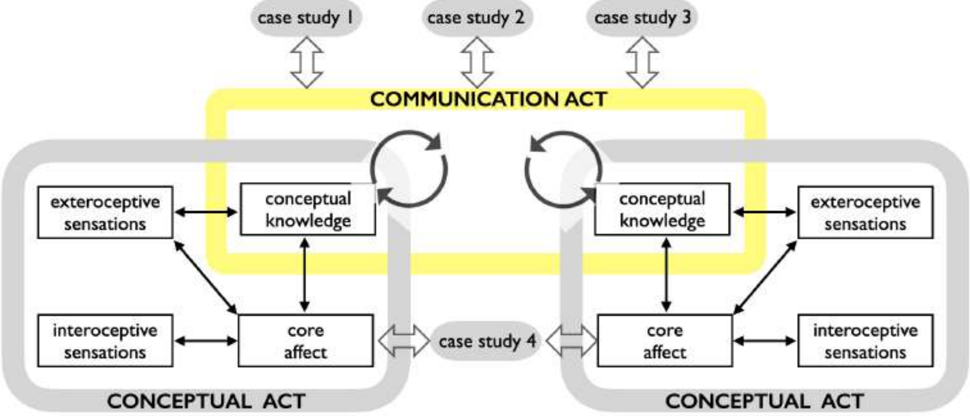


Figure 7.1: *The organization of the simulations (case studies 1, 2, 3 and 4) related to the in-context communication acts between agents as grounded in their conceptual acts*

7.1 Case study 1: Hyperbole

People do not always mean what they say. Imagine meeting a friend in the city on a warm summer day, and as they arrive they say: *"It's a million degrees today!"*. Their utterance is not likely to be interpreted literally, i.e. it's unlikely that the listener would actually believe the air temperature to be a million degrees. The speaker is obviously referring to the weather being unusually and overwhelmingly hot, and are exaggerating to express their annoyance with it. Such non-literal interpretation of exaggerated utterances intended to convey affect is called hyperbole.

The RSA model can be modified in such a way to account for hyperbole. We proceed by first introducing this model, and then altering it so that it includes a richer communicative context with an additional channel of communication revealing the speaker's affect.

Implementation details The context of our example is that of a listener waiting on the speaker that went to buy a bottle of wine. As they approach they utter to the listener: *"It cost me s dollars!"* with s being its price. The belief state space in this example just refers to the range of actual costs of a bottle of wine

$$\mathcal{B} = \{B_s\}_{s_0}^{S}$$

and it relies on the joint concept state $\mathcal{C} = \{C_{bottle}, C_{wine}, C_{price}\}$ of bottle, wine, and price. In the simulation, it is an increasing series of integers between $s_0 = 50$ and a $S = 10000$.

In virtue of previous experience on affordable prices of wine bottles, the listener has at time t a prior belief which we denote $P(B_s) = P(s)$. In what follows we drop the time index t when not explicitly needed. The listener assumes that such beliefs are common ground with the speaker.

At a single level of recursion to which we'll limit this example, the pragmatic listener reasons about a pragmatic speaker reasoning about a literal listener as per the original RSA model, but we introduce two modifications. Other than wanting to communicate an objective state of the world B_s, the speaker might have a different communicative goal, that of communicating his affect state $F = V \times A$. In this example this space is restricted only to arousal values $a \in A$. Under such circumstances, the interpretation space for the utterance has an additional dimension. The RSA model we introduced in the previous chapter is thus extended with a $P(a \mid s)$ with $a \in A$, a prior on affect given a world state, and a goal function $g_t : \mathcal{M} \to M_X$ projecting the speaker's communicative goal from the full space of meanings $\mathcal{M} = S \times A$ to the appropriate subspace M_X.

To keep things simple, we assume the following:

- Words are sampled one-to-one according to agent's beliefs state concerning the prices of wine bottles:

$$P(w) \approx P(w \mid \mathcal{L}) \sum_{\mathcal{C} \neq \mathcal{C}^{price}} P(\mathcal{C} \mid \mathcal{B}, \mathcal{L}) = \delta_{w=s},$$

  with words denoting prices $w \in \{\text{“}\texttt{fifty}\text{''}, \text{“}\texttt{fifty one}\text{''}, \ldots, \text{“}\texttt{ten thousand}\text{''}\}$ and we are not considering a role for the language model LM at this level

- the external action plan concerning the utterance action, $\mathcal{A}^{utt} \sim P(\mathcal{A}^e \mid w, \mathcal{G}, \mathcal{B})$, to be sampled at this point simply is the one necessary to operationalize the decision of uttering a specific word under a given goal; the external motor execution, since we are not dealing with speech production issues, is assumed to precisely provide the exact corresponding utterance,

$$P(\mathcal{X}^{utt} \mid R^e, \mathcal{A}^{utt}) = \delta_{u=w},$$

such that $u \in \mathcal{UT} = \mathcal{X}^{utt} = \{\text{“}\texttt{fifty}\text{''}, \ldots, \text{“}\texttt{ten thousand}\text{''}\}$

- the listener has perfect perception of the uttered sound in the environmental context, $u = Z^{utt}(\mathcal{O}(\mathcal{X}^{utt}))$

Thus, given the above steps the execution of the plan for uttering u boils down to the following:

$$\mathcal{A}^{utt} \sim P(\mathcal{A}^e \mid w, \mathcal{G}, \mathcal{B})$$
$$u \sim P(\mathcal{X}^{utt} \mid R^e, \mathcal{A}^{utt})$$

Let us now consider the literal listener's perspective as simulated by the pragmatic speaker. The literal listener, endowed with a prior on world states and the speaker's affect, by way of the meaning function $[[u]](s)$ interprets the utterance without taking into account the speaker's communicative goals:

$$P_{Lit}(s, a \mid u) \propto \delta_{[[u]](s)} P(a \mid s) P(s) \tag{7.1}$$

The pragmatic speaker then maximizes the epistemic utility considering the literal speaker's inferred meaning, while minimizing their utterance cost $cost(u)$:

$$\begin{aligned} P_S(u \mid s, a, g) &\propto \exp \alpha U_S(u \mid s, a, g) \\ U_S(u \mid s, a, g) &= \log P_{Lit}(g(s, a) \mid u) - cost(u) \end{aligned} \tag{7.2}$$

The pragmatic listener L_1 finally performs Bayesian inference by simulating the pragmatic speaker S_1 and marginalizing over the intended communicative goal g:

$$P_{L_1}(s, a \mid u) \propto P_{S_1}(u \mid s, a, g) P(a \mid s) P(s) P(g). \tag{7.3}$$

We assume $A = \{0, 1\}$ to represent at a very coarse grain (low/high) the affective dimension of arousal. Furthermore, the prior over prices $P(s)$ as a distribution is decreasing, with most probability mass lying at lowest prices, and very high prices being far less likely. The prior of the arousal being high conditioned on the state $P(a = 1 \mid s = \tilde{s})$ is the reverse, and increases as the communicated price increases. The speaker may choose to communicate the price s, his arousal a consequent on the price paid, or both s, a, with the prior over the three goals $P(g)$ being uniform. We ignore the utterance cost term $cost(u)$.

The above concerns the basic model of hyperbole. We now consider the full complex communicative context where the speaker's utterance is accompanied by a facial expression. In this case, the speaker beyond the conscious decision of making the external utterance action $\mathcal{A}^{utt}$, consciously or unconsciously “decides” for a non verbal external action $\mathcal{A}^f$, a facial expression, that will generate an appropriate motor plan of facial muscles, $R^f \sim P(R^f \mid \mathcal{A}^f)$, giving rise to an executed facial expression action $X^{face} \sim P(\mathcal{X}^f \mid R^f)$. Thus, a specific facial expression, say $\mathcal{X}^f$ will be produced.

In addition, now the listener, under the assumption that she holds the same, common ground affective perspective of the listener (prior on F), will take advantage of the fact that she actually observes the non verbal behaviour of the speaker, in particular his facial expression.

The exteroceptive unimodal representation Y^{NV} is thus obtained through the backward inference $Y^{NV} \leftarrow \mathcal{O}^{NV}(X^{face})$ and giving rise to an actual exteroceptive instantiation of the RV Y^{NV} in terms of the set of facial action units AU:

$$Y^{NV} = AU$$

In the experimental setup, the facial expression represented via the AU set is inferred from an actual image depicting a face that exhibits one of the standard expressions that are considered to be related to basic categorical emotions in the affective computing literature.

To sum up, the listener, other than using the prior on arousal and the utterance for inference over the speaker's intended meaning has additional information on the speaker's affective state relayed through the facial expression.

The listener exploits the joint distribution over facial expressions and affect $P(AU, a_f)$ with a_f denoting the arousal (here subscripted with f to distinguish it from the prior), and upon observing the the speaker's expression $\tilde{AU}$, uses her internal generative model to infer a distribution of the speaker's affect $P(a_f \mid AU = \tilde{AU})$. To include the inferred affect a_f in their reasoning about the speaker's intended meaning, the listener performs weak cue integration over the conditional prior on arousal, and the one inferred from the observed facial expression to arrive at the integrated arousal $\bar{a}$:

$$P(\bar{a} \mid s, AU) = P(a \mid s)P(a_f \mid AU) \tag{7.4}$$

The only piece of the RSA inferential machinery that needs adjustment is the pragmatic listener's in eq. 7.3:

$$P_{L_1}(s, a \mid u) \propto P_{S_1}(u \mid s, \bar{a}, g)P(\bar{a} \mid s, f)P(s)P(g) \tag{7.5}$$

whereby the pragmatic listener simulates the pragmatic speaker by sampling from $P(\bar{a} \mid s, AU)$ instead of $P(a \mid s)$. Note that the literal listener's behaviour is unaltered and sampled from the conditional prior on arousal.

As per the theoretical model, the surmised affective state underlying the speaker's beliefs cascades to the unfolding of an internal action plan which generates a physiological regulation of the body, and, most importantly, an interoceptive prediction to make meaning of the outcomes of the world, $\mathrm{Y}^i_n \sim P(Y^i_n \mid Z^i_t, F_{t+1})$. The listener, grounding in the same conceptual act structure, even in the absence of a concrete measurement of the speaker physiological status, can rely on her own simulation.

Thus, the affect inferred from the facial expression sets off an internal simulation loop where physiological signals are simulated in the listener's system, and these in turn end up reinforcing the listener's estimate of arousal. Let $P(Y^i_{1:N}, a_p)$ be the joint distribution of N physiological signals Y^i and arousal a_p (subscripted for distinction). The listener performs the following internal simulation to arrive at their estimate of arousal, which in brief for this example boils down to the following steps:

1. Sample the arousal inferred from the facial expression:

$$\tilde{a_f} \sim P(a_f \mid AU = \tilde{AU})$$

2. Sample the physiological signals conditioned on $\tilde{a_f}$:

$$\tilde{Y}^i_{1:N} \sim P(Y^i_{1:N} \mid \tilde{a_f})$$

3. Use $\tilde{Y}^i_{1:N}$ to get $P(a_p \mid \tilde{Y}^i_{1:N})$

The distribution from step three is then incorporated as a cue in Eq. 7.4 which then becomes:

$$P(\bar{a} \mid s, AU, \tilde{Y}^i_{1:N}) = P(a \mid s)P(a_f \mid AU)P(a_p \mid \tilde{Y}^i_{1:N}) \tag{7.6}$$

And the pragmatic listener's posterior instead becomes the following:

$$P_{L_1}(s, a \mid u) \propto S_1(u \mid s, \bar{a}, g)P(\bar{a} \mid s, AU, \tilde{Y}^i_{1:N})P(s)P(g) \tag{7.7}$$

We implement both $P(AU, a_f)$ and $P(\tilde{Y}^i_{1:N}, a_p)$ with a supervised probabilistic principal component analysis model (sPPCA) (Yu et al., 2006) with Markov chain Monte Carlo inference using the Pyro probabilistic programming language (Bingham et al., 2019) in Python. The data used for training the two models was artificially generated based on experimental relationships of the phenomena involved. As to the physiological signals we use electrodermal activity (EDA), since it correlates well to the state of arousal.

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

$$b(t) = c\left(e^{\frac{-t}{\tau_1}} - e^{\frac{-t}{\tau_2}}\right) \tag{7.8}$$

where τ_1 measures the steepness of rise and τ_2 its decay, while c is a constant term for the gain. The deconvolution of original EDA signal with the IRF above permits to extract the SCR components.

We followed the work of (Kreibig, 2010) to form the relationships between autonomic nervous system activity parameters and the relevant affective dimensions. We used the number of EDA peaks as a relevant identifier of the level of arousal, and the two together form the inputs of the sPPCA model. The simulations of EDA are performed using the neurokit2 package (Makowski et al., 2021). The relationship between the facial expression and affective dimensions used for generating the data was formed based on the facial action coding system (Ekman and Friesen, 1978), with action units and affective dimensions forming the two inputs of the model. Exact Bayesian inference was used on the RSA models.

Results We first examine the results of the base model that are presented in figures 7.2 to 7.4. In fig. 7.2 the speaker utters 10000 as the price and the model correctly infers that the speaker's communicative intention was that of transmitting arousal, and not the actual price. There is substantial probability mass on prices 50 and 51 with the communicative goal of arousal. This is due to the prior on prices that is naturally skewed towards lower prices. In contrast, a more certain posterior is seen in the case of the utterance 500 in fig. 7.3 where the model again correctly infers the correct communicative intention. The same is true of all other prices in between the two not shown here. Finally, in fig. 7.4, where the speakers utters 50 the inference is again correct with the model inferring that the communicative intention was to communicate the price, and places a low probability mass over other prices.

These results show that in principle the RSA model is capable of modelling hyperbole. We now turn the modified model where other than the utterance, the listener perceives a facial expression communicating arousal, which they then utilize for simulating physiological signals that in turn reinforce that arousal. The resulting arousal value is then integrated as a cue with the prior, resulting in more intricate inferences of the posterior.

In fig. 7.5 the listener hears 10000 and observes an angry facial expression. The listener using their internal model infers a high probability of high arousal and the physiological simulation reinforces it. Based on the utterance and on the final probability of high arousal after cue integration, the model infers as in the base case that the speaker's communicative intention was that of communicating arousal. Instead, the same utterance but with a sad expression is represented in fig. 7.6. With sadness being marked by low arousal the model interprets the speaker's communicative intention as that of communicating price. This is a reasonable and expected result, where an emotional expression made a statement with a low prior be much more likely. In other words, it is the emotional expression, or rather the communicated emotional state of the speaker that made an otherwise improbable inference probable.

The same is conclusion is evident in figures 7.7 and 7.8 that represent the posterior for the utterance "500" with a happy facial expression and a sad expression, respectively. A happy facial expression is commonly associated with a high value of arousal, and in this case the inference is correct that the speaker wanted to communicate their annoyance with the unreasonably high prices, instead of the literal meaning. Instead, that same utterance accompanied by a sad facial expression turn the rather unlikely communicative goal of the price actually being that high the most probable interpretation.

Finally, for a reasonable price of 51 the result for the expression *anger* is depicted in fig. 7.9. In this case, the strong prior on the price actually being that amount prevents the model from inferring the speaker's utterance as hyperbole, despite the expression high in arousal.

The base case of hyperbole works because of two of its fundamental constituents: the recursive Bayesian reasoning of the RSA model, and the priors over the prices and arousal. The modifications introduced herein and the results thereof further underline their importance, but also reveal the additional complexity an emotional dimension to an exchange brings.

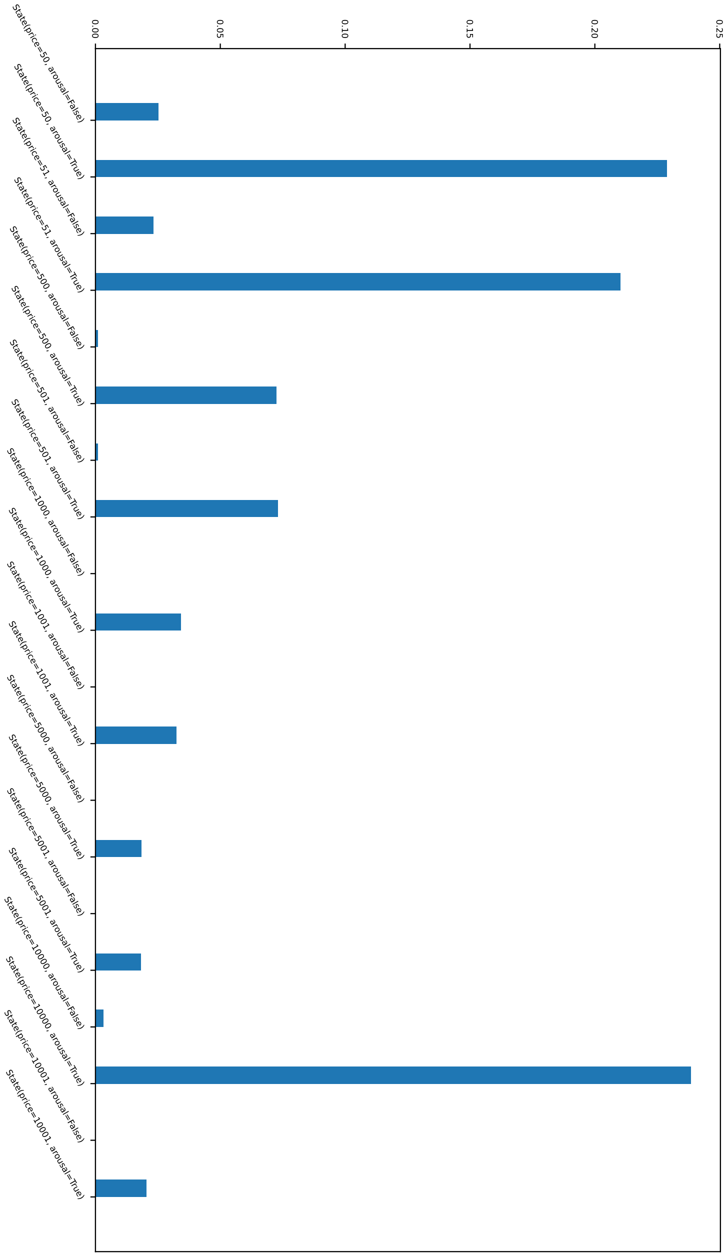


Figure 7.2: *The posterior probability over all states for the hyperbole base model for the utterance "10000".*

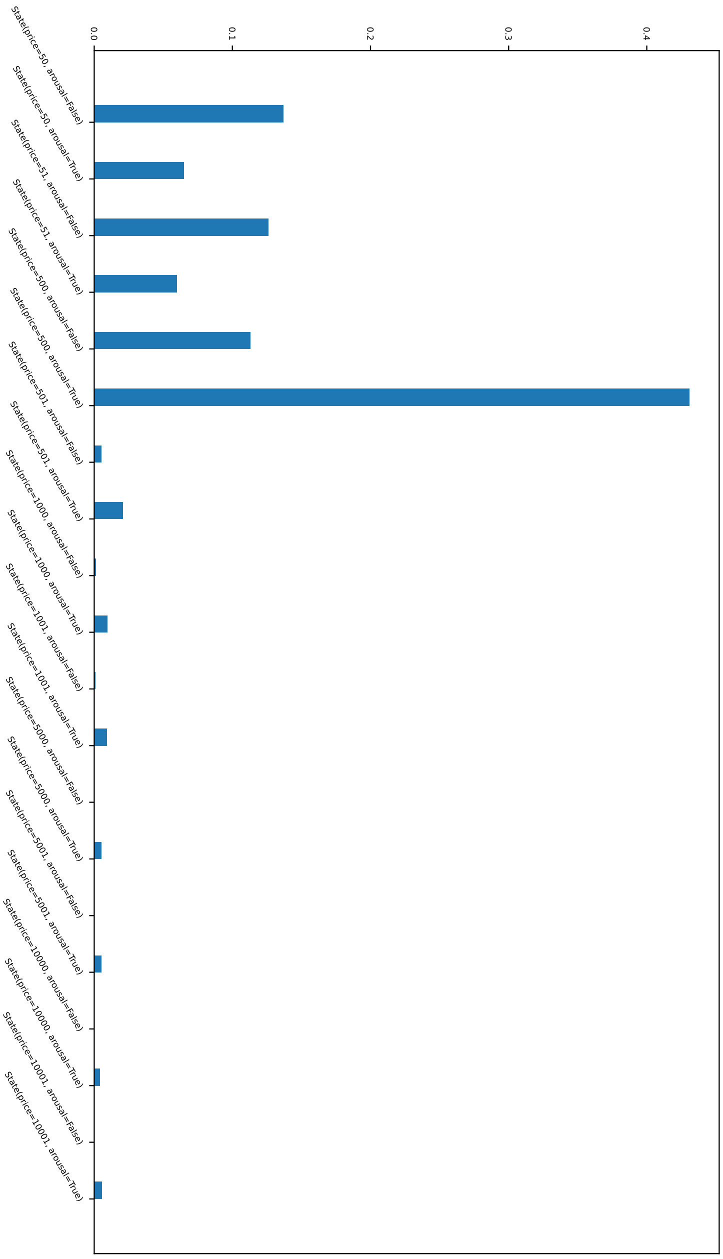


Figure 7.3: *The posterior probability over all states for the hyperbole base model for the utterance "500".*

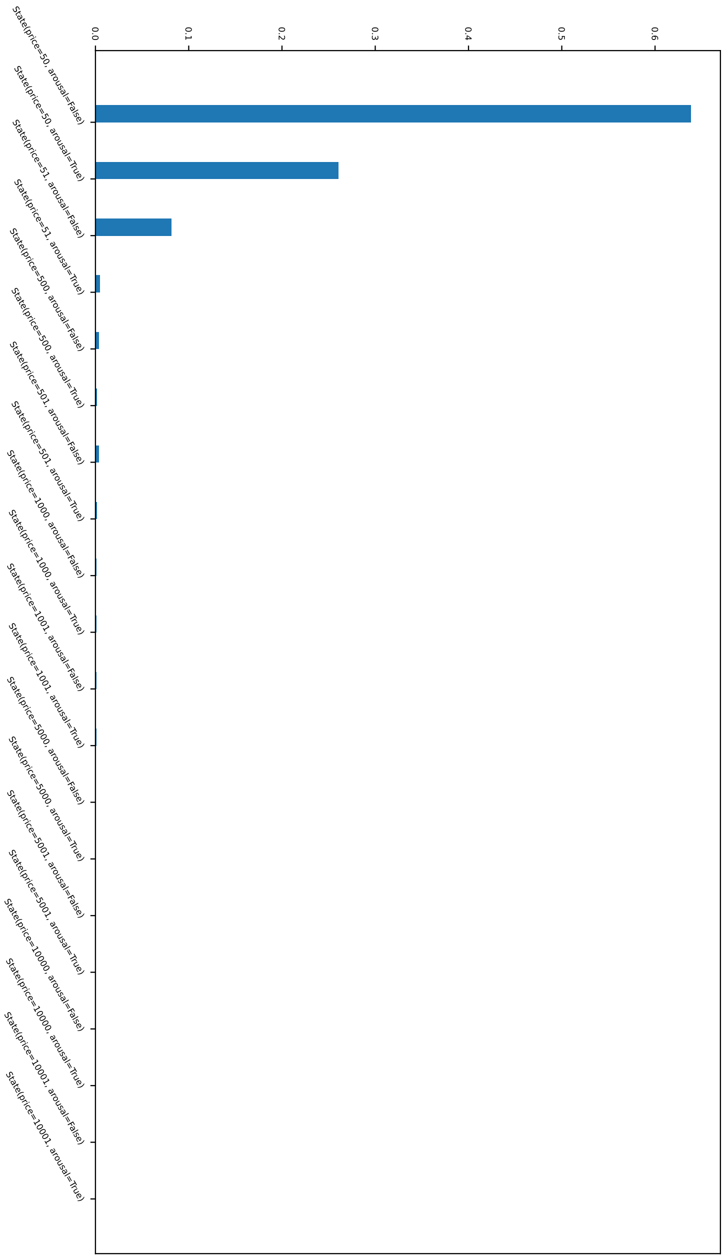


Figure 7.4: *The posterior probability over all states for the hyperbole base model for the utterance "50".*

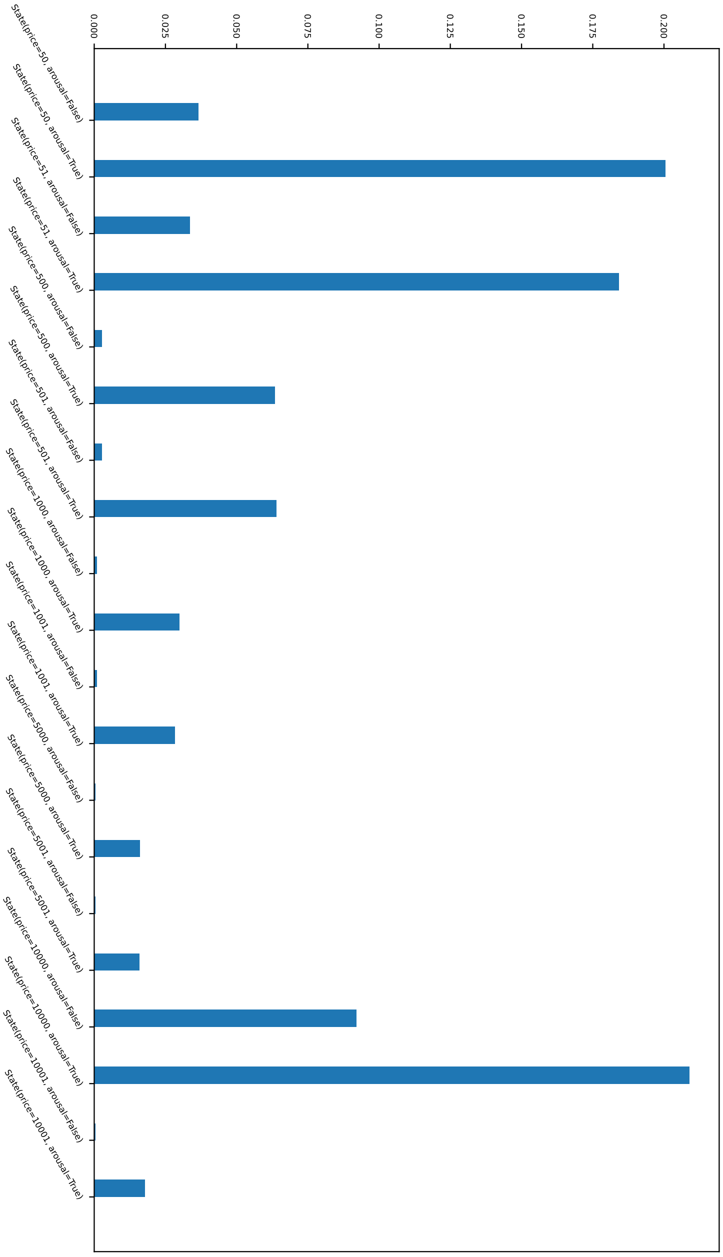


Figure 7.5: *The posterior probability over all states for the hyperbole base model for the utterance "10000" and expression anger.*

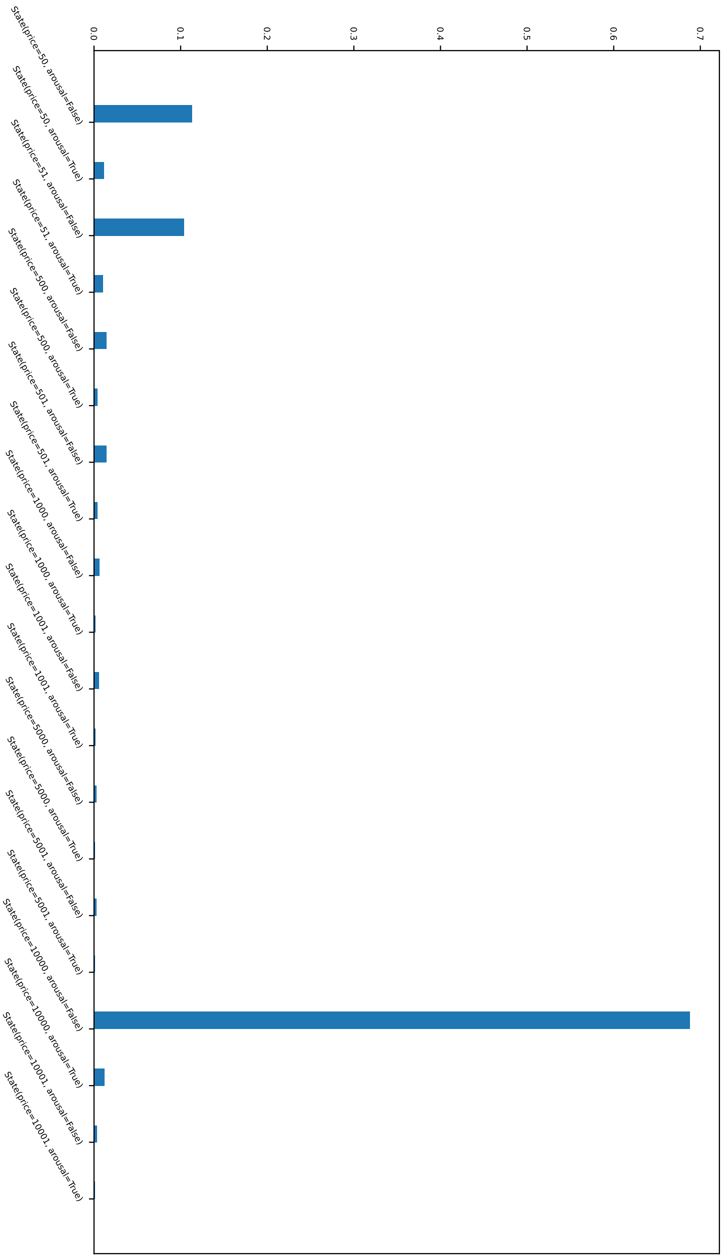


Figure 7.6: *The posterior probability over all states for the hyperbole base model for the utterance "10000" and expression sad.*

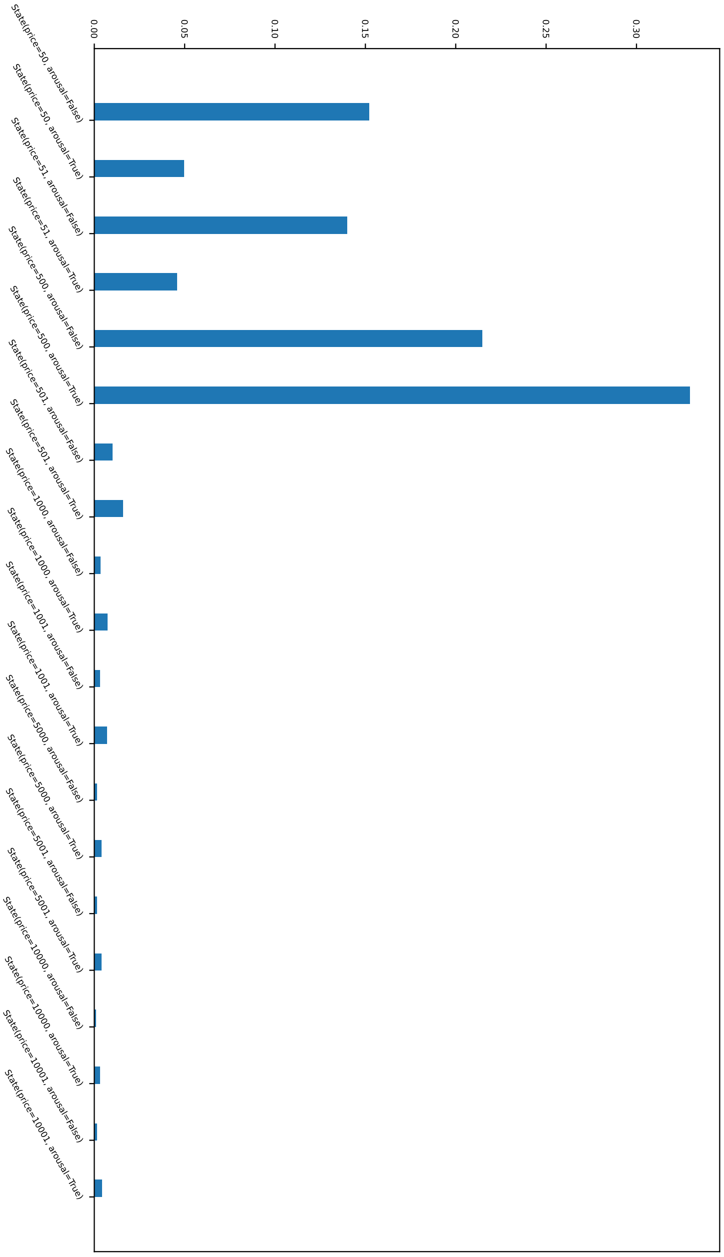


Figure 7.7: *The posterior probability over all states for the hyperbole base model for the utterance "500" and expression happy.*

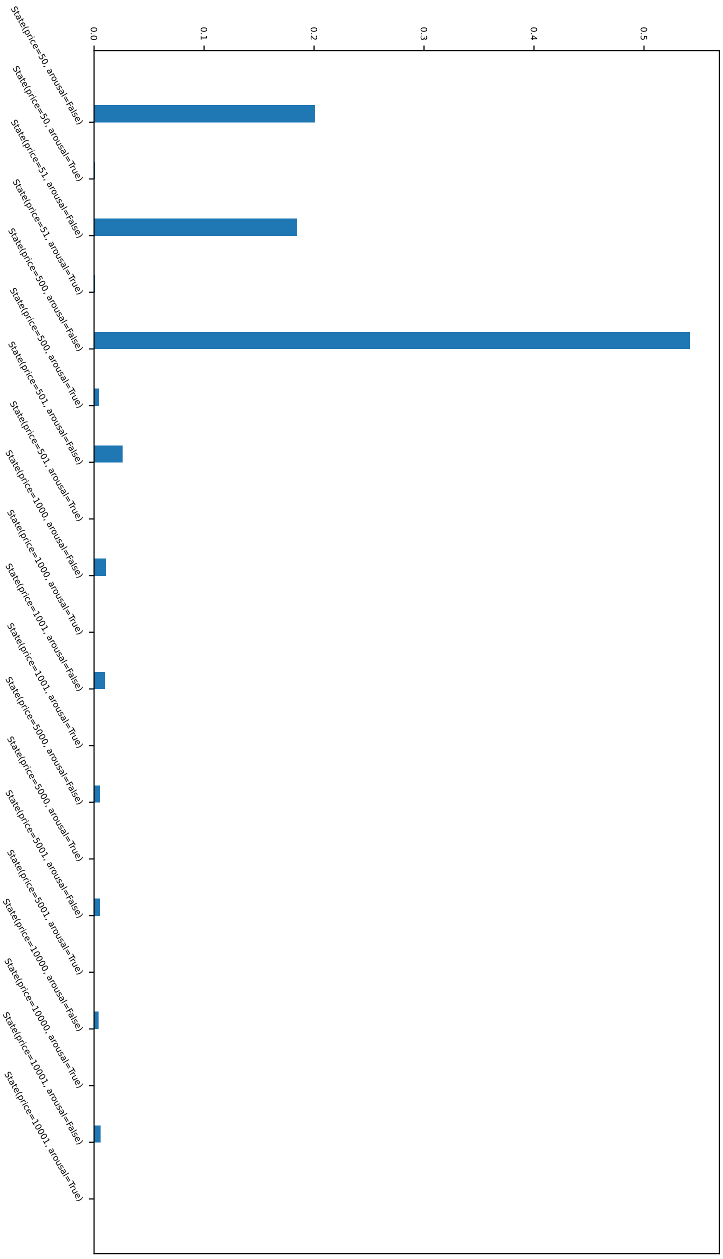


Figure 7.8: *The posterior probability over all states for the hyperbole base model for the utterance "500" and expression sad.*

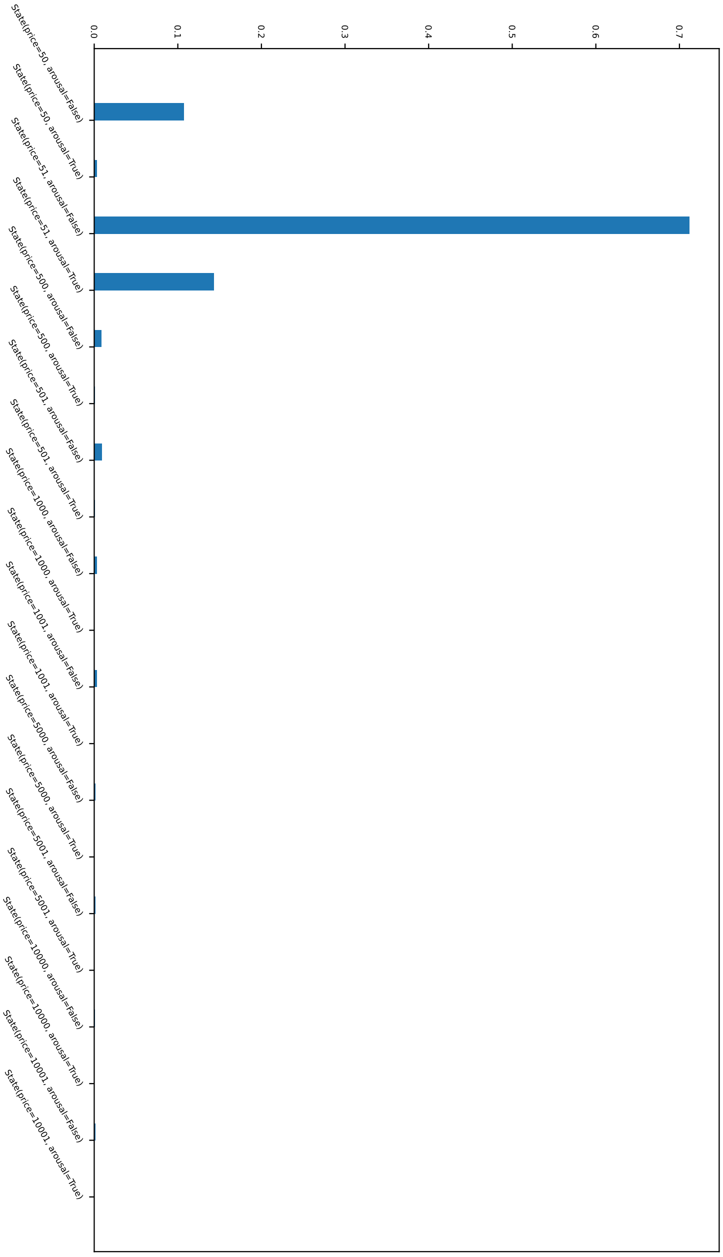


Figure 7.9: *The posterior probability over all states for the hyperbole base model for the utterance "51" and expression anger.*

7.2 Case study 2: Irony

In a non-ironic context, a speaker may want to communicate negative affect about a situation (e.g. unhappiness about the cool temperature outside) instead of the precise situation (the temperature outside), in which case choosing an exaggerated utterance (*"It's freezing outside!"*) effectively communicates negative affect and addresses the question under discussion (QUD) or topic of conversation. A listener who reasons about the speaker and QUD is then able to use his background knowledge to correctly infer that the speaker is upset about the temperature. Irony, instead, is defined as utterances whose apparent meanings are opposite in polarity to the speaker's intended meaning. In particular, since verbal irony involves expressing negative meanings with positive utterances and vice versa, it calls for a rich space of affect that includes both positive and negative affect states. For example, uttering *"Oh, what lovely weather, let's go the beach!"* while presented with below freezing atmospheric conditions would classify as ironic.

We first present a case of how irony can be modelled in general by RSA, and then proceed to extend the model to include inference on part of the speaker about both the context in which the interaction is happening and its implications on the affect the speaker might have experienced, and also similarly to the case of hyperbole, an additional channel of communication through the listener observing the speaker's facial expression.

Implementation details The context of our example is that Alice (the listener) and Bob (the speaker) are on a beautiful beach on a sunny day. A smiling Bob utters: *"Terrible!"*. What should Alice infer from the utterance and the context? What was Bob's communicative intention?

We assume a modelling perspective where the speaker's intention is effectively conveyed in a single time instant, so we disregard the dynamics of the theoretical model. First, assume at time t a discrete set of states corresponding to simple belief states of weather states as conceptualized/inferred from viewed scenes (actual or captured by photographs). This is the result of a belief update due to an observation action over the external world $\mathcal{A}_{t-1}$; i.e., $\mathcal{B}_t \sim P(\mathcal{B}_t \mid \mathcal{B}_{t-1}, \mathcal{A}_{t-1})$, such that

$$\mathcal{B} = \{B_{s_1}, B_{s_2}, B_{s_3}\} = \{\texttt{terrible}, \texttt{ok}, \texttt{amazing}\},$$

where, for notational convenience we have dropped the time-index t, which in what follows we will use only when needed.

Denote for simplicity $P(s)$, $s \in \{s_1, s_2, s_3\}$ the prior probability over belief states which can be instantiated in a given context. For instance, in a beach location, whether in California or Sicily, higher prior probability may be placed on `amazing` and `ok` than on `terrible`.

Clearly, in the given situation, which is observed by both Bob and Alice, if Bob (the speaker) chooses to communicate `amazing`, he is just providing a plain, honest description of the current state of affairs; differently, if Bob conveys `terrible`, he might be making an ironic statement, or just lying, etc.

Performing the most probable inference is indeed the task of Alice (the listener), who will use all the information available, including the possible affective states of the speaker.

At the mental level the core affect state $F = V \times A$ provides a discretized/clustered representation of the low-level continuous core-affect space, reflecting the affective granularity of the subject. In this example, to simplify, we shall consider a rather coarse-grained, binary representation of F:

$$v \in V = \{-1, 1\}$$

$$a \in A = \{\texttt{low}, \texttt{high}\}$$

Under such circumstances, priors over binary RV (V, A), conditionally on state s, might be simply written,

$$P_V(v \mid s) = Bern(v \mid \pi^V(s_i))$$

$$P_A(a \mid s) = Bern(a \mid \pi^A(s_i))$$

where $Bern()$ denotes the Bernoulli distribution, and

$$(\pi^V(s_i), \pi^A(s_i)) = P(F \mid \mathcal{B} = s_i, \mathcal{L})$$

Turning back to the speaker, in order to perform the speech action A_1, Bob has to choose a realizable communication goal $g \in G \subset \mathcal{D}$, among his possible desires $\mathcal{D}$ (in the BDI agent jargon). Formally, the goal represents the projection from the full meaning space to the subset of interest to the speaker. The result might be any possible subset M_X of states and affect. For instance,

$$g(s, v, a) = \begin{cases} g_s(s, v, a) = s, \\ g_v(s, v, a) = v, \\ g_a(s, v, a) = a. \end{cases}$$

The prior distribution on goals is chosen as a categorical distribution, whose parameters π_i^G represent the probability of sampling goal $g_i \sim Cat(g_i \mid \pi_i^G, i = s, v, a)$, where the simplest case is represented by all π_i^G being equal (uniform sampling).

Much like to the hyperbole simulation, here we assume the following:

- Words are sampled one-to-one according to agent's beliefs state concerning the weather situation

  $$P(w) \approx P(w \mid \mathcal{L}) \sum_{\mathcal{C} \neq \mathcal{C}^{weather}} P(\mathcal{C} \mid \mathcal{B}, \mathcal{L}) = \delta_{w=s},$$

  where $w \in \{\text{“}\texttt{terrible}\text{''}, \text{“}\texttt{ok}\text{''}, \text{“}\texttt{amazing}\text{''}\}$ and we are not considering a role for the language model LM at this level

- the external action plan $\mathcal{A}^{utt} \sim P(\mathcal{A}^e \mid w_{t+1}, \mathcal{G}, \mathcal{B})$, to be sampled at this point, is simply the one necessary to operationalize the decision of uttering a specific word under a given goal; the external motor execution, since we are not concerned with speech production issues, is assumed to precisely provide the exact corresponding utterance,

  $$P(\mathcal{X}^{utt} \mid R^e, \mathcal{A}^{utt}) = \delta_{u=w},$$

  such that $u \in \mathcal{UT} = \mathcal{X}^{utt} = \{\text{“}\texttt{terrible}\text{''}, \text{“}\texttt{ok}\text{''}, \text{“}\texttt{amazing}\text{''}\}$

- the listener has perfect perception of the uttered sound in the environmental context, $u = Z^{utt}(\mathcal{O}(\mathcal{X}^{utt})$

In the mind of the actual listener (Alice), the communicative action plan $\mathcal{A}^{utt}$ of pragmatic Bob, under the assumptions stated above, unfolds as follows. The speaker addresses a literal listener who represents an idealized, simple minded listener. The literal listener L_0 provides the base case for recursive social reasoning between the speaker and listener. L_0 interprets u literally without taking into account the speaker's communicative goals. Define the literal interpretation function $[[u]] : S \to Bool = \{0, 1\}$:

$$[[u]](s) = \delta_{u=s}$$

The L_0 inferential process, given the utterance u and communicative goal g, relies upon the posterior

$$P_{Lit}(s, v, a \mid u, g) \propto \delta_{u=s} P_V(v \mid s) P_A(a \mid s) P_S(s)$$

The pragmatic speaker S_1, by generating an utterance u, aims at being informative, as in Gricean theories of communication, but only with respect to a particular goal g or topic, thus realizing a kind of relevance principle. This relevance is critical for deriving non-literal interpretations.

To such end S_1 relies on the inference of the literal listener's posterior $P_{Lit}(s, v, a \mid u, g)$, to generate the most convenient utterance u under the given goal g via the posterior:

$$P_{S_1}(u \mid s, v, a, g) \propto P_{Lit}(s, v, a \mid u, g) P(u)$$

Recall that in RSA, S_1 chooses utterances according to a softmax decision rule that describes an approximately rational planner

$$P_{S_1}(u \mid s, v, a, g) \propto e^{\alpha U_1(u|s,v,a,g)}$$

where the parameter α tunes the speaker's rational choice. The larger α is, the more the speaker's choice probabilities converge to a strict maximization of utility. The two equations are equivalent in the case f $\alpha = 1$.

At the top layer of inference, the pragmatic listener, L_1, interprets utterance u to update prior beliefs on meaning $P(s)$ by taking into account how likely the speaker would have been to produce the observed utterance u in various states and goals. To such end L_1 arrives at the following posterior:

$$P_{L_1}(s, v, a, g \mid u) \propto P_{S_1}(u \mid s, v, a, g) P_G(g) P_V(v \mid s) P_A(a \mid s) P_S(s) \tag{7.9}$$

We now complete the baseline irony inference with non-verbal cues. In this case, the pragmatic listener is able to perceive the speaker's facial expression which, other than through the prior, provides an additional clue as to the affective state of the speaker. Again, the speaker and listener share an external stimulus related to the context. They see a wonderful picturesque beach in front of them on a sunny day. This stimulus provides a category with its own affective meaning, to both listener and speaker, that the listener exploits for setting the background mental simulation of the speaker.

More formally, prior beliefs at current time are based on the conditional posterior at previous time and are conditionally formed based on concepts available after the action, $P(\mathcal{B}_{t+1} \mid \mathcal{C}_t, \mathcal{L}_{t+1})$.

Concepts are the formal result of the conceptualization act, which is summarized by the conditional posterior distribution $P(\mathcal{C}_t \mid \mathcal{O}_t)$. In this example we consider a simple set of concepts states $\mathcal{C} \in \mathcal{L} \times F$, where each concept is a joint state of a world label and a valence/arousal, state concerning the external context, e.g,

$$\mathcal{C} = \{C_{\texttt{beach}}, C_{\texttt{sunny}}\}$$

with $C_{\texttt{beach}} = (\text{“}\texttt{beach}'', v_{\texttt{beach}}, a_{\texttt{beach}})$, $C_{\texttt{sunny}''} = (\text{“}\texttt{sunny}'', v_{\texttt{sunny}}, a_{\texttt{sunny}})$, with “$\texttt{beach}''$, “$\texttt{sunny}'' \in \mathcal{L}$. In brief, beliefs are the inferential results due to the mapping $\mathcal{O}^{env} \to \mathcal{B}$ that both interlocutors experience and share over the common environmental outcomes $\mathcal{O}^{env}$. Belief states B_{s_i} are in parallel supported by conceptual states $\mathcal{C}^{weather} = \{C_{s_i}\}$

In this simulation, differently from the previous one, we actually rely on the backward inference $P(\mathcal{B}_t, \mathcal{C}_t \mid \mathcal{O}_t^{env}, \mathcal{O}_t^{S_{NV}})$, where the environmental outcomes $\mathcal{O}_t^{env}$ are presented to the agent's in the form of static images representing beaches in a sunny day.

More concretely, the implementation of the backward inference in the case of the external stimulus related to the context is through an image captioning model. When presented with an image, an image captioning model provides a verbal description of image contents. In our case, for simplicity's sake we keep only lemmatized nouns of the generated description. In our example these words turn out to be `beach` and `sunny`. These are looked up in the lexicon $\mathcal{L}$ containing English words and their associated affective values based on experimental data. The three values, that is, the word, valence and arousal, each form a concept (e.g. $C_{\texttt{beach}}$), and the set of all such concepts is the aforementioned $\mathcal{C}$. The lexicon used is the NRC lexicon (Mohammad, 2018), and the image captioning model used is a generic publicly available pretrained model.

Consequently, the probability of the overall affective state of the agents at this point can be obtained by marginalizing with respect to the available lexicon

$$P(F \mid \mathcal{B}, \mathcal{L}) = \sum_{\mathcal{L}} P(\mathcal{C} \mid \mathcal{B}, \mathcal{L}).$$

In addition, now the listener, under the assumption that she holds the same, common ground affective perspective of the listener (prior on F), will take advantage of the fact that she actually observes the non verbal behaviour of the speaker, in particular his facial expression. As in this example we take into account both the affective dimensions, we denote as $P(AU, F) = P(AU, v, a)$ the joint distribution of core affect F and the perceived facial expression, with which the listener is equipped. The implementation model of $P(AU, F)$ and the data used to train the model are the same as in the case of hyperbole.

The contextual cues give rise the internal simulation loop of physiological states, similar to the case of hyperbole, that evolves according to the following sampling steps in the case of arousal:

1. Sample the arousal inferred from the observed context:

$$\tilde{a_c} \sim P(a_c \mid \mathcal{O}^{env} = \tilde{\mathcal{O}^{env}})$$

2. Sample the physiological signals conditioned on $\tilde{a_c}$:

$$\tilde{Y}^i_{1:N} \sim P(Y^i_{1:N} \mid \tilde{a_c})$$

3. Use $\tilde{Y}^i_{1:N}$ to get $P(a_c \mid \tilde{Y}^i_{1:N})$

and analogously for valence. The contextual cues, the non-verbal cues and the physiological cues are integrated according to the following expressions:

$$P(\bar{a} \mid \mathcal{O}^{env}, AU, \tilde{Y}^i_{1:N}) = P(a_c \mid \mathcal{O}^{env})P(a_f \mid AU)P(a_p \mid \tilde{Y}^i_{1:N})P_A(a \mid s) \quad (7.10)$$

$$P(\bar{v} \mid \mathcal{O}^{env}, AU, \tilde{Y}^i_{1:N}) = P(v_c \mid \mathcal{O}^{env})P(v_f \mid AU)P(v_p \mid \tilde{Y}^i_{1:N})P_V(v \mid s) \quad (7.11)$$

The pragmatic listener's posterior then becomes the following:

$$P_{L_1}(s, v, a, g \mid u) \propto P_{S_1}(u \mid s, \bar{v}, \bar{a}, g)P_G(g)P(\bar{a} \mid \mathcal{O}^{env}, AU, \tilde{Y}^i_{1:N}) \quad (7.12)$$

$$P(\bar{v} \mid \mathcal{O}^{env}, AU, \tilde{Y}^i_{1:N})P_S(s) \quad (7.13)$$

For what concerns the arousal estimate from physiological signal, we again rely on EDA.

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

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

At the end of the aforementioned basic steps, the amplitude of the sample is computed as the inverse of the time difference between consecutive R peaks and is placed at the instant of the second R peak. In other experimental settings, the recorded cardiographic signal is the result of a blood volume pulse (BVP) sensor placed on a finger. This is a non invasive mean to obtain an indirect measure of the heart rate, via the arterial oxygen saturation of hemoglobin. This information, obtained via photoplethysmography (PPG), consists in sending two lights at different wavelengths, namely $660nm$ (red light spectrum) and $940nm$ (infrared light spectrum). The first wavelength is absorbed by deoxyhemoglobin (Hb) and the second by hemoglobin (HbO_2), which together affect the blood stream. By considering the absorption levels it is possible to calculate the heart rate with alternating vasodilatation and vasoconstriction. This signal is strictly correlated with the heart rate measured via ECG and increases in presence of pleasant stimuli Selvaraj et al. (2008). The value provided by PPG, indeed, corresponds to the so called “instantaneous heart-rate” (Electrophysiology, 1996), namely the number of times the heart would beat in one minute if the duration of successive cardiac cycles were constant. Heart rate values are provided as the number of contractions of the heart per minute (BPM) and directly relates to the RR distance, by following the basic formula from the textbooks of physiology and medicine (Braunwald et al., 1998; Hall, 2010): $RR[ms] = 60[sec] * 1000/BPM$.

Again, as in the arousal/EDA case, a synthetic simulation is performed such that the agent has available the likelihood of the interoceptive cue about the current cardiac status.

Results As in the previous example, we first examine the workings of the base model, then proceed with a discussion of the modified model. Finally, we change the state prior to see how it influences the posterior distribution.

The prior on the states is such that it is 50 times more likely to observe *"ok"* weather and *"amazing"* weather than *"terrible"* weather. Figures 7.10, 7.11, 7.12 depict the posterior for the three possible utterances respectively: *"amazing"*, *"ok"* and *"terrible"*. For the utterance *amazing* the pragmatic listener correctly infers the state and that the speaker's communicative intention was to communicate the state with a high valence and high arousal. There is however some probability mass on the same state but with the goals of communicating their valence and arousal. In the case of *"ok"*, the pragmatic listener places almost the same probability mass on the speaker wanting to communicate the state *ok* with high arousal, but is unsure of the valence being high or low. We interpret the probability on the state *amazing* to be in this case due to the strong prior on that state. Finally, when the speaker utters *"terrible"*, the same is interpreted by the listener as an ironic statement. The listener, using her prior knows it is highly unlikely that the speaker was referring to the weather literally being terrible, and instead infers the speaker had a different communicative goal, that of communicating their arousal.

These results demonstrate the capability of the model to make correct inference over ironic statements. Now we proceed to the extended model in which, just as in the previous case study, the listener perceives the speaker's facial expression along with the utterance. However, in this case the inference is over both valence and arousal instead of only arousal, and the speaker is furthermore able to see the surroundings which all together comprise a set of cues over the affective dimensions. These are integrated in the inferential process and influence the posterior.

In the previous example uttering *"terrible"* produced an interpretation of the statement as ironic. By definition of irony, negative utterances accompanied by positive affect should be interpreted as such. If the speaker chooses to accompany their utterance of *"terrible"* with a smile, it should make the pragmatic listener even more certain of the statement being ironic. The listener in addition has a strong prior for the weather not being terrible and perceives the weather actually not being terrible. Indeed, the posterior for this case is presented in fig. 7.13 and a larger probability mass is placed on the state being amazing, with high valence and arousal, and the goal of communicating arousal. Instead, under the same conditions, should the speaker instead of a smile express sadness, the statement should be perceived as less ironic as the facial expression is more congruent with the speaker's utterance. The posterior for that case is presented in fig. 7.14. The pragmatic speaker confidently infers that the state the speaker refers to is *ok*, with low valence and arousal, and that the speaker actually wanted to communicate their valence. One might expect the inferred state to be *terrible*, however we remind the reader that the prior is strongly skewed away from that state. Instead, the speaker's utterance is interpreted as an exaggeration in terms of its literal meaning with the aim of communicating valence, in this case a disappointment the weather is not as nice as they expected.

An emotional expression need not always alter the intended state or communicative goal, but instead communicate concomitantly and parallelly an affective state alongside an utterance. In the base example of the utterance *"ok"* the pragmatic listener was indecisive about the speaker's valence. Fig. 7.15 is depicted the posterior for the same utterance accompanied by a happy expression. The posterior is largely similar, but places a far greater probability mass on the valence being high due to the happy facial expression of the speaker. The two channels of communication may interact in certain contexts, but may also be parallel.

The prior over states influences heavily the inference the listener will make. It is indeed a prerequisite for the successful interpretation of irony in the case of the base model, but also in the case of the modified model plays a considerable role. To illustrate its role we analyse the following three simulations with different state priors.

The first case is that of having a 10 times higher prior probability of the state being *ok* instead of the other two states. The speaker utters *"terrible"* with a sad facial expression. The posterior, as depicted in fig. 7.16, reveals that the most likely interpretation for the pragmatic listener is for the state *ok* and for the goal of communicating valence. It should however be kept in mind that also in this simulation the observation of the interlocutors' surroundings which are that of perfect weather are included. In the second case the prior over states is uniform, the speaker utters *"terrible"* with a neutral facial expression. As might be expected, the inferred communicative goal is that the speaker actually wanted to communicate the state *terrible*, as visible in fig. 7.17. Finally, in the third example we have the converse case of irony. With a state prior for the state *terrible* five times that of the other two states, the speaker utters *"amazing"* with a sad facial expression, one that is incongruent to their utterance. The posterior, presented in fig. 7.18, places confidently most of the probability mass on the state *terrible* with the goal of communicating arousal, a result that is nearly the mirror image of the results in figures 7.12 and 7.13.

From the above considerations we draw two main conclusions. Firstly, while a working computational model of irony such as RSA is already a great advancement towards better human-computer interaction, the complexity of human behaviour far exceeds that of only the linguistic domain. The alterations to the posterior introduced by the modified model and the resulting interpretations highlight the importance of non-verbal communication. Secondly, the importance of the prior knowledge shared between the speaker and listener is paramount to successful interpretations of irony, and other non-literal language uses. In our implementation of irony the priors are hard-coded, but effort towards more automatic, general and diverse modelling of general shared knowledge is central to further developments.

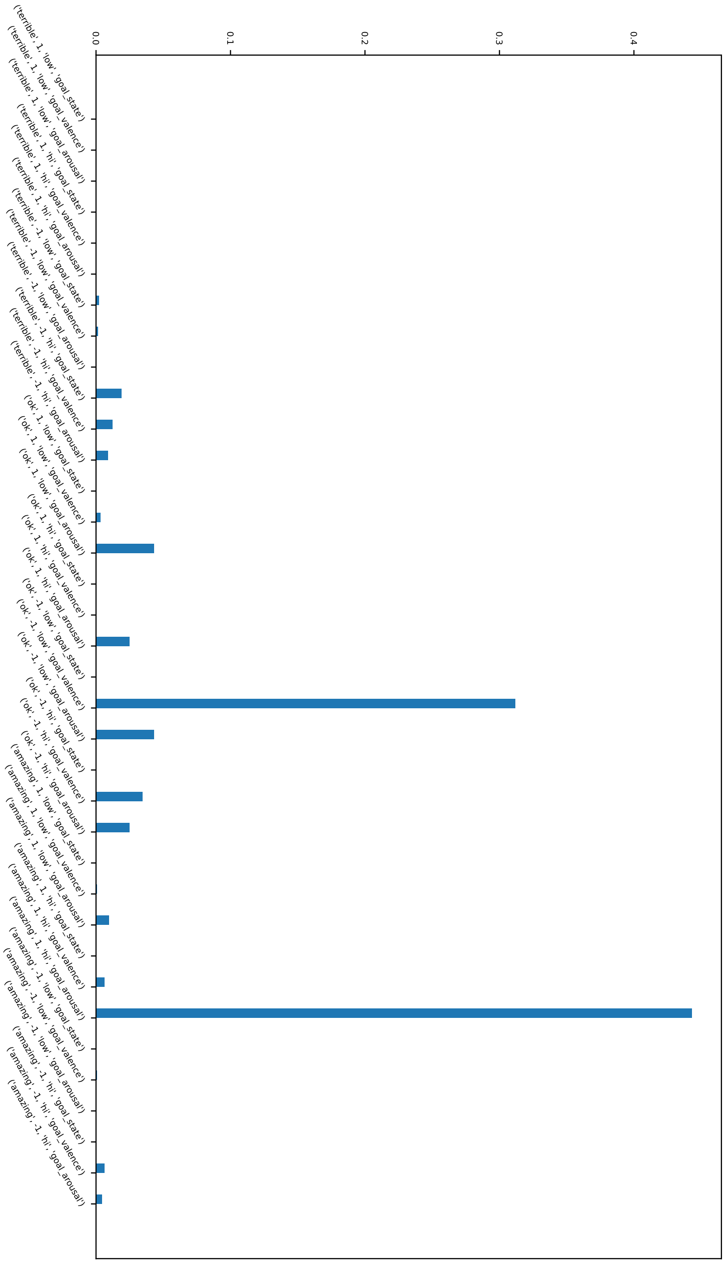


Figure 7.10: *The posterior probability over all states for the irony base model for the utterance "amazing".*

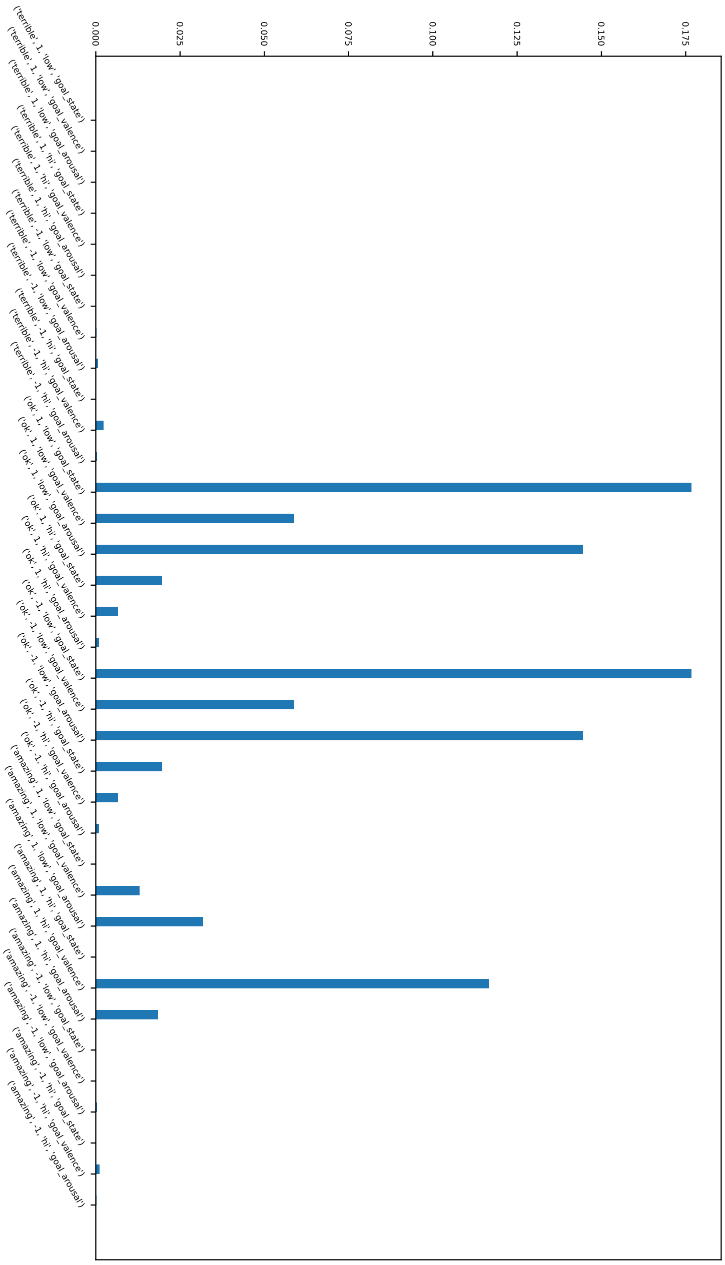


Figure 7.11: *The posterior probability over all states for the irony base model for the utterance "ok".*

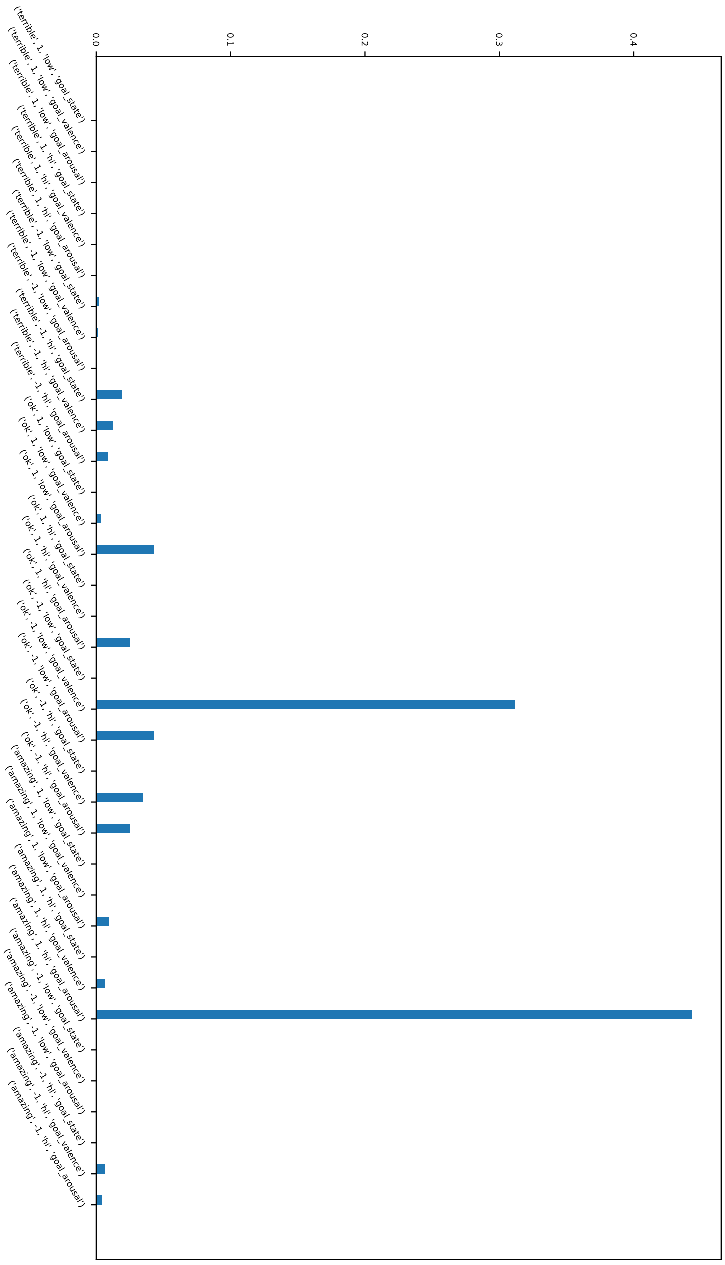


Figure 7.12: *The posterior probability over all states for the irony base model for the utterance "terrible".*

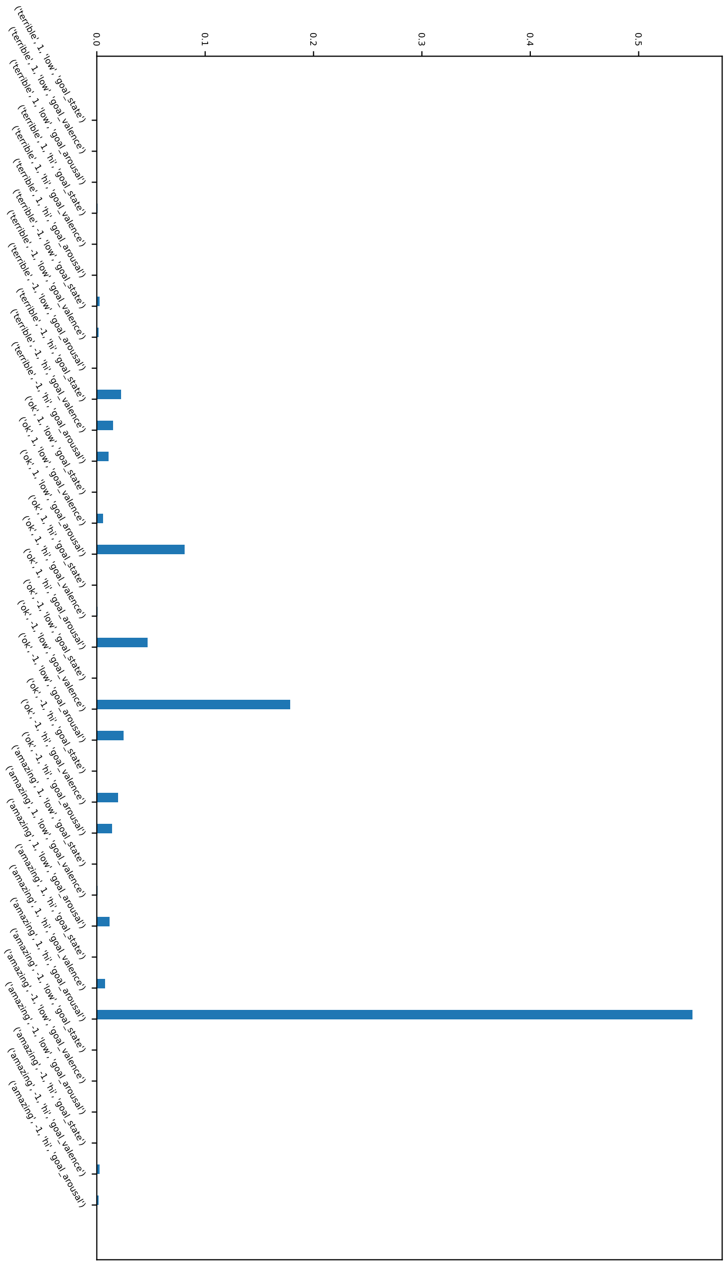


Figure 7.13: *The posterior probability over all states for the modified irony model for the utterance "amazing" and the expression happy.*

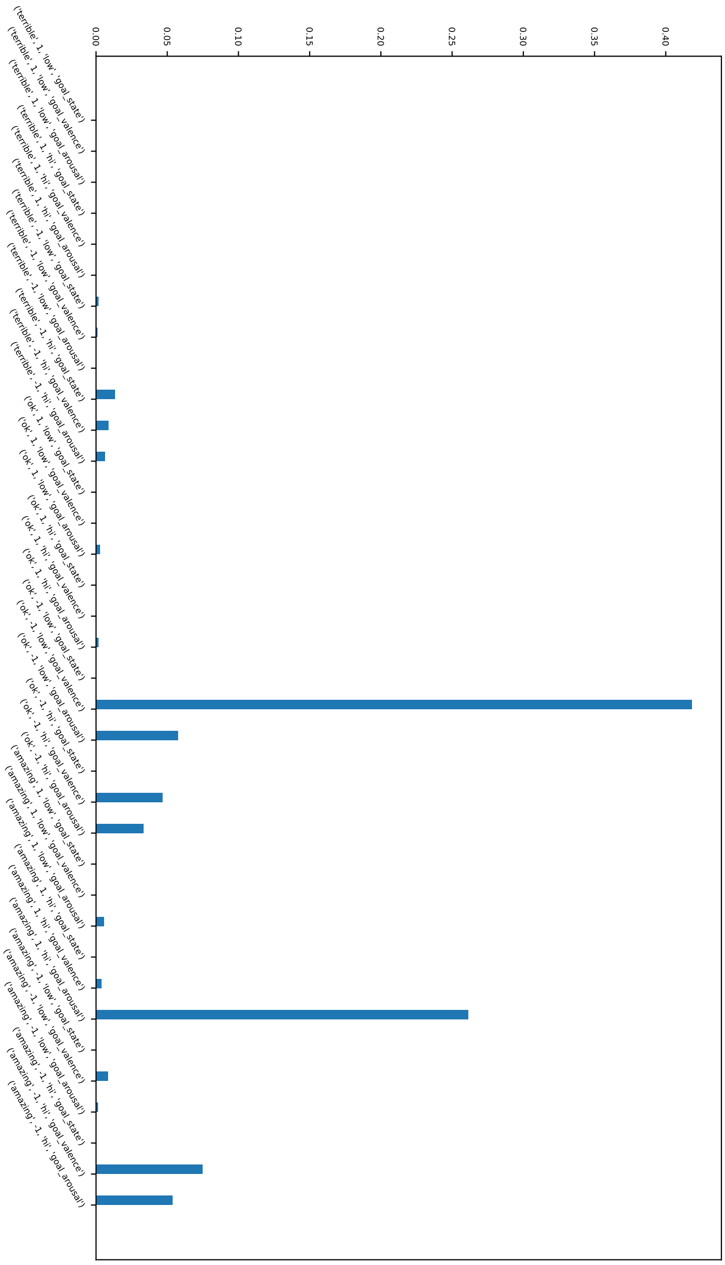


Figure 7.14: *The posterior probability over all states for the modified irony model for the utterance "amazing" and the expression sad.*

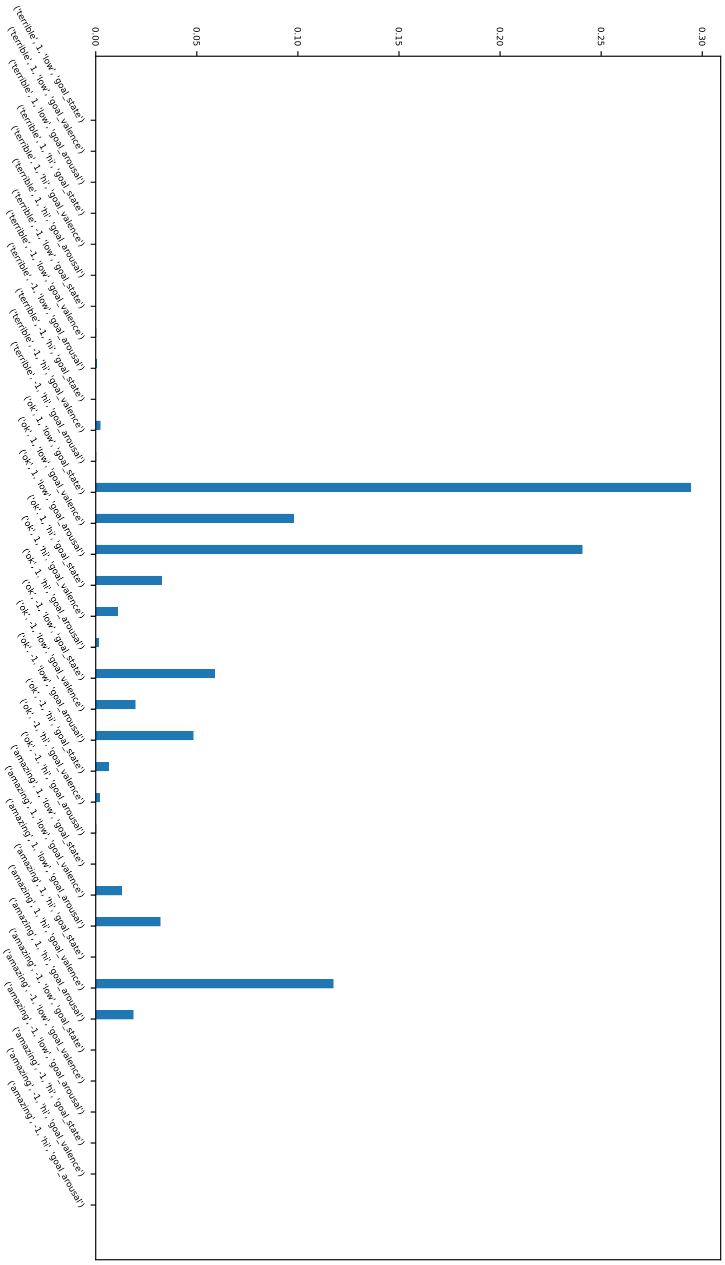


Figure 7.15: *The posterior probability over all states for the modified irony model for the utterance "ok" and the expression happy.*

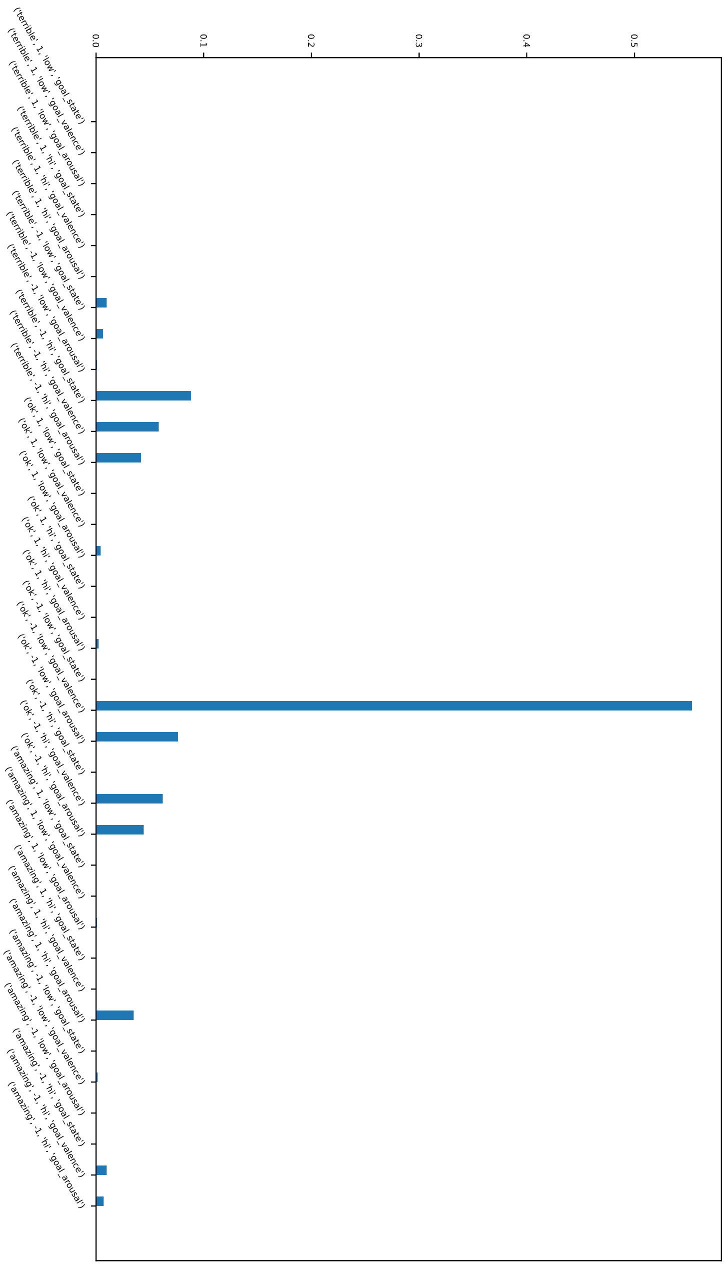


Figure 7.16: *The posterior probability over all states for the modified irony model for the utterance "terrible" and the expression sad where the prior probability of state "ok" is ten times that of the other states.*

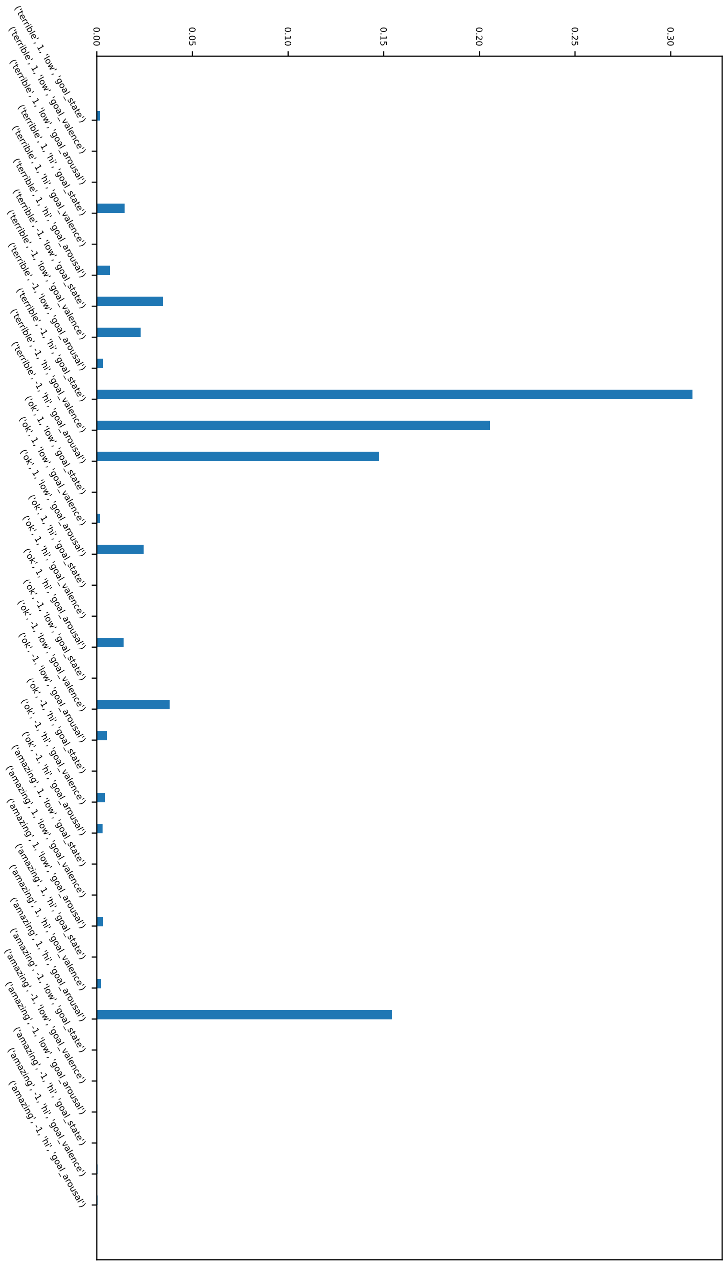


Figure 7.17: *The posterior probability over all states for the modified irony model for the utterance "terrible" and the expression neutral with a uniform prior.*

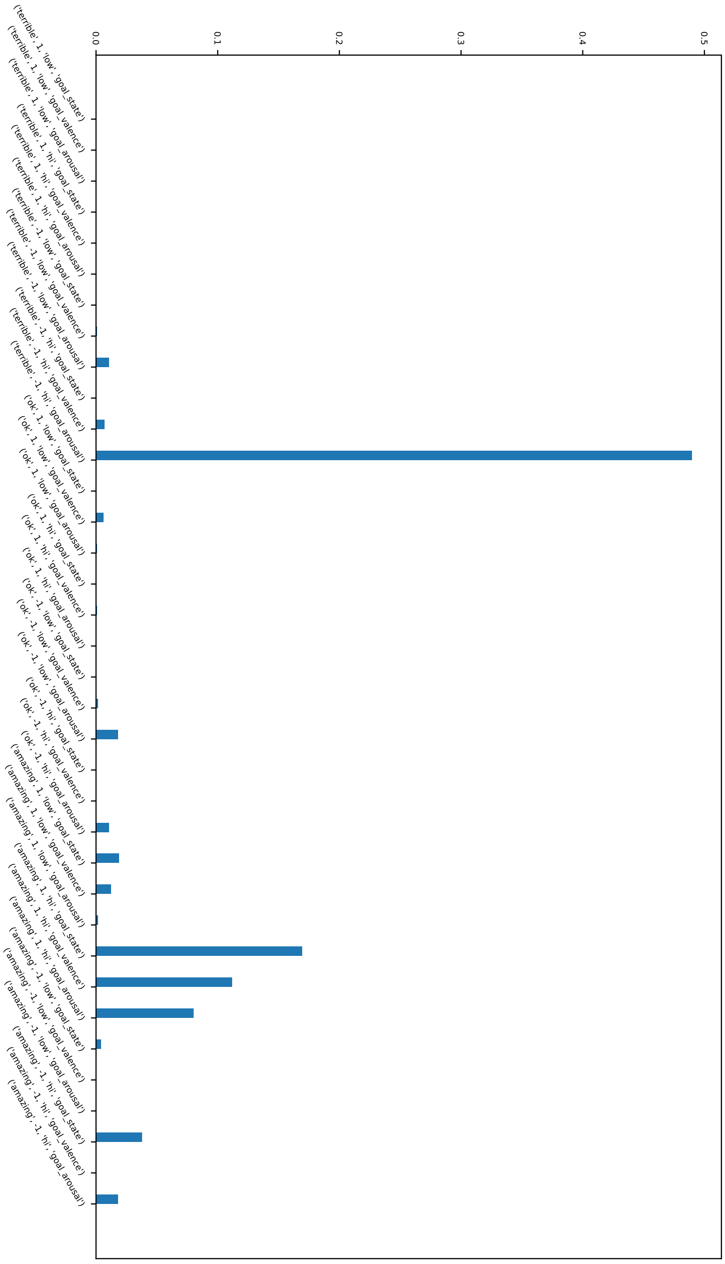


Figure 7.18: *The posterior probability over all states for the modified irony model for the utterance "amazing" and the expression sad where the prior probability of state "terrible" is five times that of the other states.*

7.3 Case study 3: Politeness

Your friend has a passion for the visual arts and enjoys painting, regardless of them lacking notable talent in the field. One fine day they enthusiastically invite you for coffee and proudly but shyly present you their latest artistic masterpiece and ask for your opinion. You are presented with a piece of questionable aesthetic and artistic value but you hear yourself utter *"I like it!"*. They smile and say *"Thank you!"*, knowing at the same time that at least in part you said it out of politeness, being their friend and caring for their feelings.

Most language models are not equipped to handle inference over such communicative intentions as conveying information indirectly, falsely, or out of care for the listener's feelings, as the assumption underlying them is that of an epistemic goal on part of the speaker - conveying information accurately. However, the RSA model is capable of modelling a speaker engaged in such an interaction with a listener who infers the speaker's intended motivations. We present the general model first, and then as in the previous examples consider an extension that provides an additional channel of communication, the speaker's facial expression, revealing the speaker's affective state in regard to an external stimulus.

Implementation details The listener is curious about the rating the speaker gives the drawing on a scale from 1 to 5, which defines the set of possible word states $S = \{1, 2, 3, 4, 5\}$. The speaker may utter any of the following five utterances

$$\mathcal{UT} = \{\text{"terrible", "bad", "okay", "good", "amazing"}\}$$

The literal listener will reason about the state s given the speaker's utterance u in terms of the literal meaning of the utterance (i.e. the literal interpretation function $[[u]](s)$), employing their prior beliefs over the states $P(s)$ to arrive at a posterior distribution over world states:

$$P_{Lit}(s \mid u) \propto \delta_{[[u]](s)} P(s) \tag{7.14}$$

The pragmatic speaker considers the pragmatic listener seeking to maximize their utility. However, instead of valuing only epistemic utility, the speaker is also interested in not hurting the listener's feelings. The speaker's utility function is thus composed of two terms: the usual surprisal based epistemic utility, and social utility:

$$U_e(u, s) = \log P_{Lit}(s \mid u) \tag{7.15}$$

$$U_s(u, s) = \mathbb{E}_{P_{Lit}(s|u)}[V(s)] \tag{7.16}$$

Social utility is calculated as the expected value of a transformation of the state of the world inferred under the literal listener's posterior. The transformation V is in our case an affine transformation:

$$V(s) = as + b \tag{7.17}$$

$$a > 0 \tag{7.18}$$

The pragmatic speaker is thus presented with a trade-off, they can either be sincere and convey the true state of the world, or they can choose to be polite and lie by inflating their rating of the drawing. The choice is itself parameterized as a linear combination of the two through the parameter ϕ. The speaker's utility function then becomes:

$$U^*(u, s, \phi) = \phi U_e(u, s) + (1 - \phi)U_s(u, s) \tag{7.19}$$
$$0 \leq \phi \leq 1 \tag{7.20}$$

which they then use to choose their utterance:

$$P_{S_1}(u \mid s, \phi) \propto \exp \alpha U^*(u, s, \phi) \tag{7.21}$$

The parameter ϕ represents the inclination of the speaker to be sincere instead of being polite, a speaker with a value of ϕ closer to one will weigh being epistemic much more in choosing the utterance, and vice versa. The pragmatic listener, instead, is endowed with a prior distribution over the parameter ϕ which they then use to simulate the speaker, finally arriving at a posterior distribution over the states s and the parameter ϕ:

$$P_{L_1}(s, \phi \mid u) \propto S_1(u \mid s, \phi)P(s)P(\phi) \tag{7.22}$$

We now turn to the case where the listener other than hearing the speaker's utterance, also perceives the speaker's facial expression. As in the previous simulations, the exteroceptive unimodal representation Y^{NV} is obtained through the backward inference $Y^{NV} \leftarrow \mathcal{O}^{NV}(X^{face})$ and gives rise to an actual exteroceptive instantiation of the RV Y^{NV} in terms of the set of facial action units. The listener is thus equipped with a model of the joint distribution of the facial expression and valence $P(AU, v_f)$. The facial expression gives an additional cue to the listener as to the speaker's actual liking or disliking of the object through the valence generating the expression, and the pragmatic listener uses it to infer the posterior on the state and parameter ϕ, i.e. the speaker's propensity for sincerity. This stimulus sets in motion an internal simulation loop of physiological signals that in turn end up reinforcing the inferred affective state, analogously to the previous two experiments.

Intuitively, the listener simulates a speaker who has two contradicting objectives: that of correctly informing the listener, and that of being polite. With the added channel of communication through the facial expression, the speaker will in the case of epistemic utility seek to express that which is coherent with their utterance. A bad review will be accompanied by a facial expression communicating low valence, and vice versa, a good review will be accompanied by an expression of high valence, e.g. a smile. In the case of social utility, the speaker will nonetheless value more to facially express happiness and contentment through high valence than otherwise, believing that it will make the listener happy.

More formally, the speaker and listener are endowed with a prior over valence conditional on the state. As in the example of irony we represent valence as a binary variable, i.e. as being either high or low $v \in V = \{-1, 1\}$. The prior can thus be expressed as

$$P_V(v \mid s) = Bern(v \mid \pi^V(s_i)) \tag{7.23}$$

The literal listener uses this prior given an utterance u to arrive at the posterior over meanings and valence:

$$P_{Lit}(s, v \mid u) \propto \delta_{[[u]](s)} P_V(v \mid s)P(s) \tag{7.24}$$

The pragmatic speaker simulates the pragmatic listener using their utility function U^* to choose an utterance:

$$P_{S_1}(u \mid s, v, \phi) \propto \exp \alpha U^*(u, s, v, \phi) \tag{7.25}$$

The speaker's epistemic utility now has to also account for valence. We keep the spirit the same as the base case and use the surprisal based epistemic utility, but in this case over the joint posterior of the listener of meaning and valence. Should for a particular utterance the speaker choose an unlikely facial expression under the literal listener's posterior, their epistemic utility will be lower. The speaker's social utility is again the expected value of a transformation of the state of the world inferred under the literal listener's posterior, but here we additively and analogously include valence. Note that to calculate the expectations the speaker has to first marginalise the literal listener's posterior over valence and world states, respectively:

$$U_e(u, s, v) = \log P_{Lit}(s, v \mid u) \tag{7.26}$$

$$U_s(u, s, v) = \mathbb{E}_{P_{Lit}(s|u)}[V^M(s)] + \mathbb{E}_{P_{Lit}(v|u)}[V^V(v)] \tag{7.27}$$

with V^S being the same function from eq. 7.17, and V^V also being a linear transformation with different parameters. The two then form part of the speaker's composite utility $U^*(u, s, v, \phi)$ analogously to eq. 7.19:

$$U^*(u, s, v, \phi) = \phi U_e(u, s, v) + (1 - \phi) U_s(u, s, v) \tag{7.28}$$

$$0 \leq \phi \leq 1 \tag{7.29}$$

The trade off a speaker faces in choosing their action is a trade off between the two aforementioned goals, an epistemic and a social goal. Given the parameter ϕ the speaker through the optimization of $U^*(u, s, v, \phi)$ chooses the appropriate action $\mathcal{A}^{utt}$ that finally produces the utterance u. Both of the goals are intended and voluntary social signaling the speaker makes, i.e. explicit goals $\mathcal{G}^{exp}$, as described in the previous section. However, the speaker may also have implicit goals, $\mathcal{G}^{imp}$, which may come in conflict with the explicit ones. Ideally, the listener would model both, as both represent two possible sources of the observed outcome X^{face}. However, we restrict our analysis to explicit goals $\mathcal{G}^{exp}$.

Finally, exactly as in the previous examples the pragmatic speaker simulates the pragmatic listener using the integrated cues at which point the pragmatic listener's posterior becomes:

$$P_{L_1}(s, \phi \mid u, AU) \propto S_1(u \mid s, \bar{v}, \phi) P(\bar{v} \mid s, AU, \tilde{p}_{1,\ldots,N}) P(s) P(\phi) \tag{7.30}$$

The implementation models of the joint distribution of valence and face, $P(AU, v_f)$, and of the joint distribution of valence and physiological signals. $P(Y^i_{1:N}, v_p)$, are the same as in the two previous experiments.

Results As in the previous case studies, we first examine the inferences of the base model and then consider the modified model and compare the two. Figures 7.19 to 7.22 depict the posterior over states and the parameter ϕ for the utterances *"amazing"*, *"good"* and *"terrible"*, respectively. In the case of *"amazing"* the listener places the most probability on the state being *amazing*, but a considerable probability mass is dispersed on other states, most notably the state *good* that isn't far behind. In addition, the speaker is perceived as insincere, with the expected value of the posterior distribution of ϕ being 0.37. In other words, the speaker is 37% sincere. This is in stark contrast with the case of the utterance *"terrible"* where the inferred state is confidently *terrible*, and the speaker is perceived as 76% sincere. In the case of *"good"* the listener infers the most likely state to be *okay* instead of good, and the speaker is nonetheless perceived as rather insincere. We interpret this result being due to two factors: the prior over states given the utterance (which in this case is actually the literal listener's posterior), and the way the problem is fundamentally posed. Any positive statement has a possibility of "coming" from either the speaker's sincere opinion or their incentive to be polite. The two coupled together provide the result that is seen here for the utterances *"good"* and *"amazing"*, i.e. positive statements may be viewed skeptically from the point of view of sincerity.

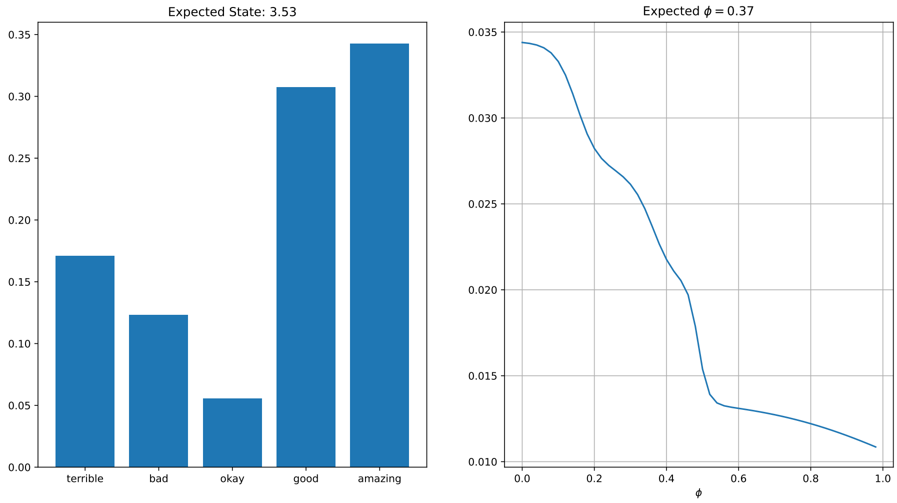


Figure 7.19: *The posterior probability of the states (left) and the parameter ϕ (right) of the politeness base model for the utterance "amazing".*

Turning to the modified model, in figures 7.23 and 7.24 are depicted the cases of the utterance *"okay"* with a happy and sad facial expression, respectively, effectively communicating high and low arousal. The effect the addition of facial expression has on the posterior over the states is evident here. While in both cases the most likely state for the listener is indeed *ok*, in the case of happy expression the rest of the probability mass is largely placed on *bad*, and in the case of a sad expression it is largely placed on more positive states. Both a high valence and low valence expression such as happy and sad, respectively, are unlikely under the set prior for the state *ok* which is that of a neutral valence. Curiously, however, only in the case of a sad expression is the speaker perceived as insincere.

In a different vein, figures 7.26 and 7.25 present the posteriors for the utterance *"amazing"* with a neutral and sad facial expression, respectively. There is little differ-

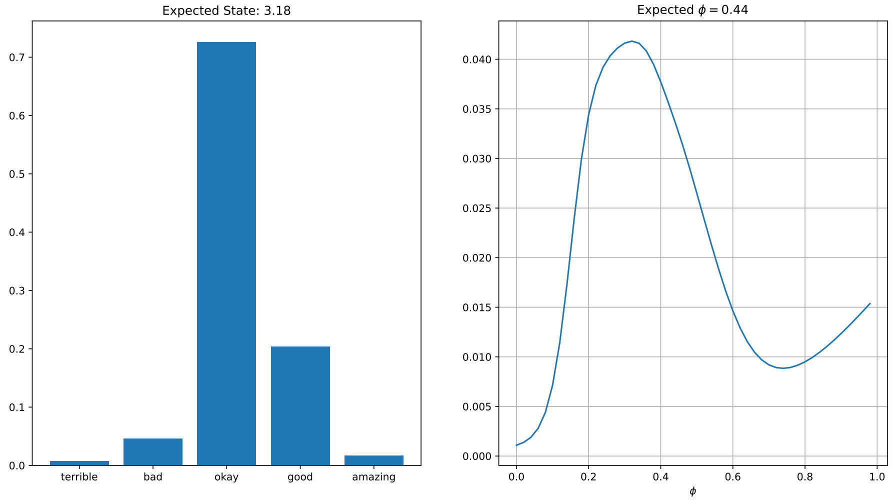


Figure 7.20: *The posterior probability of the states (left) and the parameter ϕ (right) of the politeness base model for the utterance "good".*

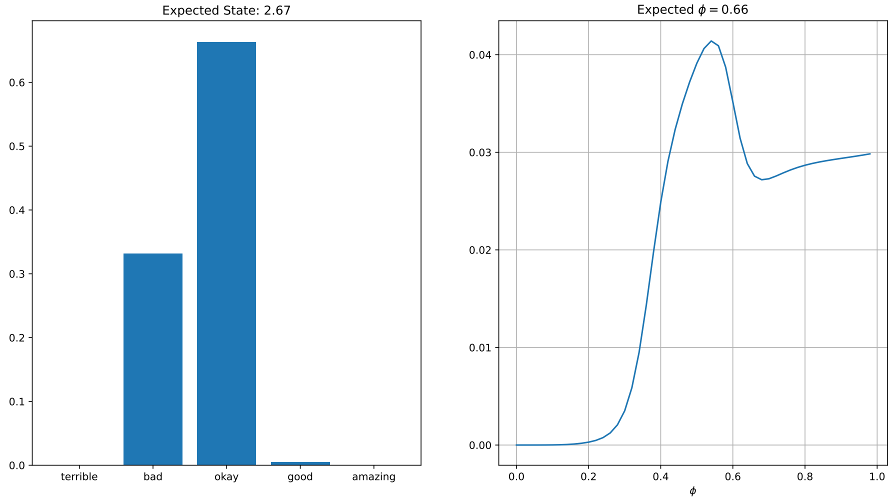


Figure 7.21: *The posterior probability of the states (left) and the parameter ϕ (right) of the politeness base model for the utterance "okay".*

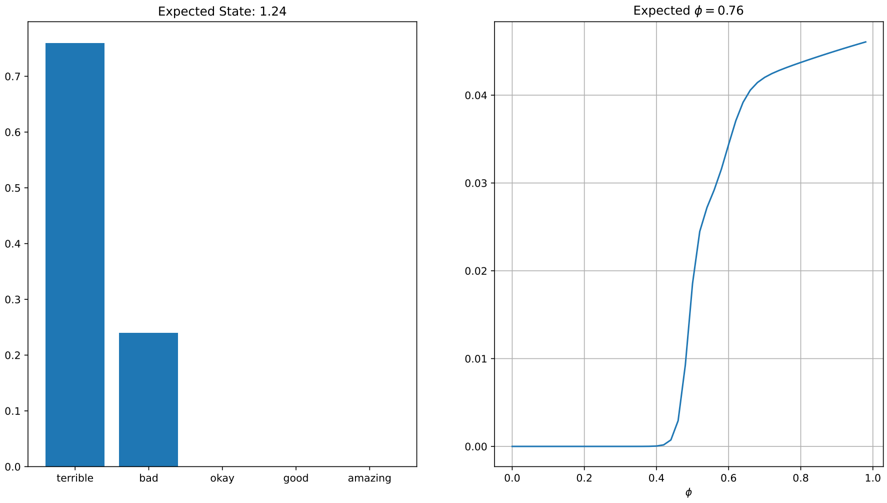


Figure 7.22: *The posterior probability of the states (left) and the parameter ϕ (right) of the politeness base model for the utterance "terrible".*

ence between the base case for *"amazing"* and the one with the neutral expression. The correct state is inferred but with a slightly higher probability placed on the state terrible. We believe this to be due to the prior of high valence under the statement being high, but at the same time observing a less likely neutral expression. However, a sad expression changes the inferred state making the state *good* more likely. Again, curiously enough, in the case of a sad expression the speaker is perceived as slightly more sincere at a mean value of ϕ of 0.48. However, the shape of the posterior on the parameter ϕ is rather different than in the other examples. A mean value in this case does not represent well the information the posterior presents. The shape of the curve is largely convex with a sharp dip and local minimum around the value 0.2. A large probability mass is placed on the values less than two, and another large portion on values higher than two. In a certain sense, for the listener it is both likely that the speaker is sincere and insincere.

While the results of the modified model are promising and do admit a sensible interpretation for the posterior over the states, in general it seems to give answers that are more difficult to intuitively interpret than in the other examples. This especially regards the inference of the posterior of the parameter ϕ. This difficulty in finding an intuitive interpretation might stem in part from the inherent complexity of the presented scenarios. It would be unclear how to interpret even in real life should a friend utter *"amazing"* with an extremely sad facial expression. Naturally one would not take seriously their review, but might rather be concerned for the friend's welfare as that reaction would be highly peculiar, even bizarre. Nonetheless, these results further testify of the general ability of the RSA approach to model highly complex linguistic phenomena, and not only, but in general to model complex multimodal human communication.

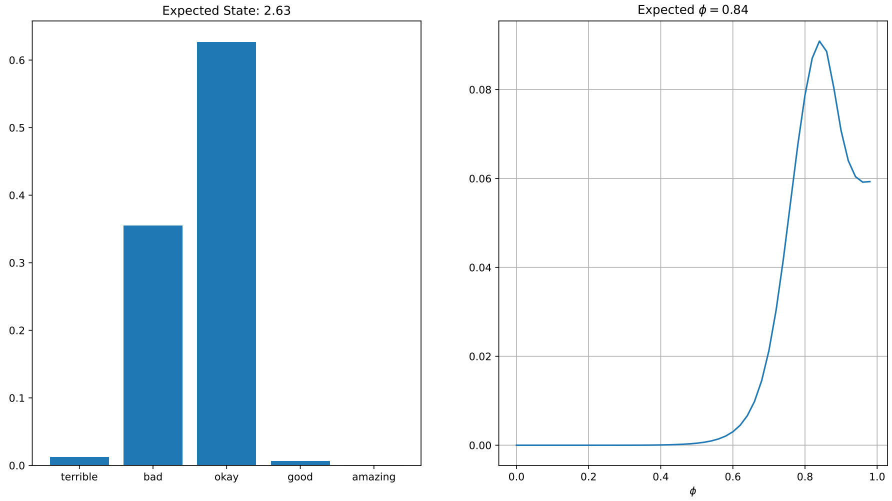


Figure 7.23: *The posterior probability of the states (left) and the parameter ϕ (right) of the modified politeness model for the utterance "okay" and expression happy.*

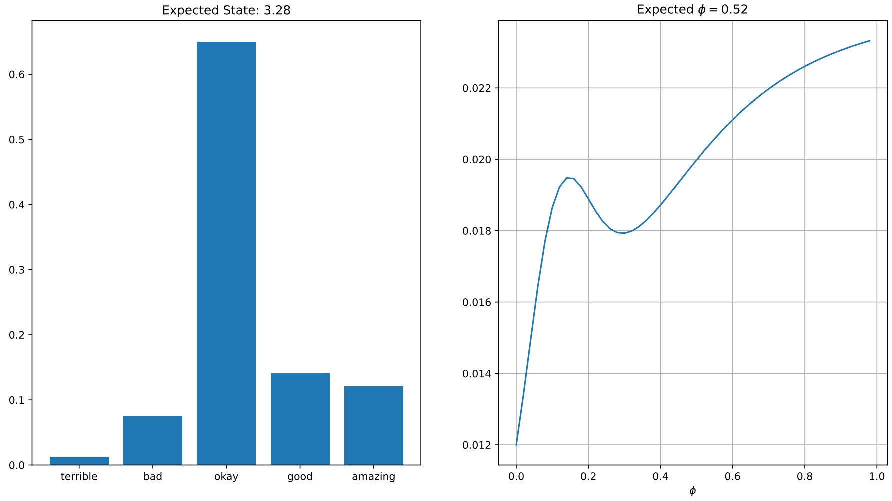


Figure 7.24: *The posterior probability of the states (left) and the parameter ϕ (right) of the modified politeness model for the utterance "okay" and expression sad.*

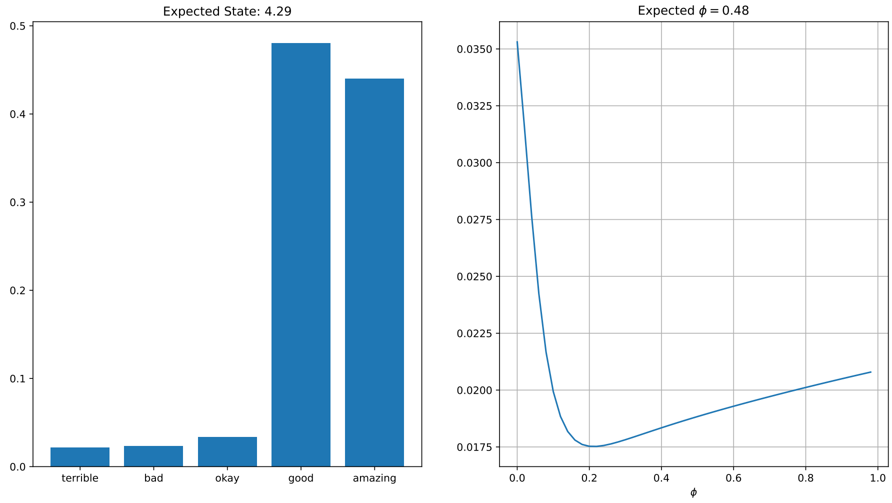


Figure 7.25: *The posterior probability of the states (left) and the parameter ϕ (right) of the modified politeness model for the utterance "amazing" and expression sad.*

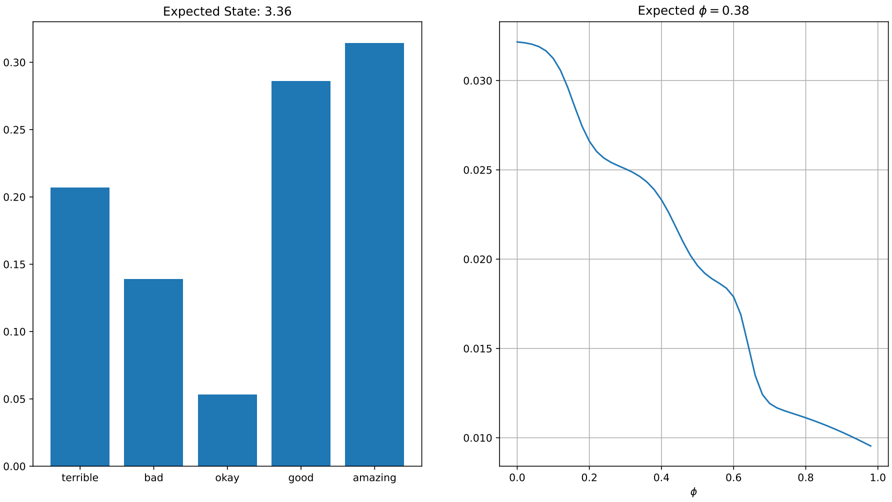


Figure 7.26: *The posterior probability of the states (left) and the parameter ϕ (right) of the modified politeness model for the utterance "amazing" and expression neutral.*

7.4 Case study 4: Exploring core affect dynamics

Differently from previous experiments, here we focus on the dynamics at the core affect level which is a key aspect of our model.

Recall that the phenomenological model by Kuppens et al. (2010), based on the analysis of human collected data, posits that core affect dynamics is consequent on the activity of a complex, open system. The latter is more suitably conceived as subject to stochastic variability resulting from the entanglement of many internal (and external) activities that influence it. Across time, the core affect unfolding, as observed from the sampling of experiential data, can be represented as a trajectory, i.e. a realisation of a stochastic process. Such random path, an OU stochastic process, reflects the typical pattern of affective changes and fluctuations that V/A levels undergo across time and that characterise an individual Kuppens et al. (2010).

The aim of this simulation is thus to assess whether, as discussed in Section 6.6.2, our model can reproduce such dynamics.

In this case, we can take advantage of a the publicly available dataset RECOLA Ringeval et al. (2013). This dataset is a multimodal corpus of spontaneous collaborative interactions in French. Aiming at studying the impact of emotional feedback on teamwork quality and efficiency, 46 participants took part in the test where several multimodal data, i.e., audio, video and physiological signals were recorded continuously and synchronously. In addition, 6 annotators concentrated on the labelling of both the affective and social behaviours that were produced by participants during their collaboration. As to affective behaviour, affect was measured continuously on the two psychological dimensions of arousal and valence (corresponding to the process $F = \{F_t, 0 \leq t \leq T\}$).

Implementation details The variational autoencoder (VAE) is an end-to-end latent variable model based on deep neural networks fist introduced by Kingma and Welling (2013). The VAE model posits a generative model as follows:

$$P(\mathcal{Z} \mid \mathcal{O}) = \frac{P(\mathcal{Z}, \mathcal{O})}{P(\mathcal{O})} \tag{7.31}$$

where the latent random variable $\mathcal{Z}$ captures the variability in the observed variable $\mathcal{O}$. The generative mapping $p(\mathbf{x} \mid \mathcal{Z})$ is commonly realized by a deep neural network which allows for highly non-linear mappings in generation. However, this makes inference of the posterior $p(\mathcal{Z} \mid \mathcal{O})$ intractable. The VAE hence uses a variational approximation $q(\mathcal{Z} \mid \mathcal{O})$ by optimizing for the evidence lower bound or ELBO (see, Appendix C):

$$\log P(\mathcal{O}) \geq -KL(Q(\mathcal{Z} \mid \mathcal{O}) \| P(\mathcal{Z})) + \mathbb{E}_{Q(\mathcal{Z}|\mathcal{O})}[\log P(\mathcal{O} \mid \mathcal{Z})] \tag{7.32}$$

The approximate posterior $Q(\mathcal{Z} \mid \mathcal{O})$ is originally a Gaussian $\mathcal{N}(\boldsymbol{\mu}, diag(\boldsymbol{\sigma}))$ whose parameters are likewise the output of a non-liner mapping, usually a neural network. As the inference model and the generative model are trained jointly, to allow the passage of the gradient through the model, the so called *"reparameterization trick"* is applied $\mathcal{Z} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is sampled from standard multivariate Gaussian distribution. The PGM of a VAE is depicted on the left of Fig. 7.27.

However, the VAE is not designed for capturing the dynamics of its latent space $\mathcal{Z}$. There are however different "flavors" of the VAE developed that tackle this issue (for a comprehensive review see (Girin et al., 2021)). The variational recurrent neural network (VRNN) model by Chung et al. (2015) models temporal dependence in the latent space by introducing a prior through the hidden state of a recurrent neural network that is motivated by the following factorization of the joint distribution:

$$P(\mathcal{O}_{1:T}, \mathcal{Z}_{1:T}) = \prod_{t=1}^{T} P(\mathcal{O}_t \mid \mathcal{Z}_{1:t}, \mathcal{O}_{1:t-1}) P(\mathcal{Z}_t \mid \mathcal{O}_{1:t-1}, \mathcal{Z}_{1:t-1}). \tag{7.33}$$

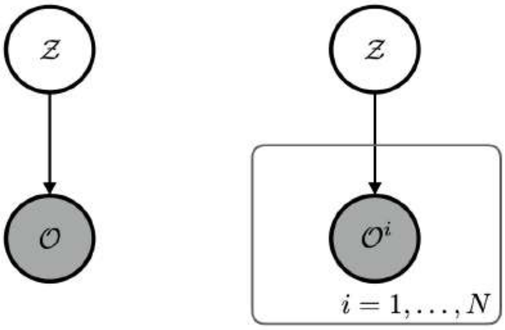


Figure 7.27: *Left: the PGM of the variational autoencoder. Right: The PGM of the multimodal variational autoencoder introduced in the simulation; a graphical plate notation is used for compacting the $i : 1, \ldots, N$ observables $\mathcal{O}^i$*

In what follows we briefly expose the workings of the VRNN model, extend it to include several modalities, and apply it the problem of jointly modelling the facial expression, prosody and affective state of a speaker through time. We introduce the VRNN through its three parts: the generative model, the inference model and the loss function.

The generative model of the VRNN is defined as:

$$\mathbf{h}_t = \mathbf{d}_\mathbf{h}(\varphi_\mathcal{O}(\mathcal{O}_{t-1}), \varphi_\mathcal{Z}(\mathcal{Z}_{t-1}), \mathbf{h}_{t-1}), \tag{7.34}$$

$$[\boldsymbol{\mu}_{\theta_\mathcal{O}}(\mathcal{Z}_t, \mathbf{h}_t), \boldsymbol{\sigma}_{\theta_\mathcal{O}}(\mathcal{Z}_t, \mathbf{h}_t)] = \mathbf{d}_\mathcal{O}(\varphi_\mathcal{Z}(\mathcal{Z}_t), \mathbf{h}_t), \tag{7.35}$$

$$P_{\theta_\mathcal{O}}(\mathcal{O}_t|\mathcal{Z}_t, \mathbf{h}_t) = \mathcal{N}\big(\mathcal{O}_t; \boldsymbol{\mu}_{\theta_\mathcal{O}}(\mathcal{Z}_t, \mathbf{h}_t), \mathrm{diag}\{\boldsymbol{\sigma}^2_{\theta_\mathcal{O}}(\mathcal{Z}_t, \mathbf{h}_t)\}\big). \tag{7.36}$$

where $\varphi_\mathcal{O}$ and $\varphi_\mathcal{Z}$ are feature extractors (implemented as neural networks) of the inputs $\mathcal{O}$ and the sample from the latent space $\mathcal{Z}$, respectively. Eq. 7.34 is the recurrence equation and effectively defines how the VRNN updates its hidden state $\mathbf{h}_t$. In other words, it's the transition function of the RNN. Equations 7.35 and 7.36 define the generative networks, i.e. the decoders that generate the observed variable x from the latent space sample $\mathcal{Z}_t$ and the hidden space of the RNN $\mathbf{h}_t$. The VRNN also has a prior distribution over the latent space $\mathcal{Z}_t$:

$$[\boldsymbol{\mu}_{\theta_\mathcal{Z}}(\mathbf{h}_t), \boldsymbol{\sigma}_{\theta_\mathcal{Z}}(\mathbf{h}_t)] = \mathbf{d}_\mathcal{Z}(\mathbf{h}_t), \tag{7.37}$$

$$P_{\theta_\mathcal{Z}}(\mathcal{Z}_t|\mathbf{h}_t) = \mathcal{N}\big(\mathcal{Z}_t; \boldsymbol{\mu}_{\theta_\mathcal{Z}}(\mathbf{h}_t), \mathrm{diag}\{\boldsymbol{\sigma}^2_{\theta_\mathcal{Z}}(\mathbf{h}_t)\}\big). \tag{7.38}$$

The inference model (the approximate posterior) is instead defined as follows:

$$[\boldsymbol{\mu}_\phi(\mathcal{O}_t, \mathbf{h}_t), \boldsymbol{\sigma}_\phi(\mathcal{O}_t, \mathbf{h}_t)] = e_\mathcal{Z}\big(\varphi_\mathcal{O}(\mathcal{O}_t), \mathbf{h}_t\big), \tag{7.39}$$

$$Q_\phi(\mathcal{Z}_t|\mathcal{O}_t, \mathbf{h}_t) = \mathcal{N}\big(\mathcal{Z}_t; \boldsymbol{\mu}_\phi(\mathcal{O}_t, \mathbf{h}_t), \mathrm{diag}\{\boldsymbol{\sigma}^2_\phi(\mathcal{O}_t, \mathbf{h}_t)\}\big), \tag{7.40}$$

with $e_\mathcal{Z}$ being the encoding network. The inference model is inspired by the mean field factorization (Eq. C.9, Appendix C), which here writes:

$$Q(\mathcal{Z}_{1:T} \mid \mathcal{O}_{1:T}) = \prod_{t=1}^{T} Q(\mathcal{Z}_t \mid \mathcal{O}_{1:t}, \mathcal{Z}_{1:t-1}). \tag{7.41}$$

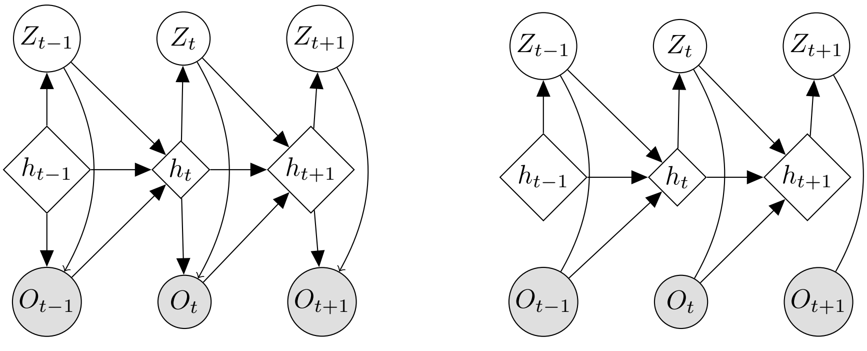


Figure 7.28: *VRNN's graphical model during generation (left) and inference (right).*

The points of contact between the VAE and RNN components of the model are manifold. To more easily understand the model the PGMs of the generative and inference model are presented in fig. 7.28 (adapted from Girin et al. (2021)).

The optimization objective of the VRNN is a modified ELBO loss, which can be obtained by Eq. 6.52 by omitting the controls $\mathcal{U}$ (but see the supplemental materials of Chung et al. (2015) for a motivated derivation):

$$\begin{aligned}\mathcal{F}\left(\theta,\phi;\mathcal{O}_{1:T}\right) = &\sum_{t=1}^{T}\mathbb{E}_{Q_\phi(\mathcal{Z}_{1:t}|\mathcal{O}_{1:T})}\big[\ln P_{\theta_\mathcal{O}}(\mathcal{O}_t|\mathcal{O}_{1:t-1},\mathcal{Z}_{1:t})\big]\\ &-\sum_{t=1}^{T}\mathbb{E}_{Q_\phi(\mathcal{Z}_{1:t-1}|\mathcal{O}_{1:T})}\left[KL\left(Q_\phi(\mathcal{Z}_t|\mathcal{Z}_{1:t-1},\mathcal{O}_{1:t})\parallel P_{\theta_\mathcal{Z}}(\mathcal{Z}_t|\mathcal{O}_{1:t-1},\mathcal{Z}_{1:t-1})\right)\right].\end{aligned} \tag{7.42}$$

For the purposes of our modelling problem, the VRNN lack a crucial component, which is multimodality. We seek to model the joint distribution over time for N observed modalities, $P(\mathcal{O}^1_{1:T},\ldots,\mathcal{O}^N_{1:T},\mathcal{Z}_{1_T})$, with the i-th modality being represented as $\mathcal{O}^i$ (cfr. Fig. 7.27, right PGM). While different variations to the VRNN model abound, there are only a few that are multimodal, such as Baruah and Banerjee (2020); Brito et al. (2020); Ong et al. (2019a). As we were not aware of their work we independently extended the VRNN to allow modelling of multimodal phenomena. We therefore in what follows introduce these modifications, that in spirit somewhat resemble the work of Brito et al. (2020), that allow it to be applied to the problem at hand.

Generally speaking, at a certain point in the model the modalities have to be fused together. There's more than one way to do that, but we took the naive approach of not interfering with the terms relating to the latent space of the VRNN nor the hidden space of its RNN as in the case of Baruah and Banerjee (2020) and Ong et al. (2019a). Instead, we instantiate an encoder and a decoder for each of the modalities and perform feature fusion before inference reaches the latent and hidden space. More specifically, the generative model is altered with respect to eq. 7.34 - 7.36 as follows:

$$\mathbf{h}_t = \mathbf{d}_h(\varphi_{\mathcal{O}}(\psi_\tau^1(\mathcal{O}_{t-1}^1), \dots, \psi_\tau^N(\mathcal{O}_{t-1}^N)), \varphi_{\mathcal{Z}}(\mathcal{Z}_{t-1}), \mathbf{h}_{t-1}), \tag{7.43}$$

$$[\boldsymbol{\mu}_{\theta_{\mathcal{O}^1}}(\mathcal{Z}_t, \mathbf{h}_t), \boldsymbol{\sigma}_{\theta_{\mathcal{O}^1}}(\mathcal{Z}_t, \mathbf{h}_t)] = \mathbf{d}_{\mathcal{O}^1}(\varphi_{\mathcal{Z}}(\mathcal{Z}_t), \mathbf{h}_t), \tag{7.44}$$

$$\vdots$$

$$[\boldsymbol{\mu}_{\theta_{\mathcal{O}^N}}(\mathcal{Z}_t, \mathbf{h}_t), \boldsymbol{\sigma}_{\theta_{\mathcal{O}^N}}(\mathcal{Z}_t, \mathbf{h}_t)] = \mathbf{d}_{\mathcal{O}^N}(\varphi_{\mathcal{Z}}(\mathcal{Z}_t), \mathbf{h}_t), \tag{7.45}$$

$$P_{\theta_{\mathcal{O}^1}}(\mathcal{O}_t^1|\mathcal{Z}_t, \mathbf{h}_t) = \mathcal{N}\big(\mathcal{O}_t^1; \boldsymbol{\mu}_{\theta_{\mathcal{O}^1}}(\mathcal{Z}_t, \mathbf{h}_t), \operatorname{diag}\{\boldsymbol{\sigma}^2_{\theta_{\mathcal{O}^1}}(\mathcal{Z}_t, \mathbf{h}_t)\}\big), \tag{7.46}$$

$$\vdots$$

$$P_{\theta_{\mathcal{O}^N}}(\mathcal{O}_t^N|\mathcal{Z}_t, h_t) = \mathcal{N}\big(\mathcal{O}_t^N; \boldsymbol{\mu}_{\theta_{\mathcal{O}^N}}(\mathcal{Z}_t, \mathbf{h}_t), \operatorname{diag}\{\boldsymbol{\sigma}^2_{\theta_{\mathcal{O}^N}}(\mathcal{Z}_t, \mathbf{h}_t)\}\big). \tag{7.47}$$

The prior distribution remains unchanged, while an additional network ψ_τ^i with parameters τ is instantiated for the i-th modality. This network serves as a feature extractor for each of the inputs effectively reducing their dimensionality. The features of each of the inputs concatenated together are then served to the encoder of the VRNN, changing equations 7.36 - 7.38 as follows:

$$[\boldsymbol{\mu}_\phi(\mathcal{O}_t^{1:N}, \mathbf{h}_t), \boldsymbol{\sigma}_\phi(\mathcal{O}_t^{1:N}, \mathbf{h}_t)] = e_{\mathcal{Z}}\big(\varphi_{\mathcal{O}}(\psi_\tau^1(\mathcal{O}_t^1), \dots, \psi_\tau^N(\mathcal{O}_t^N)), \mathbf{h}_t\big), \tag{7.48}$$

$$Q_\phi(\mathcal{Z}_t|\mathcal{O}_t^{1:N}, \mathbf{h}_t) = \mathcal{N}\big(\mathcal{Z}_t; \boldsymbol{\mu}_\phi(\mathcal{O}_t^{1:N}, \mathbf{h}_t), \operatorname{diag}\{\boldsymbol{\sigma}^2_\phi(\mathcal{O}_t^{1:N}, \mathbf{h}_t)\}\big). \tag{7.49}$$

We assume the following factorization of the joint distribution:

$$P(\mathcal{O}_{1:T}^{1:N}, \mathcal{Z}_{1:T}) = \prod_{t=1}^{T}\prod_{i=1}^{N} P(\mathcal{O}_t^i \mid \mathcal{Z}_{1:t}, \mathcal{O}_{1:t-1}) P(\mathcal{Z}_t \mid \mathcal{O}_{1:t-1}, \mathcal{Z}_{1:t-1}). \tag{7.50}$$

meaning that at time step t the observed modalities are conditionally independent given $\mathcal{Z}_{1:t-1}$ and $\mathcal{O}_{1:t-1}$. The variational lower bound from eq. 7.51 then accordingly becomes:

$$\begin{aligned}\mathcal{F}\left(\theta, \phi; \mathcal{O}_{1:T}^{1:N}\right) &= \sum_{t=1}^{T} \mathbb{E}_{Q_\phi(\mathcal{Z}_{1:t}|\mathcal{O}_{1:T})}\big[\sum_{i=1}^{N} \ln P_{\theta_{\mathcal{O}^i}}(\mathcal{O}_t^i|\mathcal{O}_{1:t-1}, \mathcal{Z}_{1:t})\big] \\ &- \sum_{t=1}^{T} \mathbb{E}_{Q_\phi(\mathcal{Z}_{1:t-1}|\mathcal{O}_{1:T})}\left[D_{\text{KL}}\left(Q_\phi(\mathcal{Z}_t|\mathcal{Z}_{1:t-1}, \mathcal{O}_{1:t}) \parallel P_{\theta_{\mathcal{Z}}}(\mathcal{Z}_t|\mathcal{O}_{1:t-1}, \mathcal{Z}_{1:t-1})\right)\right].\end{aligned} \tag{7.51}$$

A model such as the above modelling the joint distribution of all the modalities has the advantage of being able to arbitrarily condition it upon any observed subset of its inputs. To tackle the issue of missing modalities that is necessary for arbitrary conditioning (regardless of the availability of data), a predetermined value is assigned to the feature vector of that modality for that time step. Indeed, we actually condition the model in many of the possible ways during training in order to force the model to learn the meaning of a missing modality that we represent as an array of zeros of appropriate size.

We train the multimodal VRNN on the RECOLA dataset (Ringeval et al., 2013) and we are interested in jointly modelling the facial expression, voice prosody and the core affect of speakers. We first extract AUs from the video recordings using the OpenFace software package (Baltrušaitis et al., 2016), and keep only the 17 AUs whose activation intensity is measured, the scale being from 0 as the minimal activation, to 1 as the maximum. To represent prosodic features we extract the following audio features with the librosa package (McFee et al., 2015): spectral centroid, chromagram, MFCC, spectral rolloff and RMS with a window size of 0.5 seconds. Valence and arousal are instead taken from the dataset as is. They are represented on a continuous scale of -1 to 1, respectively, with -1 being the lowest, 1 the highest value, and 0 representing neutral. All the features are then resampled at a frequency of 3 Hz, synchronized, and sequences of 6 second duration are created as inputs to the model. The sequence length used, which was empirically determined, is intended to represent the mean life cycle of an emotional episode. Instead, the data frequency was adjusted empirically to be sufficiently low not to have overly slow training, and sufficiently large to not miss large variations in the data. Finally, the model was trained for 100 epochs, with a batch size of 32, and a KL-annealing factor of 10^{-3}. The dataset is composed of 24 sessions of which 4 were set aside as the test set, while the rest were used for training, which constitutes a train/test split of approximately $83\% : 17\%$. We also tried splitting the dataset in favour of having a larger test set, but RECOLA as a whole being of rather modest size, and deep neural networks being data hungry, the training performance was underwhelming under such a split. Separating by sessions instead of randomly selecting sequences from the whole dataset ensures that during testing any person specific information captured during training won't influence the testing results. All feature extractors of the VRNN were implemented as multilayer perceptrons with a hidden size of 256. The number of RNN cells were 20 with a hidden dimension of 32. The dimensionality of the feature space is 64 for both the input data features and the hidden space features, while the VAE latent space dimension is 256. The implementation was developed in the Python programming language using the PyTorch deep learning library.

7.5 Discussion

Data from the test set sessions is split in sequences of same length as used for training, and is then run through the model in three ways:

- reconstructing the valence and arousal given the AUs and audio features: $P(x_t^{VA} \mid x_{\leq t}^{F}, x_{\leq t}^{A})$, for all t in the sequence,
- reconstructing the valence and arousal given only AUs: $P(x_t^{VA} \mid x_{\leq t}^{F})$, for all t in the sequence,
- reconstructing the valence and arousal given only audio features: $P(x_t^{VA} \mid x_{\leq t}^{A})$, for all t in the sequence.

To measure the general performance of our model in reconstructing the core affect sequences conditioning on both exteroceptive modalities, and to gauge the predictive capabilities of each individual modality we reconstruct the sequences conditioning as above, and calculate Pearson's correlation coefficient (PCC) and root mean square error

	bimodal	face	audio
PCC valence mean	0.61	0.60	0.50
PCC arousal mean	0.72	0.50	0.69
RMSE valence mean	0.09	0.10	0.14
RMSE arousal mean	0.15	0.23	0.14

Table 7.1: *The PCC and RMSE between the original and reconstructed sequences of the test dataset. Columns represent the conditioning used for reconstruction: both facial expression and audio features (bimodal), only facial expression (face) and only audio features (audio), respectively.*

(RMSE) between the original and reconstructed sequences. The means of the PCC and RMSE over all the sequences in the test dataset are presented in table 7.1 and visualized in figures 7.29 and 7.30 for the bimodal case. A value of PCC closer to 1 is desirable indicating that there is a strong linear relationship between the two sequences, while the RMSE will ideally be close to 0.

In figures 7.31 and 7.32 are visualized the reconstruction and the original sequence for a single sequence from the test set for valence and arousal, respectively, conditioning on both modalities. The same is plotted for both the affective dimensions in 3D in fig. 7.33. Meanwhile, figures 7.34 and 7.35 are reconstructions conditioning only on the facial expression, and similarly for figures 7.36 and 7.37 that are conditioned only on the audio. For the first half of the sequence the model was conditioned on all the three modalities (including affective), while for the remaining part only on on the ones indicated.

The results indicate that as expected, bimodal conditioning confers more information to the model for reconstructing valence and arousal well. The PCC and RMSE are higher and lower, respectively, for the bimodal case. The PCC that theoretically ranges from -1 to 1 lies between 0.5 and 0.75 in all conditioning cases indicating that a strong linear relationship exists between ground truth and the reconstructed sequences, slightly more so, however, for arousal than valence in the bimodal case. Similarly, the RMSE is mostly lower for the bimodal case for both the affective dimensions, and in fig. 7.30 one can identify that most values lie below 0.2 for both valence and arousal, many also nearing 0, and thus indicating good performance on that metric. Unsurprisingly, it would appear from the two metrics that the facial expression is a much better predictor of valence than prosody, while prosodic features are much better predictors for arousal. We find that this is consistent with literature and is a further testament that the model has successfully learned the joint distribution of the three modalities.

Turning to a more qualitative analysis of model performance, the above conclusions are fortified by the graphs of the reconstructed and original sequences. The reconstructed sequences appear to follow the general trend of the original sequence closely, and also capture sudden changes in value. This justifies the PCC values in table 7.1 and effectively means that sudden facial and/or prosodic changes are reflected in the model's reconstruction of affective state (e.g. a sudden smile or smirk will cause a jump in the value of valence). Similarly, fig. 7.35, depicting arousal reconstruction from facial expression, and fig. 7.36, depicting valence reconstruction from prosodic features, show a much lower correspondence to the original sequence in the latter half

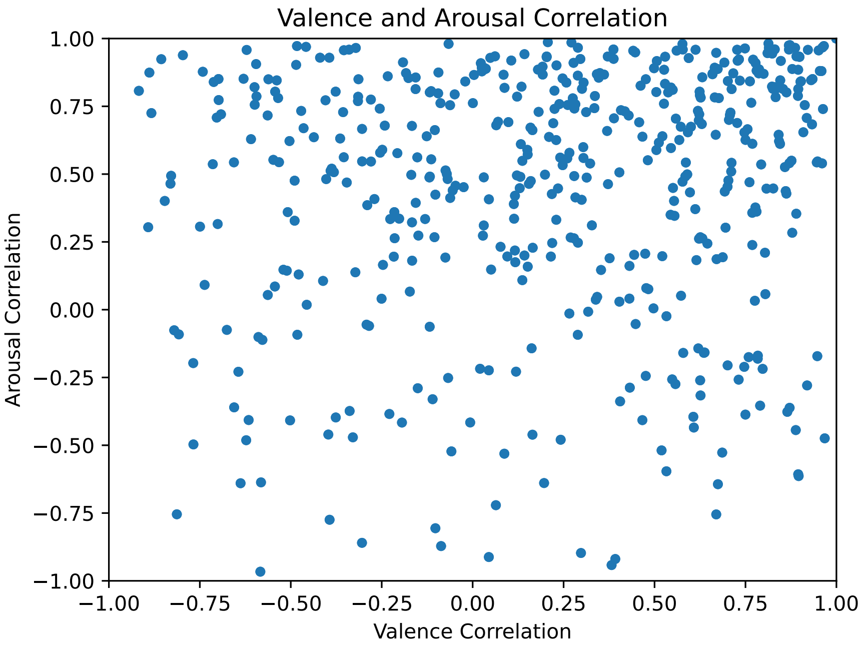


Figure 7.29: *Pearson's correlation coefficient of the reconstructed and original sequences of valence and arousal*

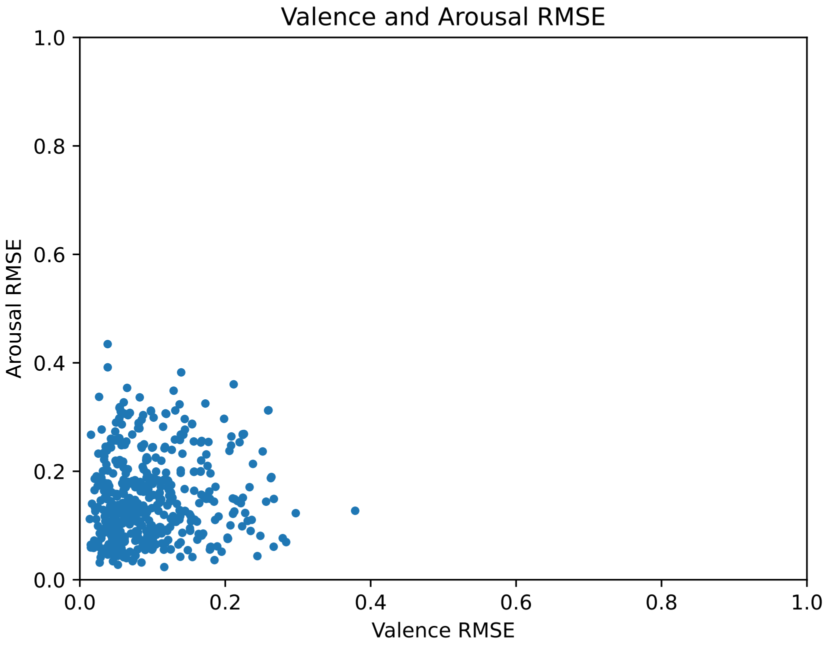


Figure 7.30: *RMSE of the reconstructed and original sequences of valence and arousal*

of the sequence, justifying further the results in table 7.1 and the conclusion that facial expressions are more relevant for inference of valence, and vice versa.

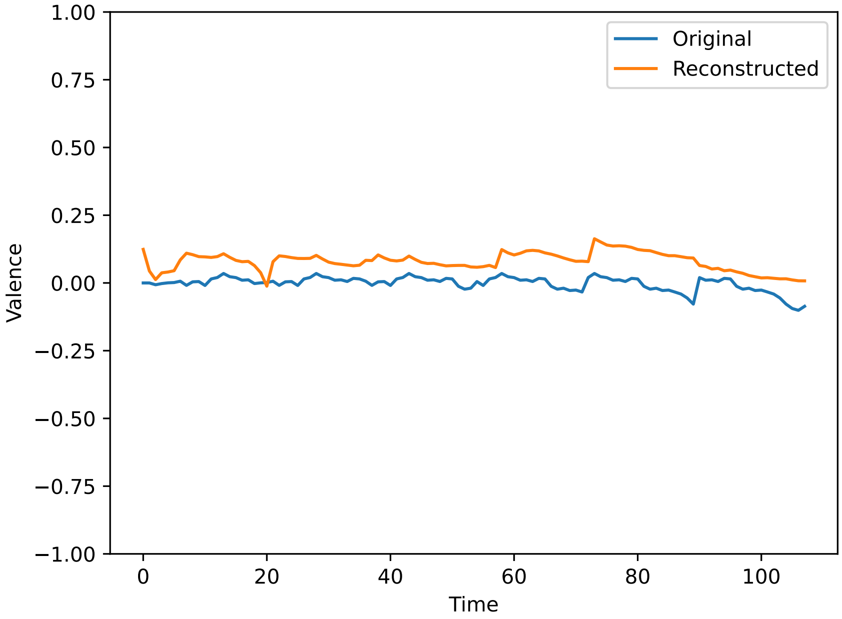


Figure 7.31: *Valence reconstructed and the original sequence for a segment of approximately* 6 *seconds. For the first half of the sequence the reconstructions were formed from all three modalities, and the rest with facial and audio modalities.*

In figures 7.38 and 7.39 are represented two plots of the original and reconstructed trajectories of valence and arousal for two segments of two different test set sessions. The reconstructions were conditioned on both facial expressions and prosodic features. One can notice that the trajectories correspond closely and largely overlap in their placement in the valence/arousal state space, and are furthermore similar in shape. Affective dynamics can be successfully modelled by the Ornstein-Uhlenbeck stochastic process (Oravecz et al., 2011). It can be argued that the observed similarity of the trajectory placement and shape between the original and reconstructed sequences would afford a similar fit of the parameters of the Ornstein-Uhlenbeck process, that is of the home base (i.e. average position) and covariance matrix. This experiment is however left as a future development of this work.

To further qualitatively examine what the model has learned, we condition on the affective state and generate the corresponding facial expression. We then visualise the expression through the generated AUs using the openFACS software package (Cuculo and D'Amelio, 2019). The neutral expression of the openFACS avatar with all the AU activations at zero is shown in fig. 7.41. In figures 7.42, 7.43 and 7.44 are presented the generated expressions for a high value of valence, high value of arousal and a low value of valence, respectively, while keeping the other affective dimension neutral. The expression of high valence is clearly a smile and upon random generation shows consistently high values of the cheek raiser (AU 6), lid tightener (AU 7), upper lip raiser (AU 10), lip corner puller (AU 12), dimpler (AU 14) and lips part (AU 25), a result consistent with the literature definition of a smile. The expression of high arousal resembles a

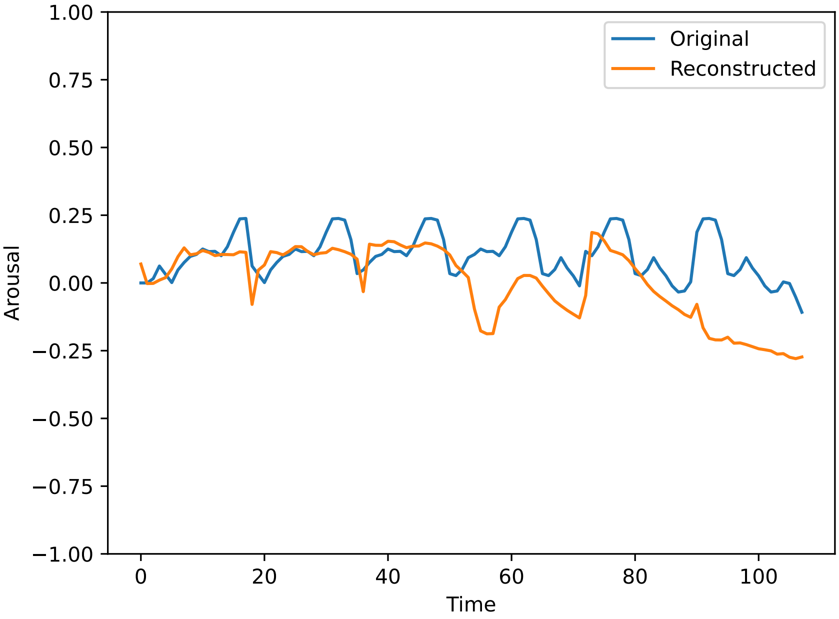


Figure 7.32: *Arousal reconstructed and the original sequence for a segment of approximately* 6 *seconds. For the first half of the sequence the reconstructions were formed from all three modalities, and the rest with facial and audio modalities.*

contemptful expression with a tightening of the eyelids and puckering of lips. Indeed, upon generation the brow lowerer (AU 4), lid tightener (AU 7) and lip suck (AU 28) are consistently strongly activated. Finally, the expression for low valence has similar activations for the eye area, but the lip area is substantially more relaxed. The expression of low arousal with valence remaining neutral doesn't visually differ from the neutral expression which could partly be because the RECOLA dataset affective annotations are scarce in that area, leading the model to not receive enough information during training for proper generation. We find these results roughly conform to what one might expect, and furthermore find the clear-cut smile an intriguing result.

Finally, fig. 7.40 depicts the training losses of the model for different conditionings of the joint distribution used during training. All losses decline in an asymptotic manner which is indicative of convergence, particularly in the case of single modalities. From epoch 50 until the end of training is present a stronger decline in the loss for bimodal and trimodal conditioning than for unimodal. We interpret this as the network first learning to reconstruct the single modalities and later in training focuses on learning the dependencies between them.

It is worth remarking that, in spite of the promising results obtained through the core affect dynamics simulation presented in Section 7.4, these should be taken with caution. In the RECOLA dataset, much like other datasets exploited in the practice of machine-learning oriented affective computing, valence/arousal continuous labelling is provided as produced by external human “labellers”. This is not an innocent step from a conceptual point of view. On the one hand, even in the case the labellers are pro-

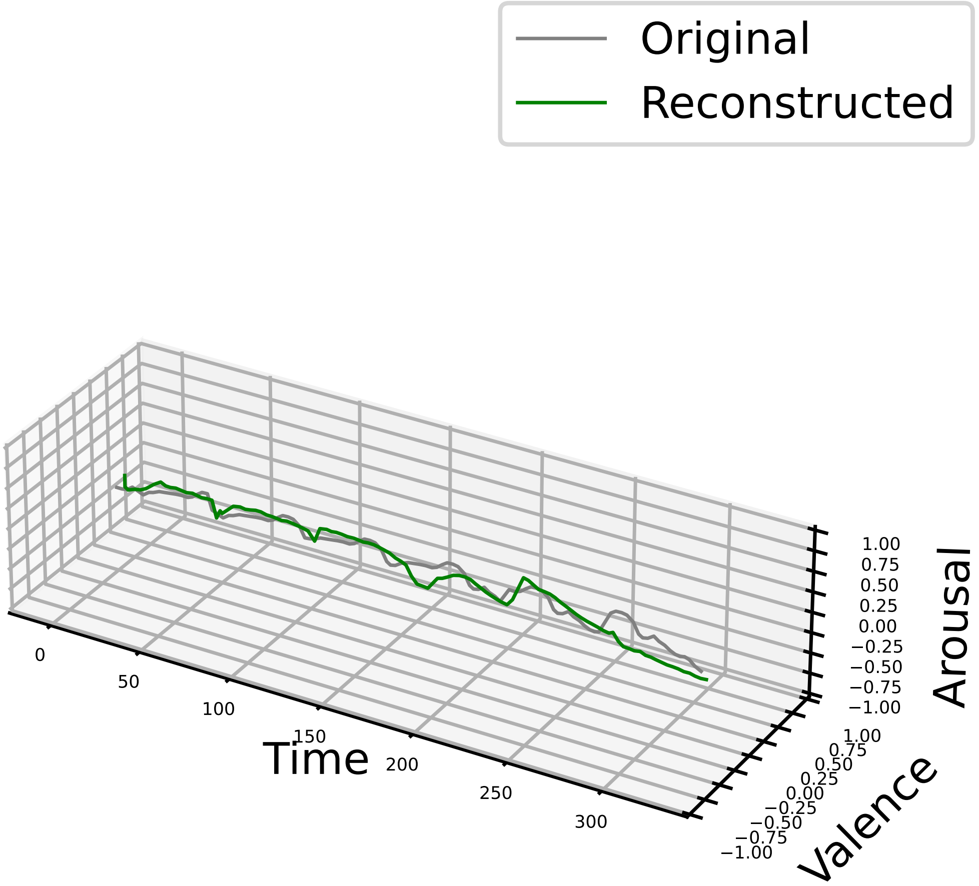


Figure 7.33: *Both valence and arousal reconstructed and the original sequence for a segment of approximately* 6 *seconds. For the first half of the sequence the reconstructions were formed from all three modalities, and the rest with facial and audio modalities.*

fessionally trained (e.g., psychologists), nothing grants that a reliable "ground truth" is eventually produced. Measuring general affective and more specific emotional changes is complex and fraught with difficulties (Quigley et al., 2013). This is a caveat that should always be taken seriously, as scrutinized in-depth, in terms of validity, by Barrett et al. (2019). Unfortunately, in the case of affective behaviour, we are far from the standard conditions that we encounter, for instance in computer vision, where classes of objects and events (e.g., actions) of interest can be objectively categorized; to the point that labelling activity, in that case, can be performed via crowdsourcing. Obviously, this is an intrinsic limitation to the effort of scaling to large datasets.

On the other hand, it might be argued that third person labelling could be replaced by participant self-labelling or self-evaluation, but even in this case results, based on the perceived affect state, can be flawed if experimental conditions are not appropriately controlled. In emotion research, experiential sampling has been proposed (Hoemann et al., 2020a; Christensen et al., 2003). The term "experience-sampling" refers to a set of empirical methods that are designed to allow respondents to document their thoughts,

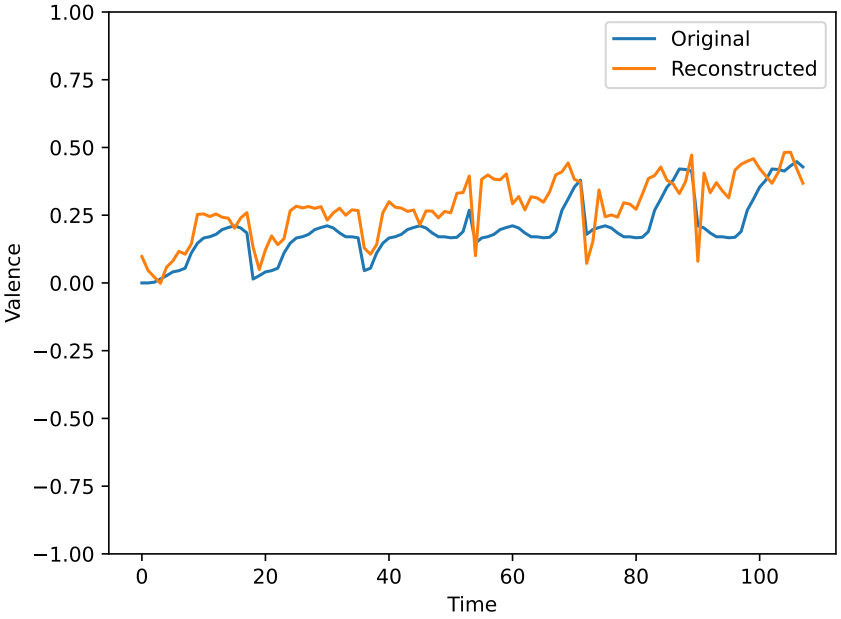


Figure 7.34: *Valence reconstructed and the original sequence for a segment of approximately* 6 *seconds. For the first half of the sequence the reconstructions were formed from all three modalities, and the rest with only the facial modality.*

feelings, and actions outside the walls of a laboratory and within the context of everyday life (the experiment by Kuppens et al., 2010 could be taken as a simplified example). However, experience-sampling is time and resource-intense participants, and provides a challenge to even the most seasoned researcher in the psychology field (Quigley et al., 2013). Clearly, this poses a real problem, probably unsolvable, if the goal is to deliver a large multimodal dataset for experimental purposes

In brief, we have here a critical and cogent issue that the affective computing field will have to confront with in the future to claim the validity of results. In parallel, as previously mentioned, an effort should be put on developing machine learning techniques capable of facing "small" sample datasets.

A final comment is worth concerning the fourth simulation (core affect). In that case, the choice of an implementation model based on the VAE architecture might *prima facie* be at variance with a possible choice of an implementation model relying on a predictive coding scheme. However, this is not the case. As it has been recently shown by Marino (2021), the computation graph for standard predictive coding and that of the VAE relying on direct amortized inference share striking commonalities. These are even more pronounced when VAE parameter learning is performed via iterative amortized inference (Marino et al., 2018).

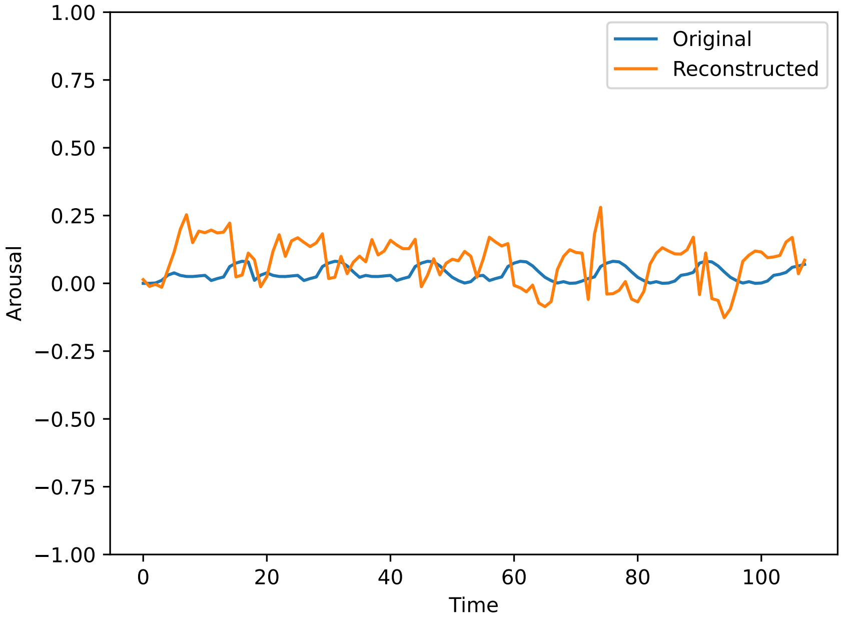


Figure 7.35: *Arousal reconstructed and the original sequence for a segment of approximately* 6 *seconds. For the first half of the sequence the reconstructions were formed from all three modalities, and the rest with only the facial modality.*

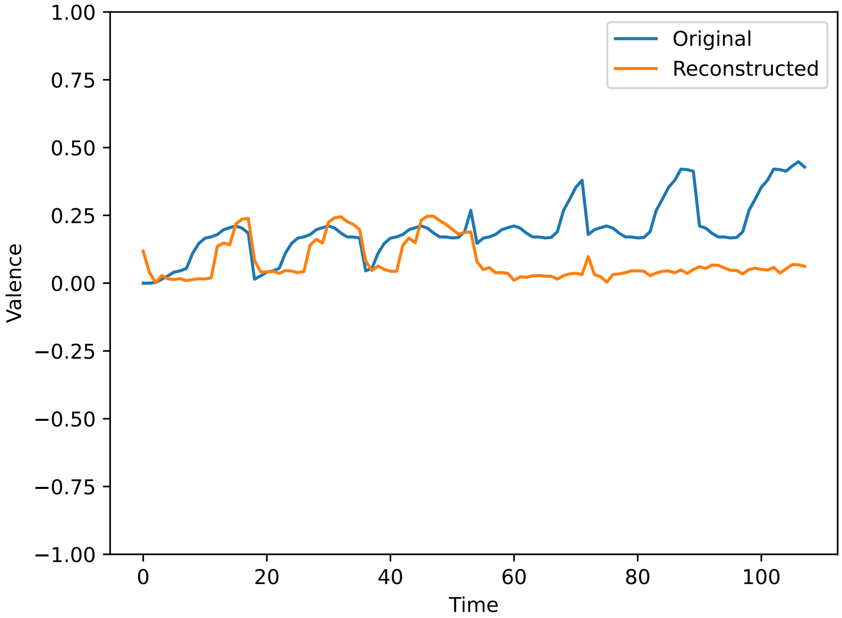


Figure 7.36: *Valence reconstructed and the original sequence for a segment of approximately* 6 *seconds. For the first half of the sequence the reconstructions were formed from all three modalities, and the rest with only the audio modality.*

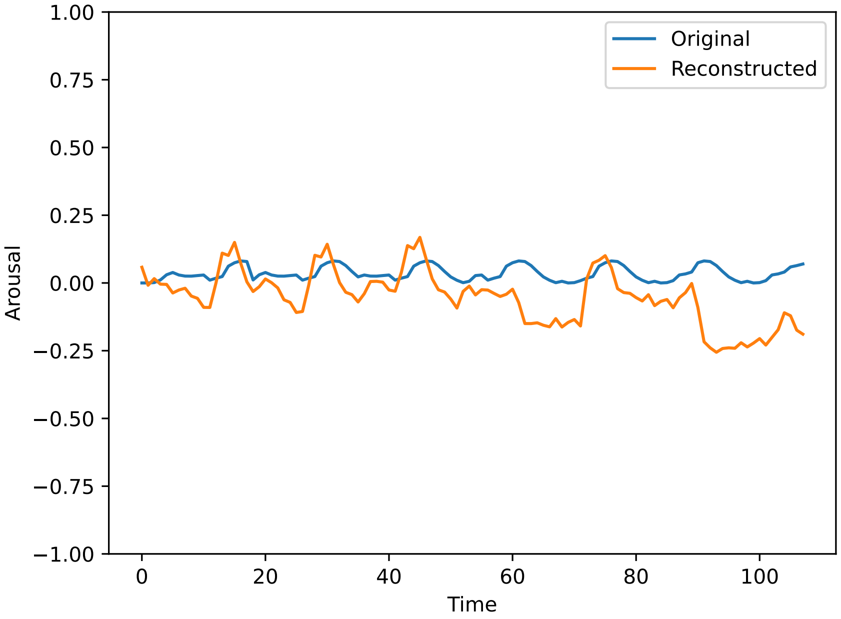


Figure 7.37: *Arousal reconstructed and the original sequence for a segment of approximately* 6 *seconds. For the first half of the sequence the reconstructions were formed from all three modalities, and the rest with only the audio modality.*

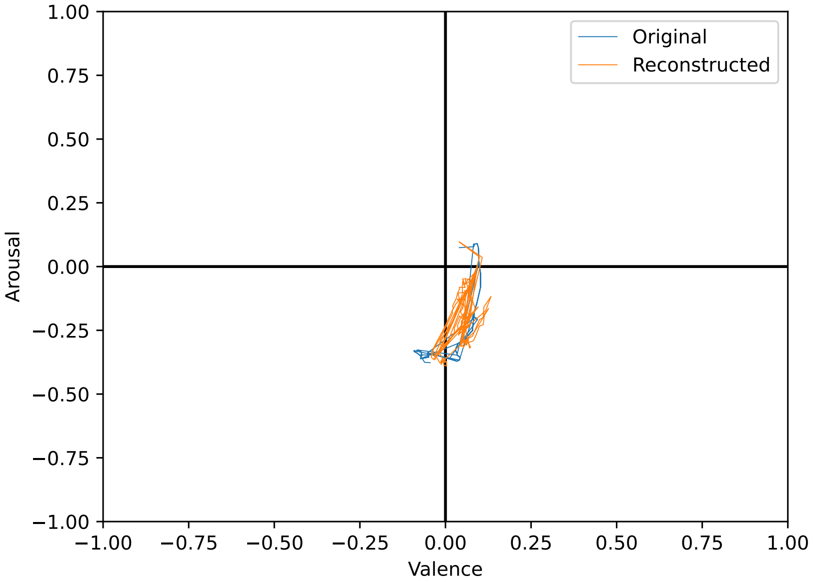


Figure 7.38: *The original and the reconstruction of the valence/arousal trajectories for a portion of a session from the test set (session P30). Both valence and arousal take values in* $[-1, 1]$*, with* 0 *being neutral and* -1 *and* 1 *the lowest and highest values, respectively.*

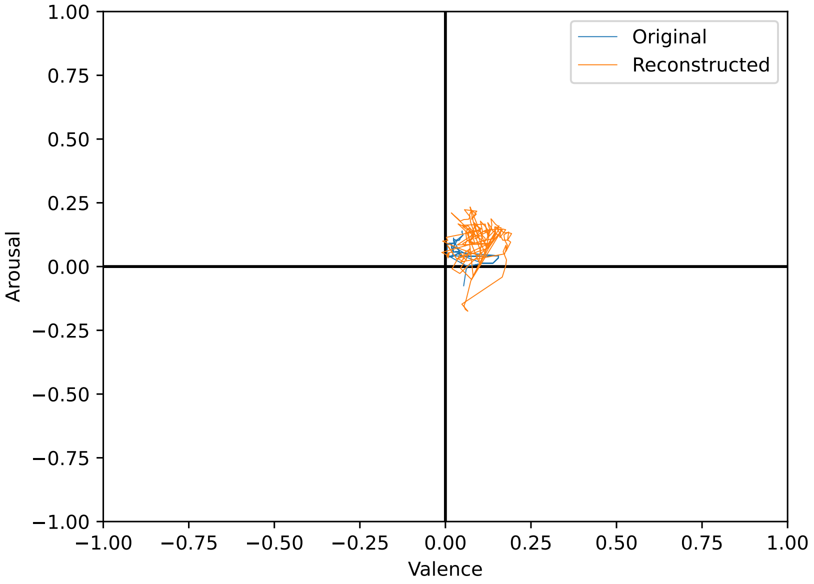


Figure 7.39: *The original and the reconstruction of the valence/arousal trajectories for a portion of a session from the test set (session P56). Both valence and arousal take values in* $[-1, 1]$*, with* 0 *being neutral and* -1 *and* 1 *the lowest and highest values, respectively.*

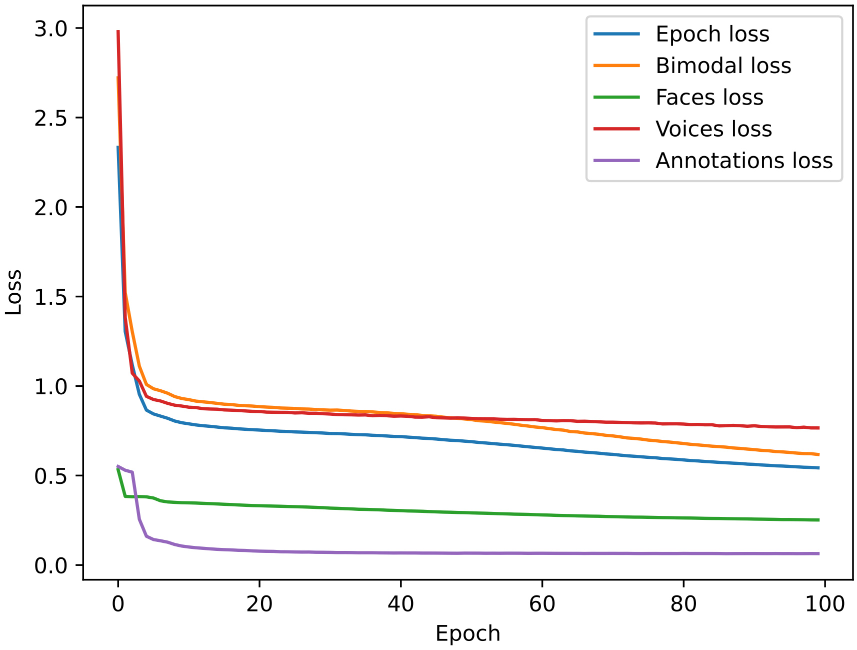


Figure 7.40: *A plot of the training loss as defined in eq. 7.51 over the various conditionings of joint distribution during training. Epoch loss is the mean loss over all conditionings in an epoch. Bimodal loss is the mean of all conditionings with two modalities. The remaining three losses are respectively over the single modalities.*

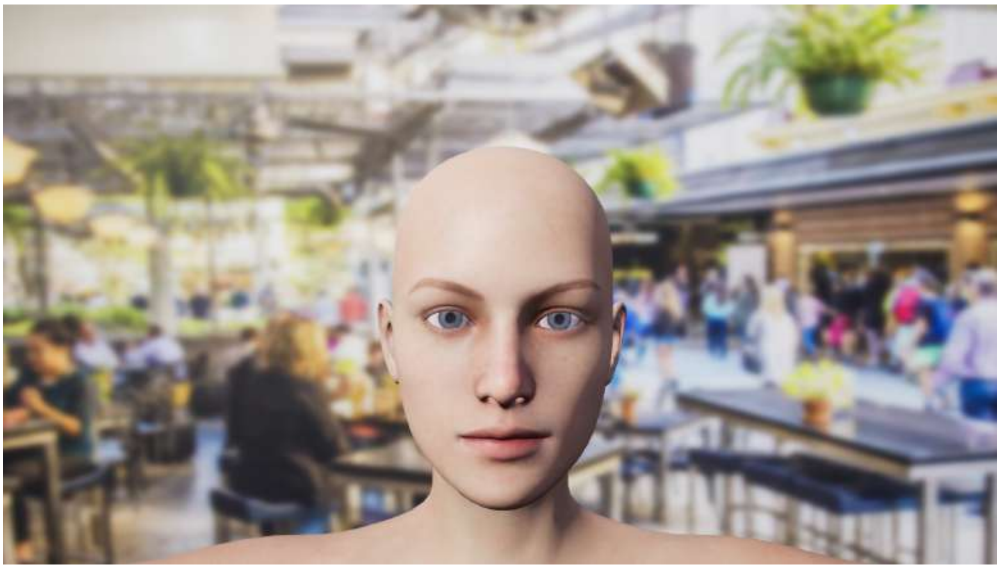


Figure 7.41: *The neutral expression of the openFACS 3D face animation system.*

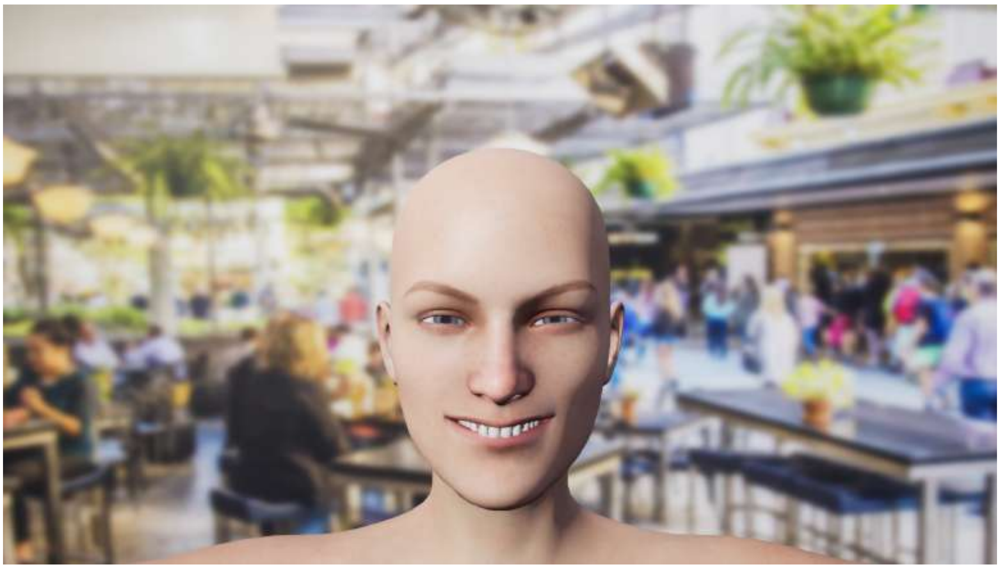


Figure 7.42: *The facial expression generated by the model using a high value of valence and visualised with the openFACS 3D face animation system.*

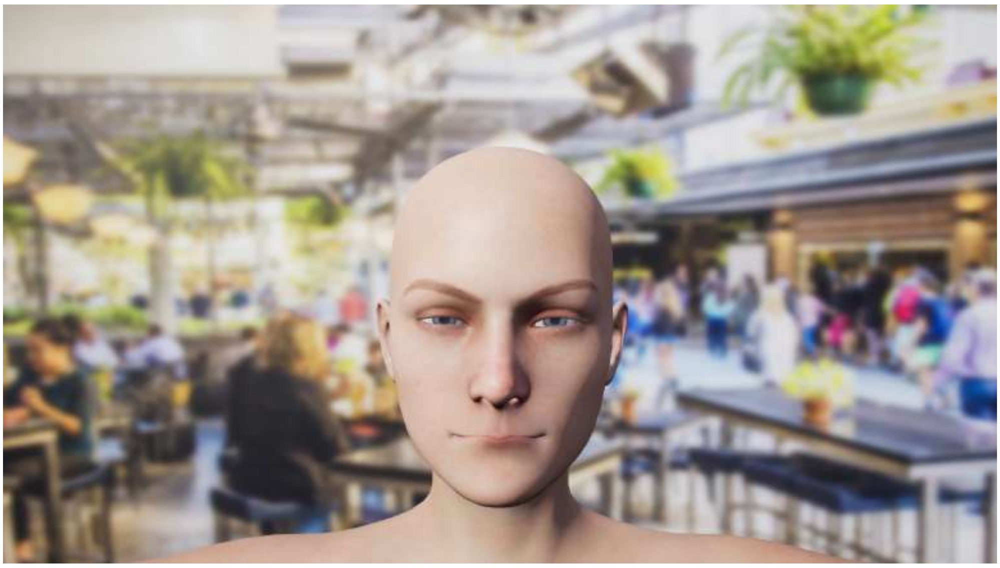


Figure 7.43: *The facial expression generated by the model using a high value of arousal and visualised with the openFACS 3D face animation system.*

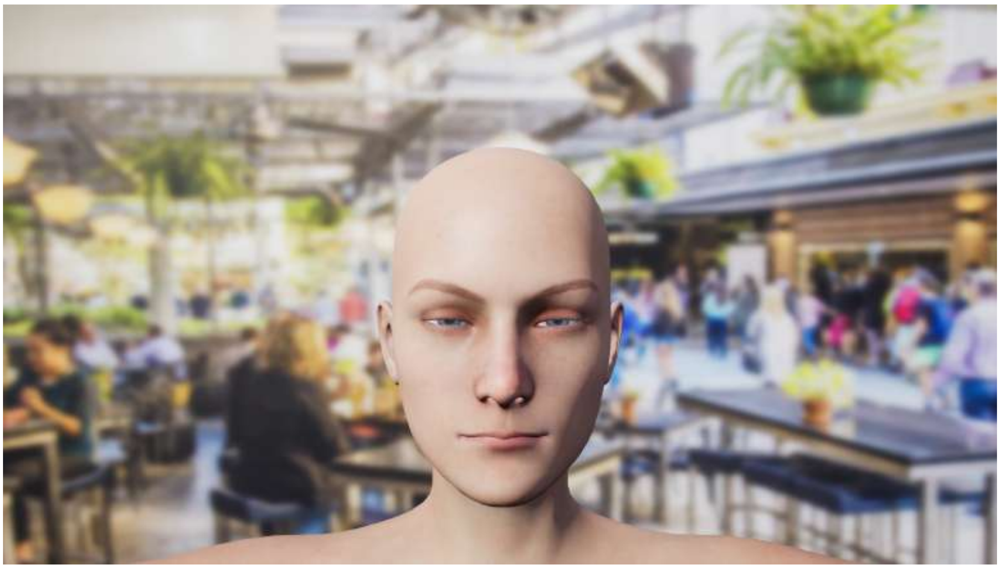


Figure 7.44: *The facial expression generated by the model using a low value of valence and visualised with the openFACS 3D face animation system.*

CHAPTER 8

Conclusions: What next?

In the present dissertation we aimed at devising a theoretical model addressing our initial research question of how language and emotions intertwine with one another.

Indeed, the main contribution achieved is that of a Bayesian model spelling in the language of probability a viable bridge that connects the Conceptual Act Theory constructivist approach to emotions and the Rational Speech Act framework for pragmatic inference based on contextual social reasoning, beyond the literal meanings of words.

To such end, we have carefully scrutinized and drawn from a broad spectrum of studies concerning most recent results achieved in the neurobiology and psychology of emotion and language. In order to take stock of such results in a principled way, our methodological stance was soundly grounded in the tradition of the rational analysis of behaviour.

To the best of our knowledge, this is the first effort in such direction.

In our view, the model offers a radical departure from the vast majority of current approaches in the affective computing field, where the deep learning wave often boils down to foster an end-to-end perspective that is likely to widen the gap with respect to the exciting results achieved in Affective Science, which require a truly interdisciplinary lens that blends questions and tools from far-ranging fields. Indeed, we have shown how a sound theoretical approach might offer a different perspective, both conceptual and practical, for exploiting most recent advances in the machine learning arena.

The work presented here, obviously, suffers from many limitations. The prominent one, at least in our view, is the lack of a unified implementation model mapping *au pair* the theoretical model so far devised. This state of affairs has at least a twofold justification. The first relates to the well known underdeterminacy of the theoretical level of explanation with respect to the implementation or realization level. On the one hand, this issue can be seen as a drawback of the rational analysis approach to modelling in general, and it has often raised criticisms in theoretical neuroscience and psychology, from an epistemological standpoint, concerning the generality of the Bayesian approach in particular. On the other hand, it is an opportunity to shed light on current controversies, markedly those concerning the very nature of the "computational brain". For instance, we have seen how the predictive coding emerges and performs a useful function in the service of Bayesian inference. This suggests that predictive coding and free energy approaches should be understood not as a computational level of explanation, but as a Marrian implementation motif; in neurobiology, a common pattern that can emerge in neural circuits subserving fundamentally different computations. In turn, predictive coding might be subject to different representational schemes. Representational borders blur when resorting to advanced machine learning tools. An implementation pluralism could even be a strategic choice for addressing different application fields. For instance, while we were completing the writing of this thesis, a work has been published by Sennesh et al. (2022), concerning interoception and allostasis in the framework of optimal control theory. This model re-conceptualizes approximate posterior belief as handled by active inference and free energy-based approaches as a feedback controller, according to the interpretation coming from the path-integral control literature. Though aimed to the interoception problem (a sub-component of our model), this could be suitably extended to our purposes. This view could be important for coping with robotics application, where the real body, hardware constraints pose severe restrictions to allowable implementation modelling. Similar concerns can rise for other applications where our model might play a role, e.g., the "hot" fields of empathic image captioning and dialogue models.

The second, more practical, is a current lack of a suitable dataset where different implementation models might be learned and be put into competition with one another. We have largely commented on this tricky issue already, but it is a cogent one for the affective computing realm. Gathering data is costly, and labeling data more so. Labeling the affective dimensions for a large dataset requires an exorbitant investment of time and resources. The alternative, as mentioned before, is to adapt the modelling approach to small sample datasets. Efforts should be invested into developing machine learning techniques capable of learning and generalizing well over datasets of modest size. Finally, current research work we are pursuing is considering these issues, and in particular the labeling/ground truth problem. Words and phrases carry affective connotations, as do nonverbal cues. Labeling the two separately, and then additionally labeling them together may provide insight into how humans convey affect, when these different modes of communication tell the same story, when they diverge, and what are the implications thereof.

Overall, we hope that this humble contribution will pave the way for novel insights and findings in the fascinating conundrum of the entanglement between language and emotions.

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 factorizations directly by inspection of the graph, and without having to perform any analytical manipulations.

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

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

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

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

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

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

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

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

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 $X_i \rightarrow X_j$ 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 $X_i \rightarrow X_{i+1}$, 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 $X_i \rightarrow X_j$ and $X_j \rightarrow X_i$.

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:

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

In other words, $\mathcal{G}$ encodes a set of conditional independence assumptions, called the *local independence*, and denoted by $\mathcal{I}_l(\mathcal{G})$.

However, a BN graph could be defined also in terms of a joint distribution P representable as a set of conditional probability distributions (CPDs) associated with the graph $\mathcal{G}$. Specifically,

Definition A.2. Let P be a distribution over $\mathcal{X}$. We define $\mathcal{I}(P)$ to be the set of *independence assertions* of the form $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ that holds in P.

Given this definition, we can derive that $\mathcal{I}_l(\mathcal{G}) \subseteq \mathcal{I}(P)$, and we say that $\mathcal{G}$ is a *I-map* (independency map) for P. More broadly:

Definition A.3. Let $\mathcal{K}$ be any graph object associated with a set of independencies $\mathcal{I}(\mathcal{K})$. We say that $\mathcal{K}$ is an *I-map* for a set of independencies $\mathcal{I}$ if $\mathcal{I}(\mathcal{K}) \subseteq \mathcal{I}$.

We can now say that $\mathcal{G}$ is an I-map for P if $\mathcal{G}$ is an I-map for $\mathcal{I}(P)$. Let note that, the direction of the inclusion requires that any independence that $\mathcal{G}$ asserts must also holds in P, but not the *vice versa*, that is P could have independencies not reflected in $\mathcal{G}$.

These key concepts allow the compact factorized representation, fundamental for the BN manipulation. Precisely,

Definition A.4. Let $\mathcal{G}$ be a BN graph over the variables $X_1, \ldots, X_n$. We say that a distribution P over the same space factorises according to $\mathcal{G}$ if P can be expressed as a product:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{pa(i)}). \tag{A.1}$$

The individual factors $P(X_i \mid X_{pa(i)})$ are the CPDs or local probabilistic models, and the whole equation is called the *chain rule for BNs*.

Definition A.5. A *BN* is a pair $\mathcal{B} = (\mathcal{G}, P)$ where P factorizes over $\mathcal{G}$, and where P is specified as a set of CPDs associated with $\mathcal{G}$'s nodes. The distribution P is often annotated as $P_{\mathcal{B}}$.

The conditional independence assumptions implied by a BN structure $\mathcal{G}$ allow us to factorize a distribution P for which $\mathcal{G}$ is an I-map into small CPDs as stated in the following theorem (see Koller and Friedman (2009) for the demonstration).

Theorem A.1. *Let $\mathcal{G}$ be a BN structure over a set of RVs $\mathcal{X}$, and let P be a joint distribution over the same space. If $\mathcal{G}$ is an I-map for P, then P factorizes according to $\mathcal{G}$.*

Theorem A.1 proves the factorization of P according to $\mathcal{G}$, but also the converse holds: factorization according to $\mathcal{G}$ implies the associated conditional independencies.

Theorem A.2. *Let $\mathcal{G}$ be a BN structure over a set of random variables $\mathcal{X}$ and let P be a joint distribution over the same space. If P factorizes according to $\mathcal{G}$, then $\mathcal{G}$ is an I-map for P.*

We now move to understand when we can guarantee that an independence $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ holds in a distribution associated with a BN structure $\mathcal{G}$.

Definition A.6. Let $\mathcal{G}$ be a BN structure, and $X_1 \leftrightharpoons \cdots \leftrightharpoons X_n$ a trail in $\mathcal{G}$. Let **Z** be a subset of *observed variables*. The trail $X_1 \leftrightharpoons \cdots \leftrightharpoons X_n$ is *active* given **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 **Z**;
- no other node along the trail is in **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 **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 **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 **Z**, then we are guaranteed that they are, in fact, conditionally independent given **Z**. To prove this it holds

Theorem A.3. *If a distribution P factorizes according to $\mathcal{G}$, then $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$.*

In other words, any independence reported by d-separation is satisfied by the underlying distribution. Also the complementary property, the *completeness*, is desirable. This holds if d-separation detects all possible independencies, that is, given two variables X and Y independents given **Z**, then they are d-separated. To formalize this property, we first introduce the notion of faithful distribution:

Definition A.8. A distribution P is *faithful* to $\mathcal{G}$ if, whenever $(X \perp Y \mid \mathbf{Z}) \in \mathcal{I}(P)$, then d-sep$_G(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$.

In other words, any independence in P is reflected in the d-separation properties of the graph. We can now introduce this result:

Theorem A.4. *For almost all distributions P that factorize over $\mathcal{G}$, that is, for all distributions except for a set of measure zero in the space of CPD 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 factorize 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 $T > 0$).

APPENDIX *B*

Bayesian Inference and Predictive coding

The idea of the brain as a predictive machine (the Bayesian brain) has gained currency in cognitive and theoretical neuroscience in contrast to traditional stimulus-response "feedforward" frameworks (see, Vilares and Kording, 2011; De Ridder et al., 2014; Aitchison and Lengyel, 2017; Chater et al., 2020; Colombo and Seriès, 2020; Yon and Frith, 2021; Marino, 2020; **?**, for general reviews and problems). According to these theories, the Bayesian brain can be conceptualized as a probability machine that constantly makes predictions about the world and then updates them based on what it receives from the senses (De Ridder et al., 2014). This idea has a longstanding history, taking roots in early cybernetics, and has fostered a variety of approaches up to most recent generative models in deep learning (Marino, 2020); its conceptual development is summarized in Figure B.1

The predictive coding model of Rao and Ballard (1999) assumes that the areas comprising the cortical hierarchy implement a hierarchical generative model of the sensory world. The neural activities at each level of the hierarchy represent the brain's internal belief of the hidden causes of the stimuli at a particular abstraction level. Furthermore, the model assumes that the top-down feedback connections from higher to lower order cortical areas convey predictions of lower-level activities. The bottom-up feedforward connections in turn convey prediction errors, calculated as the difference between the top-down predictions and actual activities. The neural activities at each level represent the beliefs about the hidden causes. These are jointly influenced by both the top-down predictions and the bottom-up error signals. Overall, the model assumes the goal of the cortex is to minimize prediction errors across all levels. The above neural operations can be interpreted within a Bayesian framework: the top-down predictions convey prior beliefs based on learned expectations while the bottom-up prediction errors carry evidence from the current input. Predictive coding combines these two sources of in-

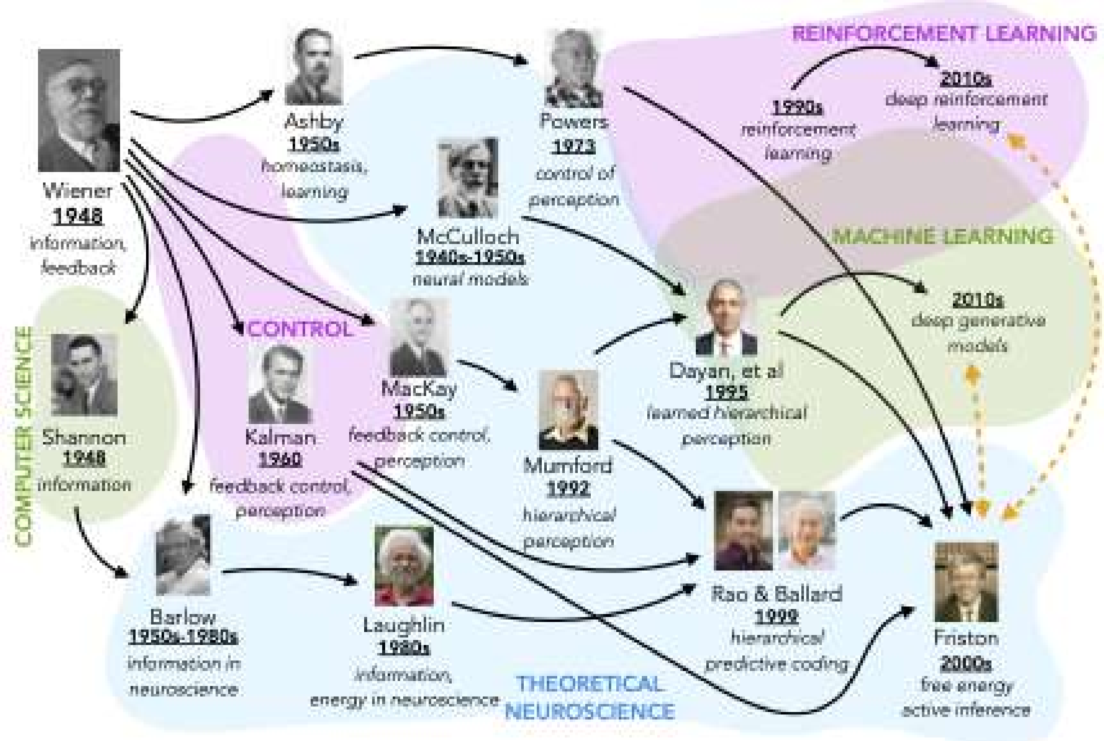

REINFORCEMENT LEARNING
MACHINE LEARNING
CONTROL
COMPUTER SCIENCE
THEORETICAL NEUROSCIENCE
Wiener 1948 information, feedback
Ashby 1950s homeostasis, learning
Powers 1973 control of perception
McCulloch 1940s-1950s neural models
MacKay 1950s feedback control, perception
Kalman 1960 feedback control, perception
Shannon 1948 information
Barlow 1950s-1980s information in neuroscience
Laughlin 1980s information, energy in neuroscience
Mumford 1992 hierarchical perception
Dayan, et al 1995 learned hierarchical perception
Rao & Ballard 1999 hierarchical predictive coding
Friston 2000s free energy active inference
1990s reinforcement learning
2010s deep reinforcement learning
2010s deep generative models

Figure B.1: *Conceptual evolution of the brain as a predictive machine from early cybernetics (top left) to the predictive coding and Bayesian Brain hypothesis in theoretical neuroscience and deep generative models in machine learning (e.g. Restricted Boltzmann Machines, variational autoencoders, etc.) Adapted from* **?**

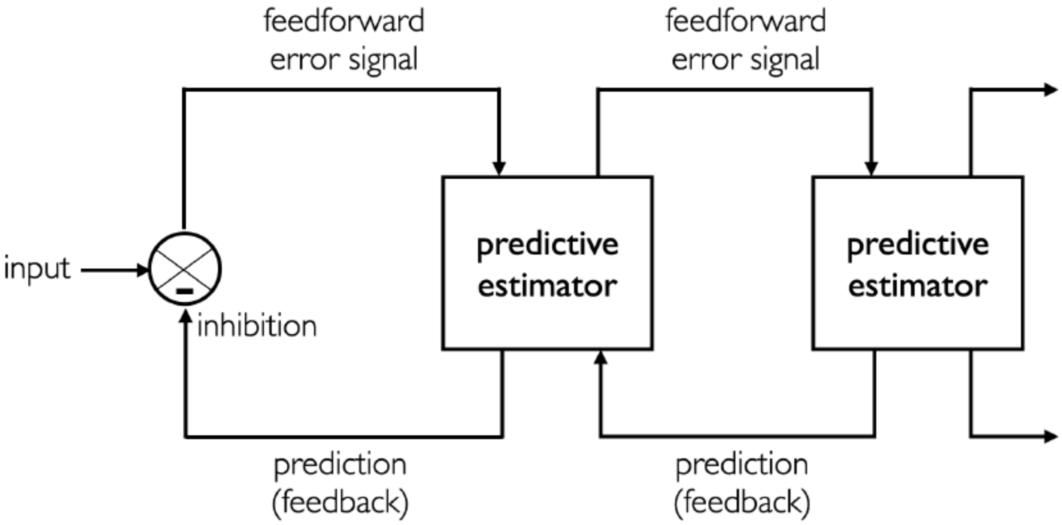


Figure B.2: *The general architecture of the hierarchical predictive coding model. Feedforward connections carry bottom-up prediction errors from the lower level. Feedback connections deliver top-down predictions to the lower level. Adapted from* **?**

formation, weighted according to their reliability (inverse variances or precisions), to compute the posterior beliefs over hidden causes at each level. The objective of minimizing prediction errors across all levels can thus be shown to be equivalent to finding the maximum a posteriori (MAP) estimates of the hidden causes. Rao and Ballard (1997, 1999); **?**.

In its bare essential, a given system (such as the human brain) harbours an internal model of the causes of its sensory input. These are hidden causes in the sense that the system does not have direct access to them but must infer them on the basis of its sensory input and prior knowledge. The model specifies hypotheses about how hidden causes generate input, used to predict what the sensory input to the system will be. Predictions are messages that descend in the internal structure of the system, to be tested against the incoming, ascending sensory signal. Any discrepancy between prediction and sensory signal $\mathcal{O}$ gives rise to prediction error messages that then ascend from the sensors and upward in the system:

$$\mathbf{e} = \mathcal{O} - \texttt{prediction} \tag{B.1}$$

This basic notion of predictive coding provides an efficient message passing scheme because prediction errors carry information about the quality of the prediction and are used to update the model, leading to new predictions:

$$\texttt{new prediction} = \texttt{old prediction} + \mathbf{e} \tag{B.2}$$

The architecture of the PC model is shown in Fig. B.2 generalized to a hierarchical arrangement.

The PC framework can include temporal predictions. Specifically, the network dynamics implements a nonlinear and hierarchical form of Bayesian inference that can be related to the classic technique of Kalman filtering **?**.

More precisely, given the joint distribution $P(\mathcal{O}_{1:T}, \mathcal{Z}_{1:T}^{1:L} \mid \mathcal{U}_{1:T}^{1:L})$, consider the specific case of a linear-Gaussian state space model

$$P(\mathcal{Z}_t \mid \mathcal{Z}_{t-1}, \mathcal{U}_t) = \mathcal{N}(\mathcal{Z}_t \mid \mathbf{A}_t\mathcal{Z}_{t-1} + \mathbf{B}_t\mathcal{U}_t, \mathbf{Q}_t) \tag{B.3}$$

$$P(\mathcal{O}_t \mid \mathcal{Z}_t, \mathcal{U}_t) = \mathcal{N}(\mathcal{O}_t \mid \mathbf{C}_t\mathcal{Z}_t, \mathbf{R}_t) \tag{B.4}$$

The Gaussian structure of the process makes it computationally tractable, and one known example of exact Bayesian filtering for the model is Kalman filtering. In that case the hidden state prediction (one-step-ahead posterior predictive distribution) and the update step can be written in closed form:

- predict:

$$\boldsymbol{\mu}_{t|t-1} = \mathbf{A}_t\boldsymbol{\mu}_{t-1} + \mathbf{B}_{t-1}\mathcal{U}_{t-1} \tag{B.5}$$

$$\boldsymbol{\Sigma}_{t|t-1} = \mathbf{A}_t\boldsymbol{\Sigma}_{t-1}\mathbf{A}_t^\top + \mathbf{Q}_t \tag{B.6}$$

- update:

$$\mathbf{e}_t = \mathcal{O}_t - \mathbf{C}_t\boldsymbol{\mu}_{t|t-1} \tag{B.7}$$

$$\mathbf{S}_t = \mathbf{C}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}_t^\top + \mathbf{R}_t \tag{B.8}$$

$$\mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{C}_t^\top\mathbf{S}_t^{-1} \tag{B.9}$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1}\mathbf{K}_t\mathbf{e}_t \tag{B.10}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{t|t-1} - \mathbf{K}_t\mathbf{C}_t\boldsymbol{\Sigma}_{t|t-1} = \boldsymbol{\Sigma}_{t|t-1} - \mathbf{K}_t\mathbf{S}_t\mathbf{K}_t^\top \tag{B.11}$$

where, to simplify the notation, we have dropped the conditioning on the inputs $\mathcal{U}_t$, and we have assumed the control provided by the upper layer is observed (known).

The Kalman filtering Eqs. B.6 andeq:Kalmobs can be directly derived from Bayes' rule by using a Maximum A Posteriori (MAP) estimate

$$\arg\max_{\mathcal{Z}_t} P(\mathcal{Z}_t \mid \mathcal{Z}_{t-1}, \mathcal{O}_t) = \tag{B.12}$$

$$\arg\max_{\mathcal{Z}_t} \mathcal{N}(\mathcal{O}_t \mid \mathbf{C}_t\mathcal{Z}_t, \mathbf{R}_t)\mathcal{N}(\mathcal{Z}_t \mid \mathbf{A}_t\mathcal{Z}_{t-1} + \mathbf{B}_t\mathcal{U}_t, \mathbf{Q}_t). \tag{B.13}$$

There are two important facts to be noted here that formally express the informal definition of PC discussed above.

1. The first step of Eq. B.11, $\mathbf{e}_t = \mathcal{O}_t - \mathbf{C}_t\boldsymbol{\mu}_{t|t-1}$ accounts for the error signal, namely the difference between the predicted observation $\mathcal{O}_{t|t-1} = \mathbf{C}_t\boldsymbol{\mu}_{t|t-1}$ and the actual observation $\mathcal{O}_t$; in stochastic processes theory, this is known as the innovation.

2. The update for the mean $\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1}\mathbf{K}_t\mathbf{e}_t$ is calculated as is the predicted new mean, $\boldsymbol{\mu}_{t|t-1}$ plus a correction factor, which is the error $\mathbf{e}_t$ weighted by the correction factor $\mathbf{K}_t$, precisely, the Kalman gain matrix computed via $\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}_t^\top\mathbf{S}_t^{-1}$. An intuition of the role of the Kalman correction can be gained by setting $\mathbf{C}_t = \mathbf{I}$. Then, $\mathbf{K}_t = \boldsymbol{\Sigma}_{t|t-1}\mathbf{S}_t^{-1}$. This is the ratio between the covariance of the prior $\boldsymbol{\Sigma}_{t|t-1}$ from the dynamic model, and the covariance of the measurement, $\mathbf{S}_t$ at

time t. This basically means that if the agent has a strong prior and/or very noisy sensors, $|\mathbf{K}_t|$ will be small, and he will place little weight on the correction term. Conversely, if we the agent has a weak prior and/or high precision sensors, then $|\mathbf{K}_t|$ will be large, and the agent will place a lot of weight on the correction term.

Note that the spatiotemporal predictive coding model allows for the possibility that the organism or agent might want to perform internal simulations of the dynamics of the external world (e.g., for planning) by predicting how future states evolve given a starting state (and possibly actions).

PC and the principle of prediction error minimization are closely related to variational inference and learning, which for instance form the basis for variational autoencoders (VAEs) in machine learning research as well as the free energy principle (FEP) in neuroscience as proposed by Friston and colleagues (but see Marino, 2020 for an in-depth overview). FEP is a unified theory of sensory-based cortical function based on predictive coding which estimated the mean and variance of predicted states (Friston and Stephan, 2007; Friston, 2008; Daunizeau et al., 2009). This more general framework was based on the ideas of hierarchical expectation maximization (EM) and empirical Bayes (a method to estimate priors from data). The concept of free energy is treated in detail in Appendix C This model is also biologically motivated along lines similar to those developed by Rao and Ballard, but is intended to address a broader range of empirical evidence.The Friston model differs from the model discussed above, but still obtains the Rao-Ballard protocol.

APPENDIX C

Free energy and variational approximations

FREE energy is a fundamental concept in statistical physics and Bayesian inference, which has gained currency in both theoretical neuroscience and modern machine learning. As to the former, it is at the core of the predicting processing view of the brain; as to the latter, all recent generative deep neural network architectures (e.g., VAE, GAN, etc) rely their learning procedure on the optimization with respect to some form/variation of the free energy, named in the literature ELBO (evidence lower bound).

Use

- $\mathbf{O} = \{O_1, \ldots, O_{j-1}, O_j, O_{j+1}, \ldots, O_m\}$, a collection of *observable* random variables;
- $\mathbf{Z} = \{Z_1, \ldots, Z_{j-1}, Z_j, Z_{j+1}, \ldots, Z_n\}$, a collection of *hidden* or latent random variables

Let $P_{\mathbf{O},\mathbf{Z}}(\mathbf{o}, \mathbf{z})$ be the joint distribution of $\{\mathbf{O}, \mathbf{Z}\}$ and $Q_{\mathbf{Z}}(\mathbf{z})$ an arbitrary probability distribution or density with respect to the Lebesgue measure.

Definition C.1 (Variational free energy)**.**

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

where $\mathcal{U}(\mathbf{Z})$ is the *internal energy* and $\mathcal{H}(\mathbf{Z})$ the Shannon or *canonical entropy* of the collection of r.v. $\mathbf{Z}$.

The following fundamental relation concerning the log evidence of observations, namely, $\log P_{\mathbf{O}}(\mathbf{o})$ and $\mathcal{F}(Q_{\mathbf{Z}})$ holds:

Prop. C.1.

$$\log P_{\mathbf{O}}(\mathbf{o}) = \mathcal{F}(Q_{\mathbf{Z}}) + KL\left(Q_{\mathbf{Z}} \| P_{\mathbf{Z}|\mathbf{O}}\right), \tag{C.1}$$

where $KL\left(Q_{\mathbf{Z}} \| P_{\mathbf{Z}|\mathbf{O}}\right) = \int_{\mathbf{Z}} Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{Q_{\mathbf{Z}}(\mathbf{z})}{P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o})}\, d\mathbf{z}$ *is the relative entropy or Kullback-Leibler divergence (Cover and Thomas, 1991) between* $Q_{\mathbf{Z}}$ *and the posterior distribution* $P_{\mathbf{Z}|\mathbf{O}}$.

Proof. Using the conditional probability definition and taking logs,

$$\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z}) = \log P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o}) + \log P_{\mathbf{O}}(\mathbf{o}) \tag{C.2}$$

which rearranges to

$$\log P_{\mathbf{O}}(\mathbf{o}) = \log P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z}) - \log P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o}). \tag{C.3}$$

Then Equation C.3 grants that

$$\log P_{\mathbf{O}}(\mathbf{o}) = \log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z})}{Q_{\mathbf{Z}}(\mathbf{z})} - \log \frac{P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o})}{Q_{\mathbf{Z}}(\mathbf{z})}, \tag{C.4}$$

and multiplying both sides by $Q_{\mathbf{Z}}(\mathbf{z})$ we obtain

$$Q_{\mathbf{Z}}(\mathbf{z}) \log P_{\mathbf{O}}(\mathbf{o}) = Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z})}{Q_{\mathbf{Z}}(\mathbf{z})} - Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o})}{Q_{\mathbf{Z}}(\mathbf{z})}. \tag{C.5}$$

By integrating with respect to $\mathbf{z}$:

$$\int_{\mathbf{Z}} Q_{\mathbf{Z}}(\mathbf{z}) \log P_{\mathbf{O}}(\mathbf{o})\, d\mathbf{z} = \int_{\mathbf{Z}} Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{o},\mathbf{z})}{Q_{\mathbf{Z}}(\mathbf{z})}\, d\mathbf{z} - \int_{\mathbf{Z}} Q_{\mathbf{Z}}(\mathbf{z}) \log \frac{P_{\mathbf{Z}|\mathbf{O}}(\mathbf{z}|\mathbf{o})}{Q_{\mathbf{Z}}(\mathbf{z})}\, d\mathbf{z}. \tag{C.6}$$

$Q_{\mathbf{Z}}(\mathbf{z})$ is an arbitrary, but normalised distribution, i.e., $\int_{\mathbf{z}} Q_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} = 1$, and using Def.C.1, we obtain (C.1) □

Recall that the basic property of the relative entropy is stated by the following (Cover and Thomas, 1991, Theorem 8.6.1)

Theorem C.2 (Cover and Thomas, 1991, Theorem 8.6.1)**.**

$$KL(Q \| P) \geq 0 \tag{C.7}$$

with equality iff $Q = P$ *almost everywhere (a.e.)*

Then, the free energy $\mathcal{F}(Q_{\mathbf{Z}})$ is a *lower* bound on the log evidence of observations $\log P_{\mathbf{O}}(\mathbf{o})$:

Prop. C.3.

$$\log P_{\mathbf{O}}(\mathbf{o}) \geq \mathcal{F}(Q_{\mathbf{Z}}) \tag{C.8}$$

Proof. Follows directly from Eqs C.1 and C.7. □

Definition C.2 (Mean field approximation). Let $Q_{Z_i}(z_i)$ be the probability distribution of Z_i, the ith element of $\mathbf{Z}$. Then

$$\mathcal{Q} := \left\{ Q_{\mathbf{Z}}(\mathbf{z}) : Q_{\mathbf{Z}}(\mathbf{z}) = \prod_{i=1}^{m} Q_{Z_i}(z_i) \right\}, \tag{C.9}$$

is called the mean-field approximation of distribution Q.

Remark. Clearly, the following trivially holds:

$$\mathbb{E}_{Q_{\mathbf{Z}}}[\mathbf{Z}] = \prod_{i=1}^{m} \mathbb{E}_{Q_{Z_i}}[Z_i]. \tag{C.10}$$

Lemma C.4. *Under the assumption that Q is factorised according to the mean-field approximation $\mathcal{Q}$ (Def. C.2), then*

$$\mathcal{F}(Q_{\mathbf{Z}}) = -KL\left(Q_{Z_j} \| \exp\left\{ \mathbb{E}_{\prod_{\substack{i=1\\ i\neq j}}^{m} Q_{Z_i}} \left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right] \right\} \right) - \sum_{\substack{k=1\\ k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right] \tag{C.11}$$

Proof.

$$\begin{aligned}
\mathcal{F}(Q_{\mathbf{Z}}) &= \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log \frac{P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})}{Q_{\mathbf{Z}}(\mathbf{Z})}\right] \\
&= \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z}) - \log Q_{\mathbf{Z}}(\mathbf{Z})\right] \\
&= \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z}) - \log \prod_{k=1}^{m} Q_{Z_k}(Z_k)\right] \\
&= \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right] - \sum_{k=1}^{m} \mathbb{E}_{Q_{\mathbf{Z}}}\left[\log Q_{Z_k}(Z_k)\right] \\
&= \mathbb{E}_{Q_{Z_j}}\left[\mathbb{E}_{\prod_{\substack{i=1\\ i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right] - \sum_{k=1}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right] \\
&= \mathbb{E}_{Q_{Z_j}}\left[\mathbb{E}_{\prod_{\substack{i=1\\ i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right] - \mathbb{E}_{Q_{Z_j}}\left[\log Q_{Z_j}(Z_j)\right] - \sum_{\substack{k=1\\ k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right].
\end{aligned}$$

Using the log-exp transformation:

$$
\begin{aligned}
&= \mathbb{E}_{Q_{Z_j}}\left[\log\left(\exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right\}\right)\right] - \mathbb{E}_{Q_{Z_j}}\left[\log Q_{Z_j}(Z_j)\right] \\
&\qquad - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right] \\
&= \mathbb{E}_{Q_{Z_j}}\left[\log\left(\exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right\}\right) - \log Q_{Z_j}(Z_j)\right] \\
&\qquad - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right] \\
&= \mathbb{E}_{Q_{Z_j}}\left[\log \frac{\exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right\}}{Q_{Z_j}(Z_j)}\right] - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right] \\
&= -\mathbb{E}_{Q_{Z_j}}\left[\log \frac{Q_{Z_j}(Z_j)}{\exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right\}}\right] - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right] \\
&= -\mathrm{KL}\left(Q_{Z_j} \middle\| \exp\left\{\mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right]\right\}\right) - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right].
\end{aligned}
$$

□

Definition C.3.

$$
\mathcal{U}_j(Z_j) = \mathbb{E}_{\prod_{\substack{i=1\\i\neq j}}^{m} Q_{Z_i}}\left[\log P_{\mathbf{O},\mathbf{Z}}(\mathbf{O},\mathbf{Z})\right] \tag{C.12}
$$

is the expected internal energy of r.v. Z_j

Theorem C.5 (Free energy theorem)**.** *The free-energy is maximised with respect to* $Q^*_{Z_j}(z_j)$ *when*

$$
Q^*_{Z_j}(z_j) \propto \exp\left\{\mathcal{U}_j(Z_j)\right\} \tag{C.13}
$$

Proof. We want to find the optimal approximating distribution

$$
Q^*_{Z_j}(z_j) = \arg\max_{Q_{Z_j}\in Q} \mathcal{F}\left(Q_{\mathbf{Z}}\right)
$$

Use Eq. C.11 to write:

$$
\max_{Q_{Z_j}\in Q} \mathcal{F}\left(Q_{\mathbf{Z}}\right) = \max_{Q_{Z_j}\in Q}\left\{-\mathrm{KL}\left(Q_{Z_j} \middle\| \exp\left\{\mathcal{U}_j(Z_j)\right\}\right) - \sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right]\right\} \tag{C.14}
$$

The term $-\sum_{\substack{k=1\\k\neq j}}^{m} \mathbb{E}_{Q_{Z_k}}\left[\log Q_{Z_k}(Z_k)\right]$ does not depend on Q_{Z_j}, further, Eq. C.7 grants that $\mathrm{KL}\left(Q_{Z_j} \| \exp\left\{\mathcal{U}_j(Z_j)\right\}\right) \geq 0$.

Thus, to maximize $\mathcal{F}\left(Q_{\mathbf{Z}}\right)$ we need to minimize the KL term. Up to constant terms:

$$\max_{Q_{Z_j} \in Q} \mathcal{F}\left(Q_{\mathbf{Z}}\right) \propto \min_{Q_{Z_j} \in Q} \left\{ \mathrm{KL}\left(Q_{Z_j} \| \exp\left\{\mathcal{U}_j(Z_j)\right\}\right)\right\} \tag{C.15}$$

The KL term in the last equation is minimised precisely when the two terms are equivalent a.e., thus the optimal distribution for which $Q^{*}_{Z_j}(z_j) = \arg\max_{Q_{Z_j} \in Q} \mathcal{F}\left(Q_{\mathbf{Z}}\right)$ is

$$Q^{*}_{Z_j}(z_j) = \exp\left\{\mathcal{U}_j(Z_j)\right\} + const = \arg\max_{Q_{Z_j} \in Q} \mathcal{F}\left(Q_{\mathbf{Z}}\right) \tag{C.16}$$

□

C.1 Variational inference

The goal of variational inference is to approximate a conditional density of latent variables given observed variables. The key idea is to solve this problem with optimization. To such end, a family of densities over the latent variables is used, parameterized by free variational parameters. The optimization finds the member of this family, that is, the setting of the parameters, which is closest in KL divergence to the conditional of interest. The fitted variational density then serves as a proxy for the exact conditional density.

From a Bayesian standpoint, the inference problem is to compute the conditional density of the latent variables given the observations, $P(\mathbf{Z} \mid \mathbf{O})$. This conditional can be used to produce point or interval estimates of the latent variables, form predictive densities of new data, and more. Using Bayes rule:

$$P(\mathbf{Z} \mid \mathbf{O}) = \frac{P(\mathbf{O} \mid \mathbf{Z})P(\mathbf{O})}{P(\mathbf{O})} \tag{C.17}$$

Here, for notational simplicity, we have dropped the distribution/density indexes over distribution RVs. The term $P(\mathbf{O}$ is the marginal density of the observations, also called the evidence. Assuming continuous RVs, for generality, the evidence can be obtained through marginalization.

$$P(\mathbf{O}) = \int P(\mathbf{O} \mid \mathbf{Z})P(\mathbf{O}d\mathbf{Z} \tag{C.18}$$

For many models, this evidence integral is unavailable in closed form or requires exponential time to compute. This is why inference in such models is hard. It is also the reason for which it is often called, in analogy with statistical physics, the partition function

Note that all unknown quantities of interest are represented as latent random variables. This includes both distribution parameters, as found in Bayesian models, and latent variables that are "local" to individual data points.

In variational inference, one specifies a family $\mathcal{Q}$ of densities over the latent variables. Each $Q \in \mathcal{Q}$ (e.g, as done in Eq.C.9 above) is a candidate approximation to the exact conditional. The goal is to find the best candidate, the one closest in KL divergence to the exact conditional. Inference the amounts to solving the following optimization problem,

$$Q^*(\mathbf{Z}) = \arg\min_{Q \in \mathcal{Q}} KL(Q(\mathbf{Z}) \| P(\mathbf{Z} \mid \mathbf{O})) \tag{C.19}$$

The $Q^*(\mathbf{Z})$ is the best approximation of the posterior, within the family $\mathcal{Q}$. However, this objective is not computable since requiring the logarithm of the evidence, $\log P(\mathbf{O}$ in Eq. C.17. In fact writing explicitly the KL divergence

$$KL(Q(\mathbf{Z}) \| P(\mathbf{Z} \mid \mathbf{O}) = \mathbb{E}\left[\log Q(\mathbf{Z})\right] - \mathbb{E}\left[\log P(\mathbf{Z} \mid \mathbf{O})\right], \tag{C.20}$$

and expanding the conditional

$$KL(Q(\mathbf{Z}) \| P(\mathbf{Z} \mid \mathbf{O}) = \mathbb{E}\left[\log Q(\mathbf{Z})\right] - \mathbb{E}\left[\log P(\mathbf{Z}, \mathbf{O})\right] + \log P(\mathbf{O}) \tag{C.21}$$

.

Thus, one optimizes an alternative objective that is equivalent to the KL up to an added constant

$$ELBO(Q) = \mathbb{E}\left[\log P(\mathbf{Z}, \mathbf{O})\right] + \log P(\mathbf{O}) - \mathbb{E}\left[\log Q(\mathbf{Z})\right], \tag{C.22}$$

namely, the evidence lower bound (ELBO). By inspecting Eqs. C.28 and C.22,

$$ELBO(Q) = -KL(Q(\mathbf{Z}) \| P(\mathbf{Z} \mid \mathbf{O}) + \log P(\mathbf{O}), \tag{C.23}$$

which shows that maximizing the ELBO is equivalent to minimizing the KL divergence.

Further manipulation gives the following expression for the ELBO and provides more insight

$$\begin{aligned} ELBO(Q) =& \mathbb{E}\left[\log P(\mathbf{Z})\right] + \mathbb{E}\left[\log P(\mathbf{Z} \mid \mathbf{O})\right] - \mathbb{E}\left[\log Q(\mathbf{Z})\right] && \text{(C.24)} \\ & \mathbb{E}\left[\log P(\mathbf{O} \mid \mathbf{Z})\right] - KL(Q(\mathbf{Z}) \| P(\mathbf{Z}). && \text{(C.25)} \end{aligned}$$

The first term is an expected likelihood: it encourages densities that place their mass on configurations of the latent variables that explain the observed data. The second term is the negative divergence between the variational density and the prior: it encourages densities close to the prior. Thus, the variational objective mirrors the usual balance between likelihood and prior. Importantly, ELBO is lower-bounds the (log) evidence. Using Eqs. C.28 and C.22

$$\log P(\mathbf{O}) = KL(Q(\mathbf{Z}) \| P(\mathbf{Z} \mid \mathbf{O}) + ELBO(Q), \tag{C.26}$$

and from Theorem C.7 (positivity of the KL), then

$$\log P(\mathbf{O}) \geq ELBO(Q), \tag{C.27}$$

i.e. ELBO is a lower-bound of the (log) evidence. By comparing the property previously stated through Eq. C.1, which we rewrite here for easiness of inspection

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

and Eq. C.28, it is readily seen that

$$\mathcal{F}\left(Q_{\mathbf{Z}}\right) = ELBO(Q), \tag{C.28}$$

i.e., ELBO is nothing but the variational free energy.

Using the ELBO and the mean-field family, we have cast approximate comnditional inference as an optimization problem. One of the most commonly used algorithms for solving this optimization problem is coordinate ascent variational inference (CAVI). CAVI iteratively optimizes each factor of the mean-field variational density, while holding the others fixed. It climbs the ELBO to a local optimum.

C.2 Relations to the physical free energy

Due to similarity with the free energy concept in statistical physics, in many papers the term is often used in a confusing way. We thus turn to an explanation of the notion of the physical free energy. Suppose that one has a system of N particles, each of which can be in one of a discrete number of states, where the states of the i-th particle are labeled by x_i. As an example, one might make a variety of simplifications and characterize the states of the atoms in a magnetic crystal by whether a given electron in each atom has an “up” spin or a “down” spin. The overall state of the system will be denoted by the vector $\mathbf{x} = \{x_1, x_2, ..x_N\}$. Each state of the system has a corresponding energy $E(\mathbf{x})$. A fundamental result of statistical mechanics is that, in thermal equilibrium, the probability of a state will be given by Boltzmann’s law

$$P(E(\mathbf{x})) = \frac{e^{-\beta E(\mathbf{x})}}{Z(T)} \tag{C.29}$$

Here, T is the temperature, and $Z(T)$ is simply a normalization constant, known as the partition function

$$Z = \sum_{\mathbf{x}\in S} e^{-\frac{E(\mathbf{x})}{T}}, \tag{C.30}$$

where S is the space of all possible states of the system. A substantial part of statistical mechanics theory is devoted to the justification of Boltzmann’s law. On the other hand, if one begins with a joint probability distribution $P(\mathbf{x})$ for some nonphysical system, one can view Boltzmann’s law as a postulate that serves to define an energy for the system, where the temperature can be set arbitrarily, as it simply sets a scale for the units in which one measures energy. We shall take this point of view and set $T = 1$ throughout the rest of this chapter. The *Helmholtz free energy* $\mathcal{F}_H$ of a system is

$$\mathcal{F}_H = -k \log Z \propto -\log Z \tag{C.31}$$

This free energy is a fundamentally important quantity in statistical mechanics, because if one can calculate the functional dependence of $\mathcal{F}_H$ on quantities like a macroscopic magnetic field H or temperature T, then it is easy to compute experimentally measurable quantities like the response of the system to a change in H or T. Physicists have therefore devoted consider- able energy to developing techniques which give good approximations to $\mathcal{F}_H$. One important technique is based on a variational approach. Suppose again that $P(\mathbf{x})$ is the true probability distribution of the system, and obeys Boltzmann’s law (Eq. C.29). It may be that even if we know $P(\mathbf{x})$ exactly, it is of a form that makes the computation of $\mathcal{F}_H$ difficult. We therefore introduce a "trial" probability distribution $Q(\mathbf{x})$, which should, of course, be normalized and obey $0 \leq Q(\mathbf{x}) \leq 1$ for all $\mathbf{x}$, and a corresponding variational free energy, which is also sometimes called the *Gibbs free energy* defined by

$$\mathcal{F}_G(Q) = U(Q) - H(Q), \tag{C.32}$$

where $U(Q)$ is the variational average energy

$$U(Q) = \sum_{\mathbf{x} \in S} Q(\mathbf{x}) E(\mathbf{x}) \tag{C.33}$$

and $H(Q)$ is the variational entropy

$$H(Q) = -\sum_{\mathbf{x} \in S} Q(\mathbf{x}) \ln Q(\mathbf{x}) \tag{C.34}$$

We measure entropy using the natural logarithm instead of the base-2 logarithm in order to be consistent with the physics literature. It follows directly from our definitions that

$$\mathcal{F}_G(Q) = \mathcal{F}_H + KL(Q \| P), \tag{C.35}$$

and that

$$\mathcal{F}_G(Q) \geq \mathcal{F}_H \tag{C.36}$$

with equality precisely when $Q(\mathbf{x}) = P(\mathbf{x})$. Thus, minimizing the variational Gibbs' free energy $\mathcal{F}_G(Q)$ with respect to trial probability functions $Q(\mathbf{x})$ is therefore an exact procedure for computing $\mathcal{F}_H$ and recovering $P(\mathbf{x})$. It can be seen from Eqs C.36 and C.22 that

$$\begin{aligned}
\mathcal{F}\left(Q_{\mathbf{Z}}\right) &= && \text{(C.37)}\\
ELBO(Q) &= -\mathbb{E}\left[\log Q(\mathbf{Z})\right] + \mathbb{E}\left[\log P(\mathbf{Z}, \mathbf{O})\right] + \log P(\mathbf{O}) && \text{(C.38)}\\
&= -KL(Q \| P) + \log P(\mathbf{O}) && \text{(C.39)}\\
&= -KL(Q \| P) - \mathcal{F}_H && \text{(C.40)}\\
&= -\mathcal{F}_G(Q) && \text{(C.41)}
\end{aligned}$$

This result tells us that our variational inference $\mathcal{F}\left(Q_{\mathbf{Z}}\right)$ or ELBO is equivalent to the *negative Gibbs' free energy*. Thus the maximization of $\mathcal{F}\left(Q_{\mathbf{Z}}\right)$ corresponds to the minimization of the Gibbs' free energy $\mathcal{F}_G(Q)$ in statistical physics systems.

APPENDIX *D*

A quick tour of intrinsic networks

The brain can be thought of as one large structural network showing continuous, intrinsic activity. Empirically, an intrinsic network is defined as those areas whose low frequency blood oxygen level-dependent (BOLD) signal correlates over time when a person is "at rest" (i.e., not being probed with an external stimulus).

As a large network the brain can be further subdivided in a set of large-scale subnetworks as shown in Figure E.6.

Each such large-scale network is a collection of interconnected brain areas, or nodes, that are linked together to perform circumscribed functions. The nodes of a network share dense interconnections among its constituent nodes when compared to connections with nodes that form other brain networks.

For instance, when coming to what is traditionally defined as cognitive control - the coordination of mental processes and action in accordance with current goals and future plans -, one might consider the six fundamental large-scale networks anchored in the prefrontal cortex (PFC, Menon and D'Esposito (2022)). These are presented in Figure D.1with their core nodes.

The main networks of interest here are the **default mode network**, (DMN, sometimes called the mentalizing network, the construction network, or semantic knowledge network), the **salience network** (SN, which bears a strong resemblance to the "ventral attention" and "multimodal" networks) and its "fraternal twin", the fronto-parietal or **central executive-control network** (CEN) that in their bare essential are presented in Figure D.2.

The SN and DMN contain a large proportion of the "rich club hubs" of the brain. Rich club hubs are the most highly connected brain areas and have been identified using diffusion tensor imaging of white matter tracts in humans and reviewing tract tracing studies in monkeys. Different intrinsic networks such as sensory networks overlap

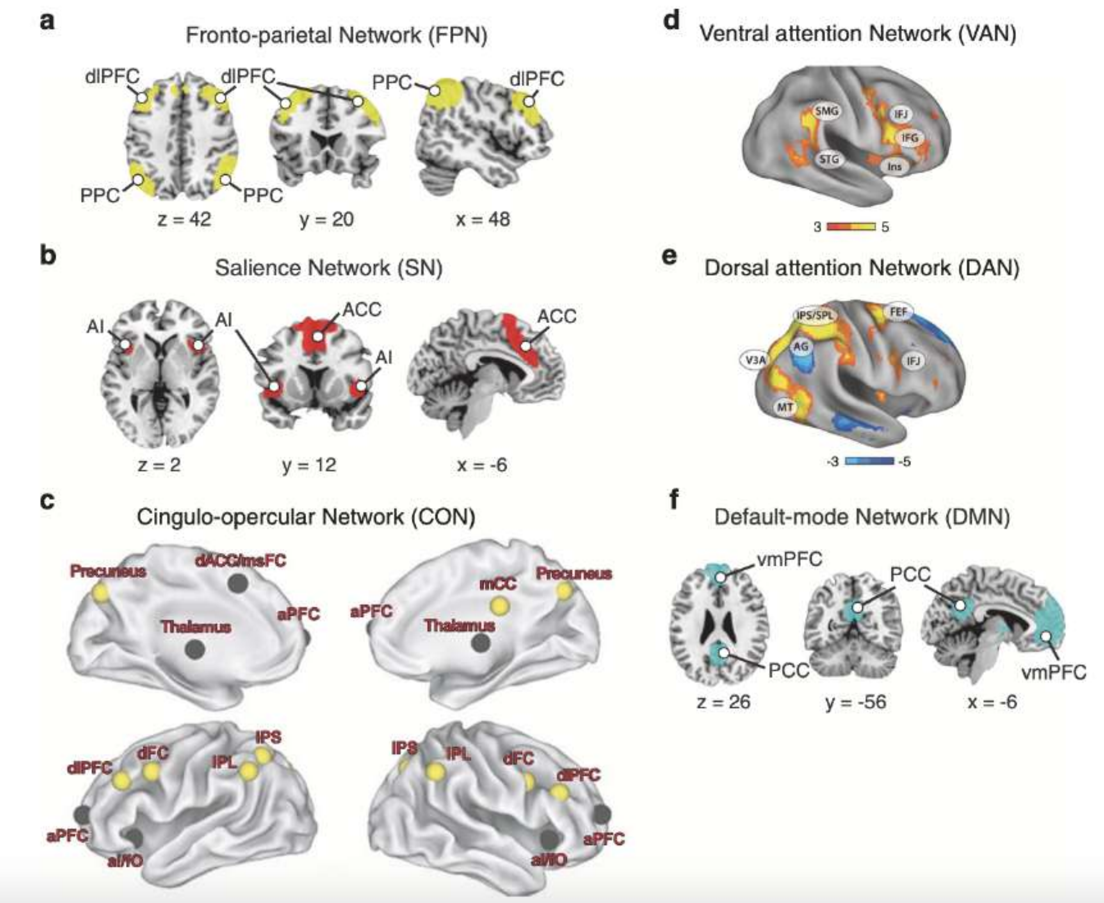


Figure D.1: *Large scale functional networks anchored in the prefrontal cortex (PFC) with key nodes.* ***a*** *Fronto-parietal network (FPN), with key nodes in dorsolateral PFC (dlPFC) and posterior parietal cortex (PPC).* ***b*** *Salience network (SN), with key nodes in anterior insula (AI) and dorsal anterior cingulate cortex (ACC).* ***c*** *Cingulo-opercular network (CON, black) with key nodes in anterior insula/frontal operculum (aI/fO), dorsal ACC and medial superior frontal cortex (dACC/msFC), anterior PFC (aPFC) and thalamus, as distinguished from the FPN (yellow).* ***d*** *Ventral attention network (VAN), with key nodes in insula (Ins), inferior frontal junction (IFJ), supramarginal gyrus (SMG), and superior temporal gyrus (STG).* ***e*** *Dorsal attention network (DAN), with key nodes in frontal eye fields (FEF), inferior frontal junction (IFJ), intra-parietal sulcus and superior parietal lobule (IPS/SPL), angular gyrus (AG), visual area 3A (V3A), and middle temporal visual area (MT).* ***f*** *Default mode network (DMN), with key nodes in ventromedial PFC (vmPFC) and posterior cingulate cortex (PCC). From Menon and D'Esposito (2022)*

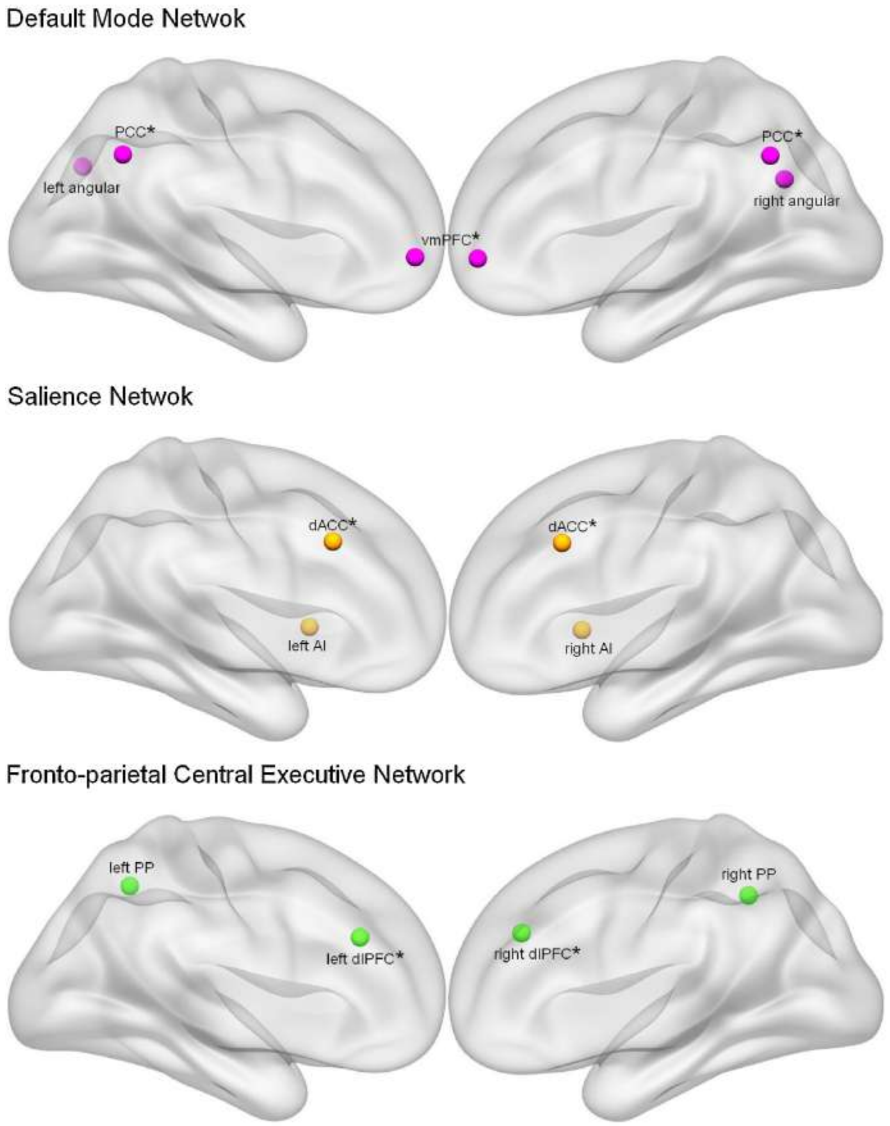


Figure D.2: *The three intrinsic networks that are of particular interest in relation to this thesis identified through their core regions (seeds). (1) The default mode network (DMN), comprised of the posterior cingulate cortex (PCC), the ventromedial prefrontal cortex (vmPFC) and the angular gyri; DMN is involved in introspective processing, social cognition (e.g., theory of mind, moral cognition), and affective cognition. (2) The salience network (SN), including the anterior insula (AI) and the dorsal anterior cingulate cortex (dACC); the SN is involved in the detection of salient stimuli and the initiation of cognitive control by influencing activation of the central executive network and the DMN. In both adults and children, these three networks work simultaneously during executive tasks, social tasks, and cognitive control tasks. (3) The frontal-parietal central executive network (CEN), anchored in the dorsolateral prefrontal cortex (dlPFC) and posterior parietal cortex (PPC), plays an important role in executive functions. Abbreviations: AI, anterior insula; dACC, dorsal anterior cingulate cortex; dlPFC, dorsolateral prefrontal cortex; PCC, posterior cingulate cortex; PPC, posterior parietal cortex; vmPFC, ventromedial prefrontal cortex. * These seeds are medial.*

in these hubs, communicating with each other through them. Indeed, structural and functional imaging in humans indicates that rich club hubs are connector nodes for intrinsic networks.

As a matter of fact, intrinsic networks cooperate with one another to build up mental functions and no one-to-one mapping between any of former and any of the latter can be plausibly defined (Barrett and Satpute, 2013).

This observation is on pair with the very fact that mental categories studied in neuroscience as rendered by the terms like “attention”, “memory”, “decision making”, “emotion” are of limited usefulness for studying and describing the relationship between brain and behavior. In particular, the functions supported by the neuroarchitecture do not align themselves well with the standard decomposition. In other words, the standard “faculty” decomposition would require an organization that is relatively modular. Instead, that fundamental principles of the neuroarchitecture indicate that it is not. In particular, the neuroarchitecture is not additive in the sense that new components are added atop an ancestral organization. Which exactly is the point of the SM.

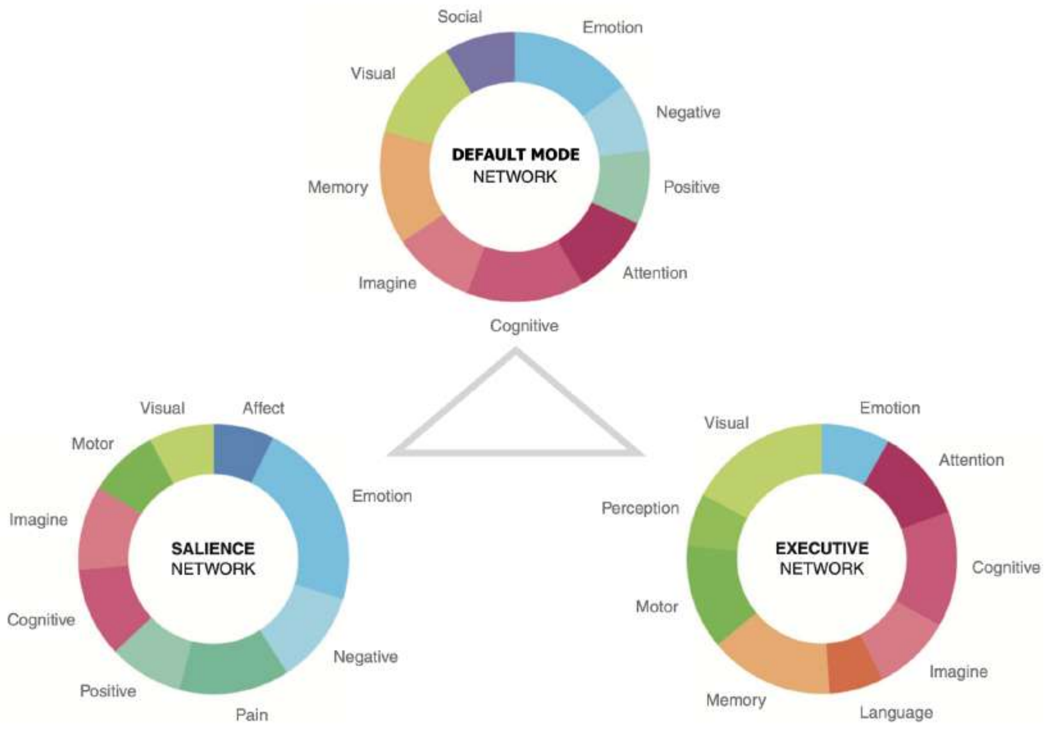


Figure D.3: *Fundamental intrinsic network cooperation for “building” mental functions. Each pie chart depicts the relative frequency with which various mental functions are discussed in the context of increased activation within the DMN (top), the SN (bottom left) and the CEN network (bottom right), obtained using the Neurosynth database including over 6000 publications from over 50 journals. Adapted from Barrett and Satpute (2013)*

D.1 Default Mode Network

Early brain imagining studies have discovered that during periods when participants are not engaged in cognitive or other external tasks, certain brain areas tend to become more activated. This network of areas was labelled the Default Mode Network (Raichle et al., 2001; Buckner et al., 2008).

Two of the main aspects of the DMN are assumed to be self-referential processing and autobiographic memory. It is particularly active when one is thinking about oneself, in past, current or future situations (e.g. anticipating outcomes, planning). As such, the DMN has also been connected to mind-wandering. Such mind-wandering or ruminations should be considered as specific instances of the general function of the DMN to cognitively simulate future scenarios.

Prima facie, decreased activation of the DMN during task engagement may be a sign of reduced mind-wandering. In light of the foregoing, findings of increased activation of the DMN during social activities also make sense because in those situations, thinking about oneself and possible outcomes of one's behaviour are relevant for adequate social performance (e.g. what is the potential impact of the things I am saying?).

The largest brain areas associated with the DMN are the posterior cingulate cortex, the medial prefrontal cortex, the angular gyrus.

More extensively regions included are: ventral medial prefrontal cortex (**vmPFC**, Broadman areas 24, 10 m/10 r/10 p, 32ac), posterior cingulate/retrosplenial cortex (**PCC** / **Rsp**, 29/30, 23/31), inferior parietal lobule / angular gyrus (**IPL/AG**, 39, 40) lateral temporal cortex (**LTC**, 21) dorsal medial prefrontal cortex (**dmPFC**, 24, 32ac, 10p, 9) hippocampal formation (**HF+**, Hippocampus proper plus entorhinal cortex, EC, and surrounding cortex, e.g., parahippocampal cortex, PH).

Probing the functional anatomy of the network in detail reveals that it is best understood as multiple interacting subsystems. The medial temporal lobe subsystem provides information from prior experiences in the form of memories and associations that are the building blocks of mental simulation. The medial prefrontal subsystem facilitates the flexible use of this information during the construction of self-relevant mental simulations. These two sub-systems converge on important nodes of integration including the PCC.

It is worth remarking that although this network has always been seen as unitary and associated with the resting state, a new deconstructive line of research is pointing out that the DMN could be divided into multiple subsystems supporting different functions. By now, it is well known that the DMN is not only deactivated by tasks, but also involved in affective, mnestic, and social paradigms, among others. Nonetheless, it is starting to become clear that the array of activities in which it is involved, might also be extended to more extrinsic functions. Indeed, the most recent developments in the research of the DMN are indicating that such network, far from being a monolithic entity, consists of multiple systems with intersecting functions and anatomies. It has been observed that the DMN is not deactivated by any task, as self-referential and emotional paradigms activated it. Since those observations, many further functions were shown to be associated to this network. Other than self- referential and emotional processes, the DMN turned out to be related to memory and mental time-travel, mental simulation and scene construction, theory of mind (ToM) and social cognition, moral

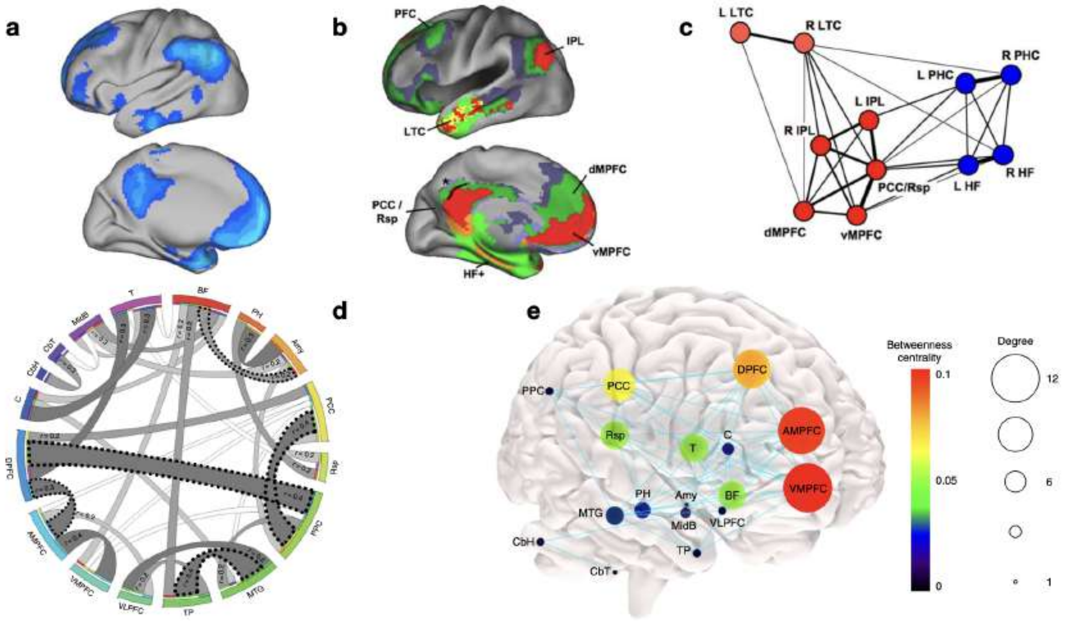


Figure D.4: *The Default Mode Network. a) Medial and lateral surface of the DMN as originally identified in a meta-analysis that mapped brain regions more active in passive as compared to active tasks; blue represents regions most active in passive task setting. b) Hubs and subsystems within the DMN mapped using functional connectivity analysis. c) The regions of the DMN are graphically represented with lines depicting correlation strengths; the structure of the default network has a core set of regions (red) that are all correlated with each other; LTC is distant because of its weaker correlation with the other structures. d) A recent graph representation of the partial correlations between regions of interest in the functional space (connections with partial correlation above 0.2 are depicted; darker grey tones represent stronger connections); statistically significant partial correlations have a dashed borderline; subcortical structures are also included. e) Graph theory analysis of structural connectivity: the node size represents node degree and the node colour illustrates node betweenness centrality; the edges denote presence of structural connection. DPFC dorsal prefrontal cortex, PPC posterior parietal cortex, VLPFC ventrolateral prefrontal cortex, Rsp retrosplenial cortex, MTG middle temporal gyrus, PCC posterior cingulate cortex, C caudate, DPFC dorsal prefrontal cortex, AMPFC antero-median prefrontal cortex, VMPFC ventro-median prefrontal cortex, TP temporal pole, BF basal forebrain, T thalamus, PH parahippocampal region, CbH cerebellar hemisphere, CbT cerebellar tonsil, Amy amygdala, MidB midbrain. Adapted from Buckner et al. (2008); Alves et al. (2019).*

judgment, semantic processing and reward mechanisms (but see Mancuso et al., 2021 for a discussion)

D.2 Central Executive Network

The key brain areas associated with the CEN are the dorsolateral prefrontal cortex (**DLPFC**) and the posterior parietal cortex (**PPC**).

More in detail, The CEN is comprised of the dorsolateral (dlPFC, BA 9/46) and dorsomedial PFC (dmPFC,BA 6), the supramarginal gyrus (BA 40) in posterior parietal cortex and subcortical regions including the dorsal caudate and anterior thalamus.

It is worth noting that, somehow confusingly, in the literature analogous labelling is often adopted to describe topographically or functionally similar neural networks: central executive network (CEN), cognitive control network (CCN), executive control network (ECN), executive network (EN), frontoparietal network (FPN), working memory network (WMN), task positive network (TPN).

The CEN consists of an array of strongly interconnected brain areas that are mainly active when engaging in external cognitive processing. That is, when one is engaged in tasks that require the active maintenance of information (or task) in working memory, a switching between task requirements and the inhibition of irrelevant information. In other words, the CEN becomes activated in situations that require focus or concentration. Notably, the CEN and the DMN often show contrasting patterns of activation. If the CEN becomes more active, the DMN decreases in activation, and vice versa.

A tight link between lateral PFC and PPC is supported by the demonstration of strong bidirectional anatomical connections with each other, as well as similar profiles of neuronal responses (Menon and D'Esposito, 2022).

Based on a vast amount of empirical evidence, the DLPFC has been referred to as the seat of working memory. Furthermore, the strength of the pathways between the DLPFC and PPC has been associated with intelligence—the ability to effectively deal with complex or novel problems and situations.

D.3 Salience Network

The main brain regions of this network are the anterior insula cortex (**AIC**) and the anterior cingulate cortex (**ACC**). Prominent subcortical nodes include the amygdala, substantia nigra, ventral tegmental area, dorsomedial thalamus, hypothalamus, and periaqueductal gray

Several accounts posit the SN role as that of switching between other brain networks, particularly between the DMN and the CEN (Menon, 2015). Figure D.7 presents the basic Network Switching Model hypothesis.

Accordingly, the SN is involved in the continuous shifting between task-related versus non-task-related and self-referential processing. This switching may be related, but is not restricted, to the switching between task concentration and mind-wandering. The network received its name due to its presumed core function, which is detecting the salience of stimuli/events. Salience, in this context, is every stimulus, *internal or external*, that the system signals as worthy of further attention and processing. The SN

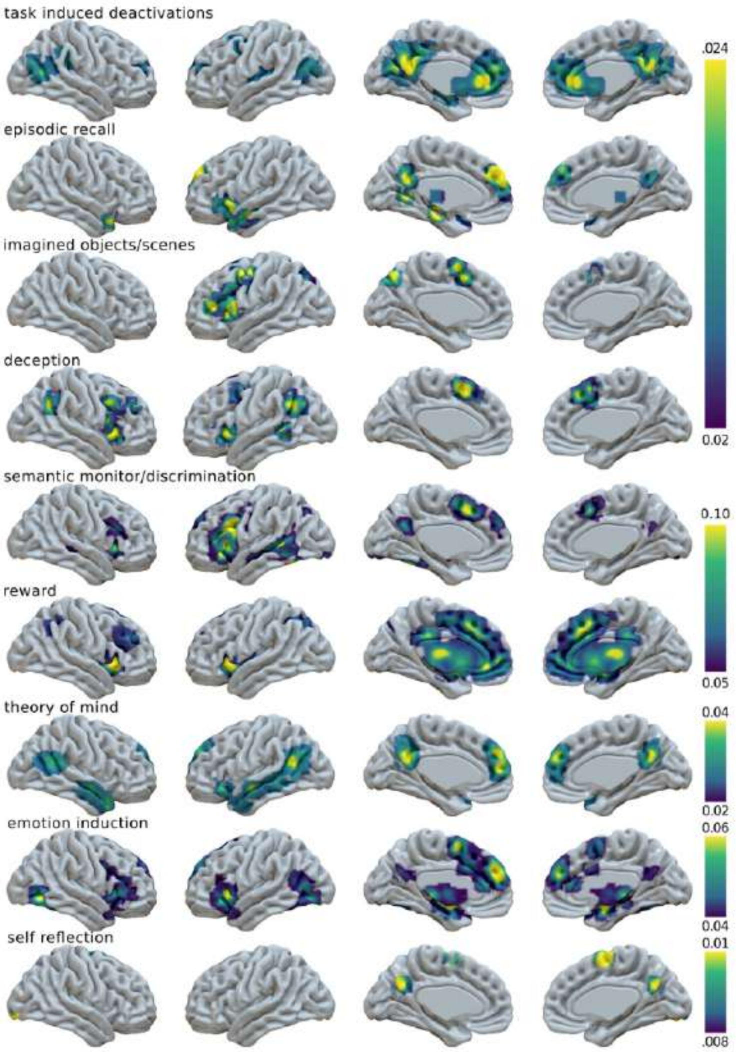


Figure D.5: *Surface mapping of 9 Activation Likelihood Estimation maps showing a nuanced involvement of the DMN in several tasks. Few maps match the prototypical representation of the DMN. Some of them show either weak activation or no activation at all in the midline core, and a strong expression of lateral areas of the network such as AG, IFG, and middle temporal gyrus. In addition, the insula and SMA/dorsal ACC, hubs of the salience network (SN) are often present. Rather than considering these as spurious findings, these an indication that, when the brain is engaged by external demands, multiple networks including DMN nodes would emerge. Although relying on intrinsic brain topology, such recruitment would be not strictly constrained by it and it might involve a flexible shift in brain hubness and a remodulation of cooperative and competitive long-range connectivity patterns. When analyzed through, MDS, PCA, and ICA, it can be seen that the activations of DMN regions are arranged along a continuum that spans from the most internal to a more external engagement. In addition, they suggest that semantic, reward, and emotional functions may be relevant elements of such outward-leaning default-mode of cognition. Lastly, and importantly, they indicate that the modulations of the DMN activations do not converge into a representative mid-point, but rather that they somewhat gravitate around it while shifting between internal, semantic, affective or motivational modes of cognition. From Mancuso et al. (2021).*

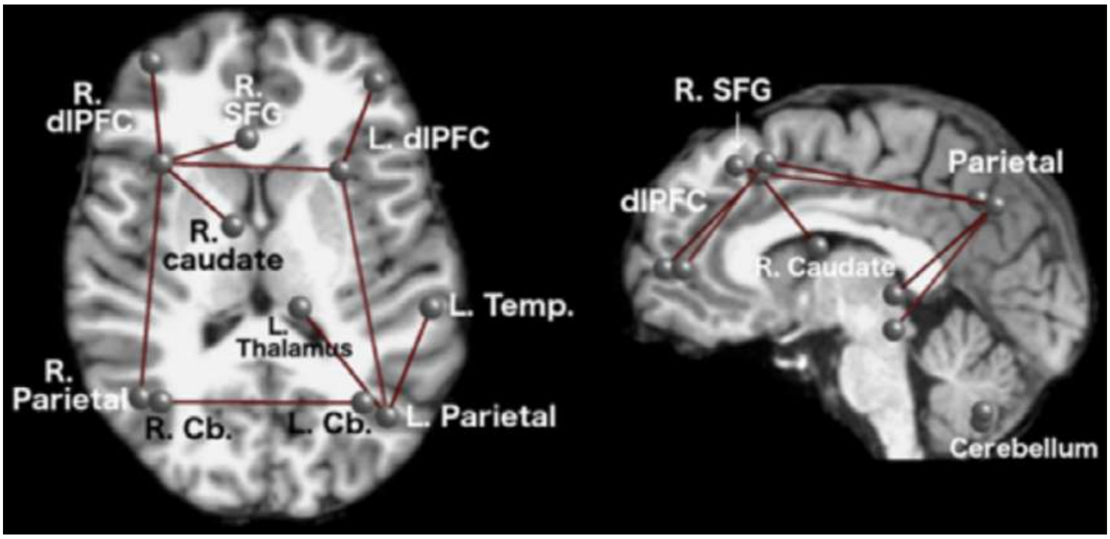


Figure D.6: *Structural core of the executive control network. Right and left dorsolateral PFC (R-dlPFC, L-dlPFC), Right and left PPC, superior frontal gyrus (SFG), right caudate and left anterior thalamus*

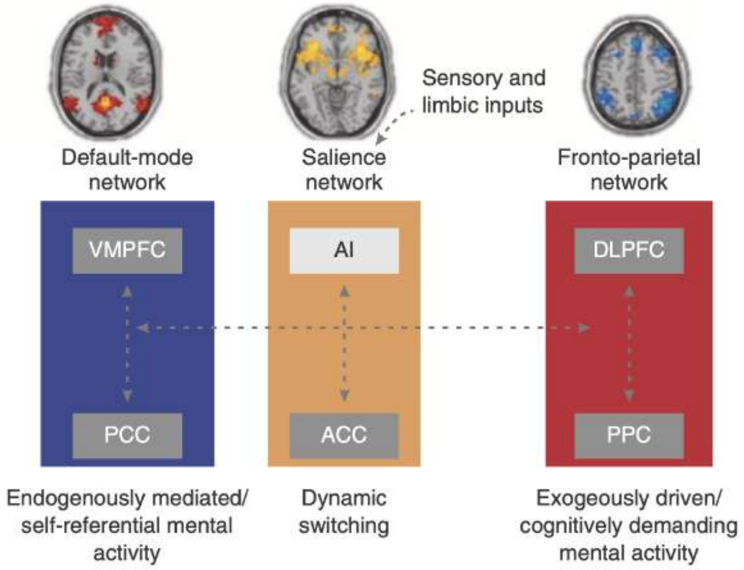

Sensory and limbic inputs
Default-mode network
Salience network
Fronto-parietal network
VMPFC
AI
DLPFC
PCC
ACC
PPC
Endogenously mediated/ self-referential mental activity
Dynamic switching
Exogeously driven/ cognitively demanding mental activity

Figure D.7: *The basic Network Switching Model in which the SN is hypothesized to initiate dynamic switching between the FPN and DMN and regulate attention to endogenous and exogenous events. Sensory and limbic inputs are processed by the anterior insula (AI), which detects salient events and initiates appropriate control signals for (i) access to resources for working memory in FPN and (ii) action selection via the anterior cingulate cortex. From Menon (2015); Menon and D'Esposito (2022).*

determines the salience of a stimulus, based on input from various other systems, including the dopaminergic reward system (Menon, 2015; Menon and D'Esposito, 2022).

AIC and ACC regions are also mentioned as the ones that are consistently active in almost all cognitive demands or tasks (Menon, 2015). Thus, they likely serve very broad functions.

The AIC, in general, presumably plays an important role in integrating information from one's internal environment, such as energy level, pain, emotions and sympathetic versus parasympathetic activation (i.e. whether one is stressed or not). The ACC has traditionally been linked to performance monitoring, which implies that it compares ongoing actions and outcomes with the direction of one's goals. In cooperation with other brain structures, such as the nucleus accumbens (NACC, part of the dopaminergic reward system), the ACC supports decisions on whether one is willing to spend effort in order to obtain a specific goal.

The central, canonical concept of the SN, according to Seeley (2019) is that the AIC is a major afferent cortical hub for perceiving viscero-autonomic feedback, whereas the ACC is the efferent hub responsible for generating relevant visceral, autonomic, behavioral, and cognitive responses. Through interactions with each other, these regions are likely to form a sort of information processing loop for representing and responding to homeostatically relevant internal or external stimuli and imbuing these stimuli with emotional weight.

Summing up, with the AI as its dynamic hub, the SN contributes to a variety of complex brain functions through the integration of sensory, emotional, and cognitive information. The mechanisms by which the SN contributes to cognitive and affective function can be recapped as follows (Menon, 2015):

1. Detection of salient events by the AI via differential sensory input and links with subcortical nodes involved in signaling reward, motivation, and affective saliency

2. Functional coupling of the AI with the dACC to facilitate rapid access to the motor system

3. Interaction of the AI with other insula subdivisions to mediate physiological reactivity to, and interoceptive awareness of, salient stimuli

4. Control signals to other large-scale networks that facilitate access to working memory resources

5. Switching between the lateral frontoparietal CEN and the medial frontoparietal DMN to keep attention focused on task-relevant goals.

A summary of the SN organization in relation to its major afferents and efferents, also including sub cortical areas, is provided in Figure D.8.

D.4 Network cooperation

The Network Switching Model presented in Figure D.7 is one but simple facet of the cooperation between intrinsic networks. Indeed, constructing and acting on mental models necessitates a brain substrate that integrates and flexibly updates many different cognitive, affective and physiological processes.

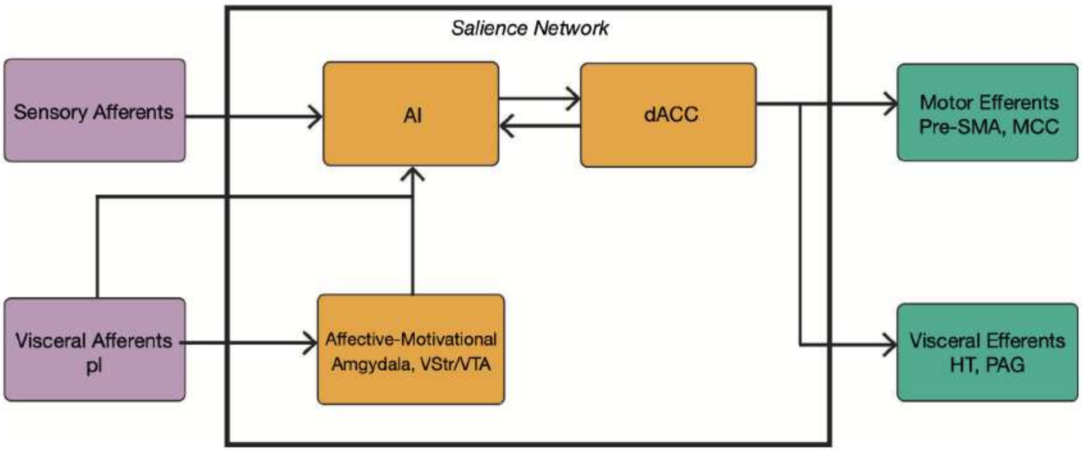

Salience Network
Sensory Afferents
AI
dACC
Motor Efferents Pre-SMA, MCC
Visceral Afferents pI
Affective-Motivational Amgydala, VStr/VTA
Visceral Efferents HT, PAG

Figure D.8: *SN organization at a glance. The AI receives convergent multisensory inputs, affective and motivational signals, and visceral afferents, reflecting biological saliency and cognitive demands. In contrast, the dACC plays a more dominant role in response selection, guiding overt behavior and modulating autonomic reactivity. AI, anterior insula; dACC, dorsal anterior cingulate cortex; HT, hypothalamus; PAG, periaqueductal gray; pI, posterior insula; VStr, ventral striatum; VTA, ventral tegmental area. From (Menon, 2015)*

Consider for instance the the ventro-medial PFC (vmPFC). The vmPFC is a cortical zone that spans multiple cytoarchitectonic regions and that is anatomically and functionally positioned to integrate conceptual thought with peripheral physiology Koban et al. (2021).

The vmPFC participates in multiple cortical networks that have been identified in resting-state functional MRI studies. The ventral vmPFC (or medial OFC) is part of the limbic network and is functionally coupled with the medial and anterior temporal lobes. The dorsal vmPFC is a core part of the DMN, and is coupled with the posterior cingulate cortex, precuneus and temporoparietal junction. Both the dorsal vmPFC and the ventral vmPFC are connected to the lateral OFC which is part of the limbic system (Figure D.9).

The vmPFC receives few direct sensory inputs. However, it has strong bidirectional links with sensory-integration regions in the lateral orbitofrontal cortex (OFC) and mediodorsal thalamus; interoceptive regions in the insula; motivational and reward-processing circuits, including the amygdala, hypothalamus and ventral striatum (including the nucleus accumbens), and circuits involved in memory and context, including the perirhinal cortex and hippocampus (Figure D.10). Strong descending projections from the vmPFC to autonomic and neuroendocrine control regions in the hypothalamus and brainstem, including the periaqueductal grey (PAG) and dorsal raphe enable the vmPFC to regulate visceromotor output.

More generally, the vmPFC in connection with other DMN regions and other brain networks enables the emergent process that concerns with the construction of mental models that integrate self and environment. Connectomics studies identify the DMN as an integrative hub network that sits at the top of a hierarchy combining multiple sensorimotor, unimodal processing and internal, multimodal processing. The vmPFC in particular is crucial for regulating physiology and behaviour, putting it in a special

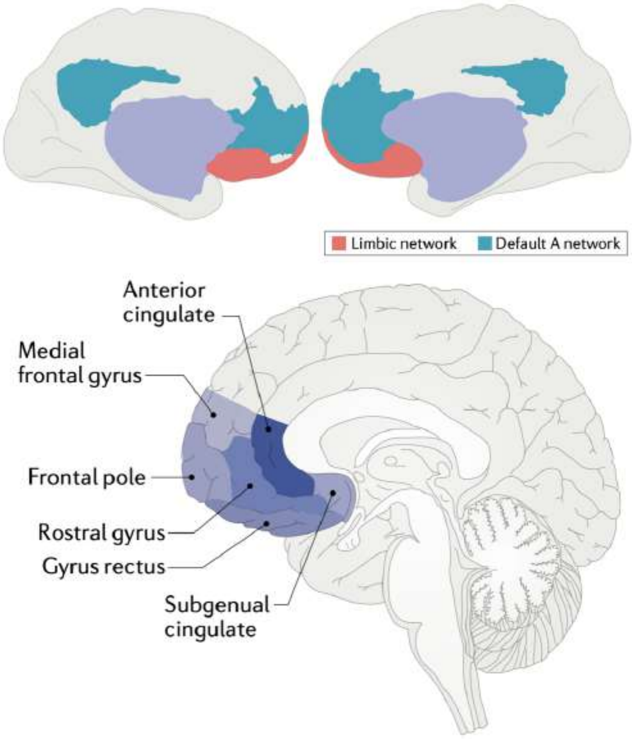


Figure D.9: *Anatomy of the vmPFC. Top panel: Limbic and DNM intrinsic networks. Bottom panel: the vmPFC includes the ventral anterior cingulate cortex and the subgenual cingulate cortex, the gyrus rectus, the medial parts of the rostral gyrus and frontal pole, and inferior parts of the superior or medial frontal gyrus; most of the vmPFC is part of the DMN, especially the default A network or core DMN, which serves as a hub between the medial temporal and dorsal subnetworks of the DMN; the most ventral part of the vmPFC (that is, the rostral gyrus and parts of the subgenual anterior cingulate cortex) is part of the limbic network. Adapted from (Koban et al., 2021)*

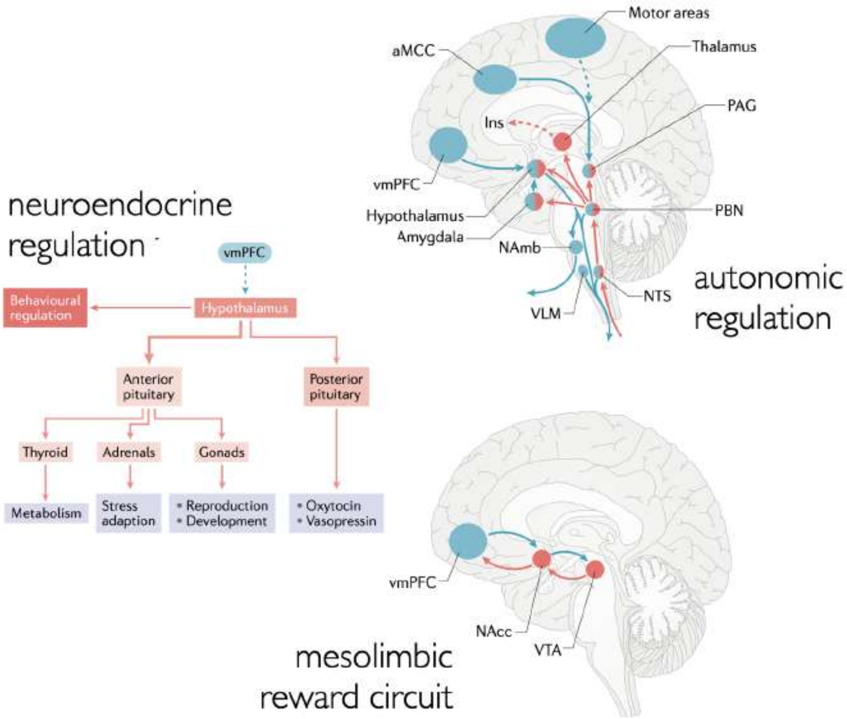


Figure D.10: *Functions of the vmPFC. Top: Brain areas associated with autonomic regulation include the vmPFC and its connections with limbic and brainstem areas (simplified overview); red denotes ascending tracts and blue denotes descending tracts; autonomic regulation involves connections from areas of different large-scale networks, including limbic, DMN, SN and somatomotor areas. Center: Via its close connections to the hypothalamus, the vmPFC can also influence the neuroendocrine system. Bottom: Together with the ventral striatum–nucleus accumbens (NAcc) and the ventral tegmental area (VTA), the vmPFC is part of the mesolimbic reward circuit (simplified here). Adapted from (Koban et al., 2021)*

position at the interface between conceptual thought, decision-making and bodily regulation (Figure D.11).

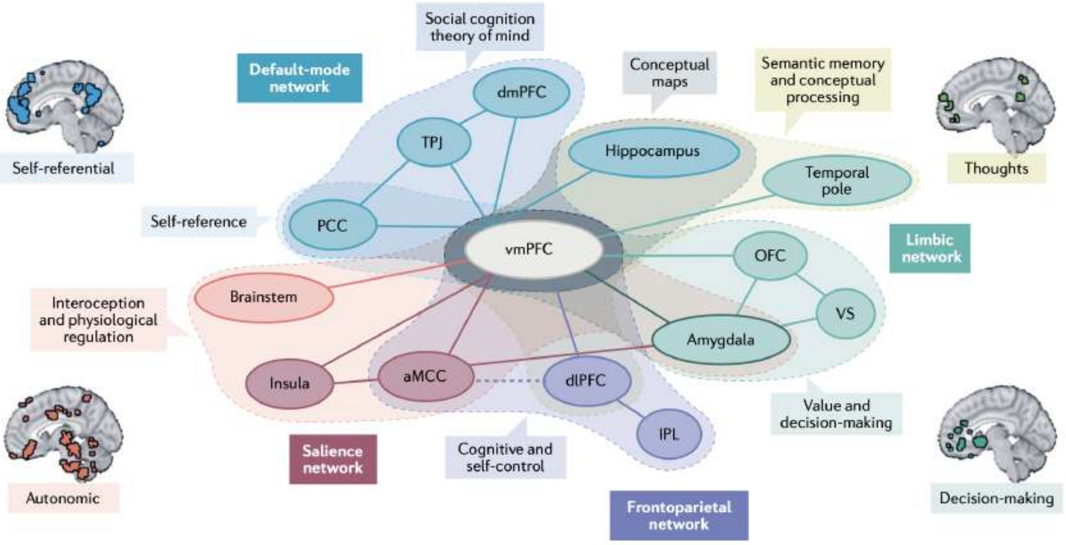


Figure D.11: *General functional associations of the vmPFC. Together with other regions of the default-mode network, including the temporoparietal junction (TPJ), the dorsomedial prefrontal cortex (dmPFC), the hippocampus and the posterior cingulate cortex (PCC), it is involved in social cognition and self- referential thought. Both the hippocampus and the vmPFC show evidence for grid-like coding of spatial and conceptual maps, and together with other temporal and frontal areas are involved in semantic memory and conceptual processing more broadly. The most ventral part of the vmPFC is connected to the limbic network, including the orbitofrontal cortex (OFC), the ventral striatum (VS) and other subcortical areas. Together with the VS, the vmPFC is important for reward processing and decision-making. Therefore, it is amenable to interactions with the frontoparietal network, especially the dorsolateral prefrontal cortex (dlPFC) and the inferior parietal lobule (IPL), involved in executive function and self-control. Together with areas of the salience network (especially the anterior midcingulate cortex (aMCC) and the anterior insula) and subcortical regions, the vmPFC is involved in interoception and physio- logical regulation. Adapted from (Koban et al., 2021)*

FigureD.11 presents a high-level view of integration (though, markedly, from the vmPFC perspective) between a subset of available intrinsic networks (limbic, DMN, SN, frontoparietal CEN) where all networks, in parallel, appear to be equally contributing to the overall activity of the brain, on the base of each network "specialization". Yet, this picture does not make justice of many aspects that have recently emerged in neuroscience research.

A puzzling problem is posed by the DMN role. In its original definition, as previously mentioned, the DMN was defined to collect regions of the brain distributed across the parietal, frontal and temporal cortex that decrease their neural activity during complex attention-demanding tasks. Their behaviour implied a neural baseline from which specific, more attention-demanding states deviate.

More recently, however, studies established that neural activity within DMN regions, such as the PMC, contain signals that relate to neural functioning in diverse systems, including those outside the DMN. These observations suggested that, as well as forming a cohesive network, the DMN can represent brain activity taking place in other cortical systems, with these representations of activity from within other neural networks often referred to as "echoes". Activity in the DMN can also provide information about the activity of task-positive systems, a pattern inconsistent with the classical view of the DMN as being intrinsically isolated from regions that are involved in external goal-directed thought.

In particular, the analysis of connectivity gradients[1], suggests that characterizing the intrinsic activity of the DMN as being primarily isolated from, or antagonistic with, that of task-positive systems does not provide a complete picture. More likely, the intrinsic behaviour of the DMN encompasses multiple modes of operation, some which are related to external tasks, and others that are not. Regions of the DMN are engaged across multiple, apparently distinct, psychological domains (episodic, linguistic, social and emotional). Subly, emerging insight into the role of the DMN in cognition comes from recent studies in which this system's activity can be related to the specific periods within a task when prior experience contributes to the broader goal of external task completion (). Cognitive neuroscience suggests that the goal-oriented control of cognition (often known as executive control) is partly implemented by regions of multiple demand cortex, which are often viewed as the apex of a cortical hierarchy that is important for organizing behaviour in a goal-orientated manner. These regions seem superficially to be the opposite of the DMN, as they enhance their responses in situations in which tasks become more difficult. However, there is growing evidence that the two systems can work together. For example, even when neural activity is reduced in the DMN because of increased external task demands, some DMN regions (such as the posteromedial cortex, PMC) show increased connectivity with regions of the CEN and support task-relevant cognition. This may also occur during autobiographical planning. and in situations in which decisions combine both prior knowledge and task goals. These interactions are made possible because the CEN is spatially fractionated into regions specialized for their interactions with the DMN and those linked to other multiple response regions more closely aligned to the external environment. Further, many processes linked to the hippocampus and parahippocampal gyrus, such as episodic memory and spatial navigation129, are also linked to activation of the DMN.

Smallwood et al. (2021) have suggested that topographical features of the DMN might account for the fact that its regions are functionally connected yet separated from sensory inputs and motor outputs. Consistent with the notion of convergence of signals from unimodal systems (vision, auditory, etc) into the DMN, studies in humans have shown that large-scale networks are organized along the cortical surface from unimodal regions to the DMN in an orderly manner, which indicates that the DMN can be understood as being located at the end of processing streams that are anchored at the cortical periphery. Forms of higher-order cognition may rely on the DMN because its location allows it to encode information about brain activity from across the cortex. For in-

[1] *Connectivity gradients* are computed using patterns of covariance within a data matrix. These gradients are ranked based on the percentage of variance that each principal component explains within the initial data (known as the explained variance). Within each gradient, brain regions are organized based on the similarity of their observed patterns of activity to each other. In these gradients, brain regions grouped at one end have similar fluctuations in activity over time, and collectively show less similarity to the groups of regions at the other end of a dimension (which are also similar in their time courses)

stance, processing streams, such as the ventral visual stream, are arranged such that the regions involved respond to increasingly abstract features of cognition as information passes along the stream85; if the DMN is located at the end of these streams, then it may be important for relatively abstract features of cognition and behaviour.

Smallwood et al. (2021) have shown that bbased on the spatial location of the peaks of connectivity gradient 1, we propose that the DMN divides the brain into mutually exclusive cortical fields, each defined by the convergence of a specific set of sensory/-motor streams towards a region of the DMN at the centre of each field.

Also, by representing the intrinsic networks organized along a connectivity gradient, they have argued that the location of the DMN at the end of processing streams suggests that it may correspond to the a functional integrative centre of the cortex. The schematic illustrates how the DMN can be thought of as the end of multiple processing streams that originate in sensorimotor cortex, and thus the functional core of the brain. The diagram shown in Figure D.13 provides a topological explanation for how the cortex balances the need for segregation between different sensory systems with the need for progressive integration of information from the periphery to the core.

The framework proposed by Smallwood et al. (2021) has many advantages.

DMN regions may represent coarse information about patterns in brain-wide activity that could be similar for many different potential configurations at lower levels of the hierarchy. This could explain why the DMN is involved in many different representational states that share broad features but differ in their specific informational content.

A DMN-based hierarchy can shape the temporal dynamics of complex systems and help integrate disparate distributed information across time, such as signals in peripheral cortical regions.

Hierarchies are also a core premise of accounts of predictive coding. In this case, DMN activity could be linked to a cycle of monitoring, and correcting for, the emergence of prediction error across the cortex. In this way, neural patterns across the DMN may provide information regarding the degree to which specific brain contexts are predictable, a metric that would be useful, for example, in shifting between exploratory and exploitative modes of foraging behaviour (Smallwood et al., 2021).

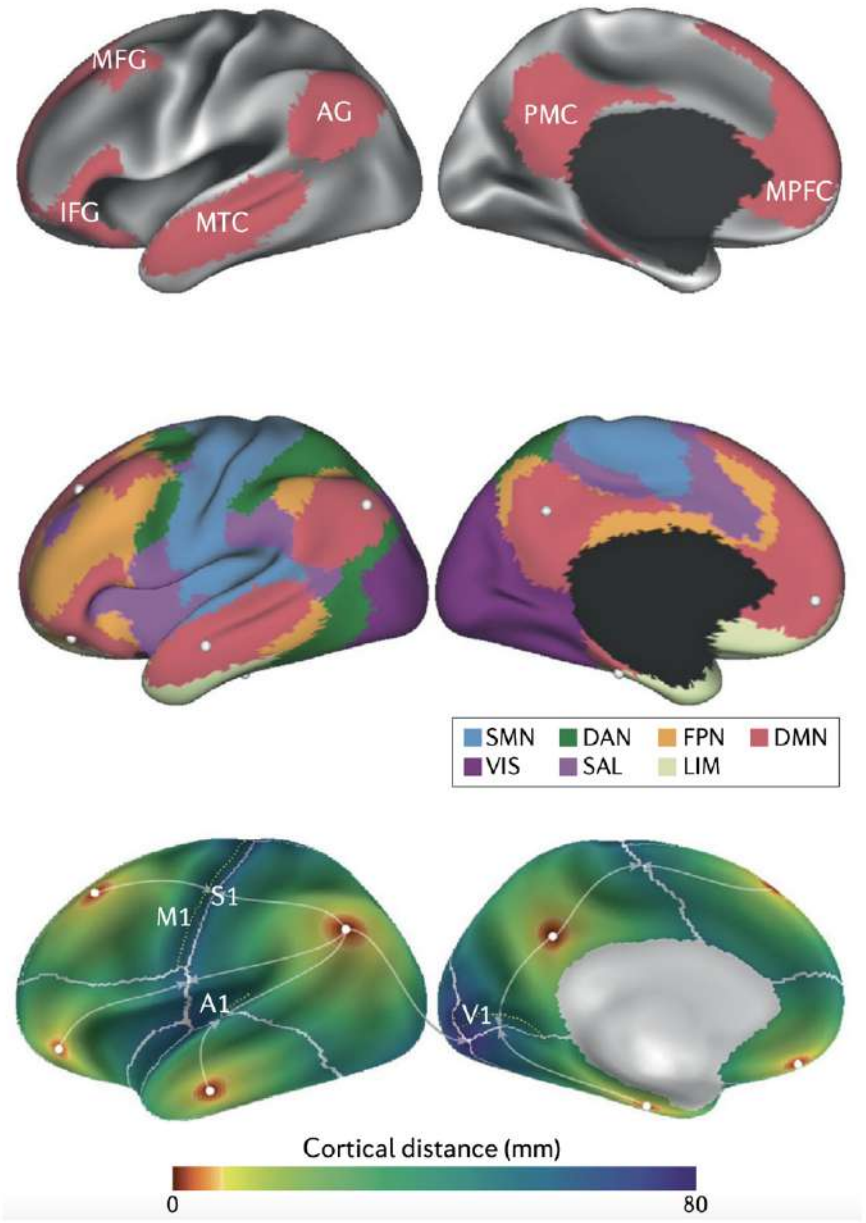


Figure D.12: *Top panel: The classic DMN (pink) regions identified as being more consistently deactivated by tasks include the posteromedial cortex (PMC), medial prefrontal cortex (MPFC), middle frontal gyrus (MFG), inferior frontal gyrus (IFG), middle temporal cortex (MTC), and angular gyrus (AG). Center panel: The DMN presented in the context of other large-scale brain networks, different colours correspond to different networks; the centroid of the regions that make up the DMN are marked by dots on this panel. Bottom panel: the centroids are the most distant from regions of unimodal sensory cortex (primary auditory A1, motor M1, somatosensory S1 and visual V1); colour gradient represents the spatial distance along the cortical surface between the peaks of connectivity gradient in the DMN and other brain regions (grey lines indicate regions of the cortex that are equidistant to two DMN regions, and arrows indicate which sensory landmarks each DMN region is closest to - e.g., PMC is equidistant between M1 and V1). Adapted from Smallwood et al. (2021)*

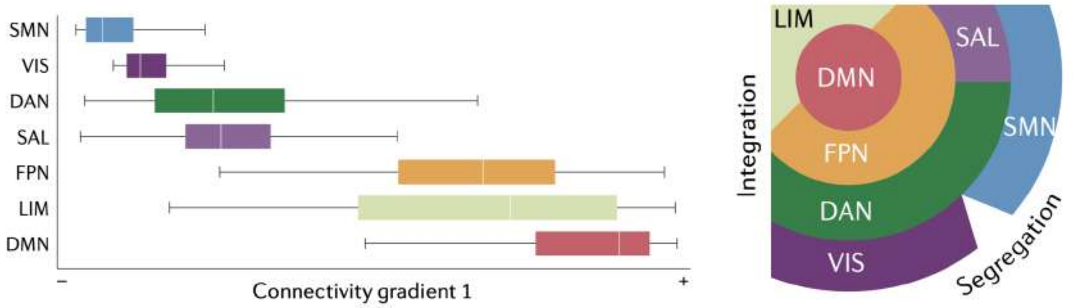


Figure D.13: *Segregation vs. integration in brain's large-scale networks. Left: Graph representing the networks from Fig. D.12 organized along the first connectivity gradient; +/– indicate the two ends of this dimension of brain activity; DAN, dorsal attention network; FPN, frontoparietal network; LIM, limbic network; SAL, salience network; SMN, somatomotor network; VIS, visual network. Right: Segregation / integration diagram showing segregation between different sensory systems (indicated by their locations on different points on the circumference of the semicircle) and progressive integration of information from the periphery to the core (illustrated by the location of different networks at different points on the radius of the semicircle). Adapted from Smallwood et al. (2021)*

APPENDIX *E*

The Structural Model of corticocortical connections and the organization of the brain

The principle of systematic variation of the cortex provides the framework for parcellating the cortex from a theoretical perspective.

The SM captures the overall laminar structure of areas by dividing the cortical architectonic continuum into discrete categories or cortical types (García-Cabezas et al., 2019; Barbas and García-Cabezas, 2016). A cortical type describes a category of cortical areas with comparable laminar differentiation, regardless of placement within a cortical sensory, high-order association or motor system.

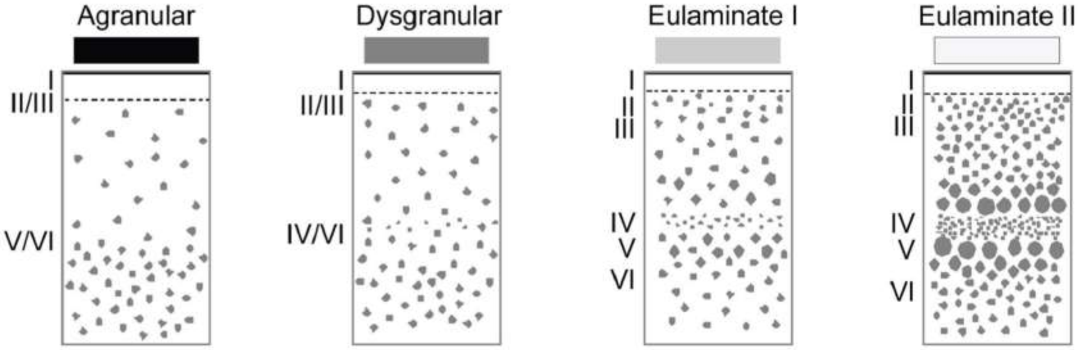


Figure E.1: *Cortical types agranular, dysgranular, eulaminate I, and eulaminate II represented in four sketches. Adapted from (García-Cabezas et al., 2019; Barbas and García-Cabezas, 2016)*

Several structural features combine to define a cortical type: the presence or absence of layer IV; the thickness or density of layer IV, when present; the distinction between adjacent layers; the relative population of neurons in the upper compared to the deep layers.

Some neocortical areas in the human cortex do not have six layers from the get-go, and remain that way during the life span. Areas lacking an inner layer IV belong to the agranular cortical type.

Nearby areas have an incipient layer IV and belong to the dysgranular type. These have a poorly developed layer IV, which is both thin and less dense than layer IV of eulaminate areas.

Eulaminate areas, instead, have a clearly identifiable layer IV. Among eulaminate areas, lamination is least differentiated in areas that are near dysgranular areas, and most distinct in areas that are the most distant from the limbic areas (eulaminate II).

A first key concept here is that each cortical system, regardless of its placement on the cortical mantle, is composed of areas that at one extreme have fewer than six layers (limbic areas), leading to adjacent areas that have six layers (eulaminate) and finally to eulaminate areas with the best delineated layers. Changes in laminar structure are exemplified by a higher density of spines and dendritic branching in pyramidal neurons in limbic than in eulaminate areas, a lower myelin density in limbic than in eulaminate areas, and other structural features.

A second key concept concerns connections. Feedforward connections originate in neurons in the upper layers (mostly layer III) and their axons terminate in the middle layers, which include layer IV. Feedback connections originate in the deep layers (V and VI) and their axons terminate in the upper layers, and especially layer I, to a lesser extent layer VI, but avoid the middle layers.Thus, for any pair of linked cortices — whether they are neighbors or not — their interconnections reflect their structural relationship. The term "feedforward" describes connections from an area with more elaborate laminar structure, which terminate in an area with less elaborate structure; "feedback" describes connections that have the opposite relationship.

From a development standpoint, since limbic areas have a lower density of neurons than eulaminate areas, especially in the upper layers the SM hypothesizes that they must have a shorter developmental period than eulaminate areas. This hypothesis is consistent with available developmental data in primates: limbic areas complete their development first, whereas the best laminated areas (which also has the highest density of neurons in the primate cortex), has the longest period of development.

According to the SM, based on laminar structure on may distinguish between:

- **Allocortex**: ancestral part of the cerebral cortex, which includes the hippocampal formation (archicortex) and the primary olfactory cortex (paleocortex); allocortex is agranular
- **Periallocortex**: neocortical areas neighboring the allocortex (agranular)
- **Neocortex**: part of the cerebral cortex with three or more layers and columnar organization. Sometimes referred to as isocortex. It can be further partitioned in
  - **Limbic cortex**: neocortical areas found on the edge (limit) of the hemisphere; they form the base or stem of the cortex. Limbic areas are either agranular (lack inner layer IV) or dysgranular (have an incipient layer IV).
  - **Eulaminate cortex**: neocortical areas with well-developed layer IV.

- **Koniocortex**: the eulaminate cortices with the most well-developed layer IV.

The overall schematic of the primate cerebral cortex according to the SM is outlined in Figure E.2

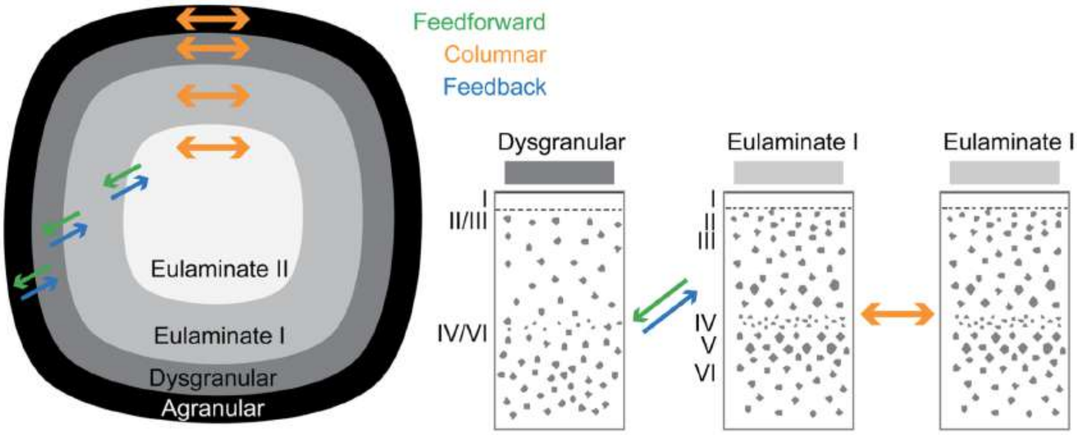


Figure E.2: *Schematic of the primate cerebral cortex shows the arrangement of cortical types in rings. Laminar differentiation progresses from the outer or basal (black and dark gray) to the inner rings (lighter shades of gray). The edge of the cortex (black and dark gray) is actually thin compared to the greatly expanded eulaminate areas in the center. Cortical areas have stronger connections with other areas in the same ring and display columnar patterns of connections (orange arrows). Connections between areas in different rings (i.e., of different cortical type) are less strong than connections within the same ring and display feedback (blue arrows) and feedforward (green arrows) laminar patterns of connections. The laminar pattern of connections is related to the cortical type difference of the connected areas. Pathways from dysgranular to eulaminate areas are feedback (blue arrow); pathways from eulaminate to dysgranular are feedforward (green arrow); pathways between areas of comparable cortical type are columnar (orange arrow). Adapted from (García-Cabezas et al., 2019; Barbas and García-Cabezas, 2016)*

E.1 Predictive processing view of the SM

Chanes and Barrett (2016) have proposed that the SM is suitable to implement predictive coding in the brain. The direction of predictions (feedback connections) and prediction errors (feedforward connections) is determined by the relative degree of laminar differentiation of the cortical areas involved. Predictions originate primarily in the deep layers of cortical areas with less laminar differentiation and terminate primarily in the superficial layers of more differentiated areas. In the opposite direction, prediction errors originate primarily in the superficial layers of cortical areas with more laminar differentiation and terminate in the deep layers of less differentiated areas.

When two areas have a comparable laminar structure, their projections originate and terminate both in superficial and deep layers (they are defined as lateral connections).

As a consequence, cortical areas, such as limbic cortices (which have the least differentiated laminar structure in the entire neocortex) primarily send predictions to better

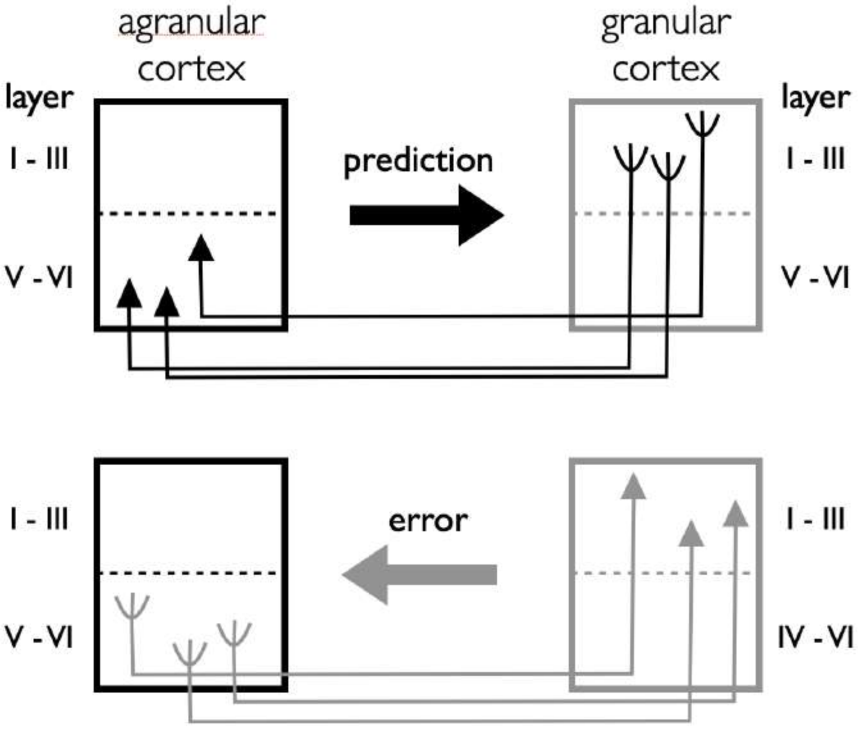


Figure E.3: *The flow of prediction and prediction-error signals between cortical columns based on cortical lamination gradients. The relative difference in laminar structure between two communicating cortical columns predicts whether the information flow is a feedback (prediction) or a feedforward (prediction-error) signal. Prediction signals (black) originate in the deep layers (Layers V and VI) of less differentiated cortical areas (such as agranular cortex with undifferentiated Layers II and III and without a Layer IV, as depicted in the black column) and terminate in superficial layers of areas with a more developed laminar structure (such as dysgranular cortices with differentiated Layers II and III and a rudimentary Layer IV or granular cortices with differentiated Layers II and III and a well-defined Layer IV, depicted in the grey column). Prediction-error signals (in grey) flow in the other direction, originating in the superficial layers (II and III) with more laminar differentiation and terminating in middle deep layers (V and VI) of areas with less differentiated laminar architecture. Adapted from (Hutchinson and Barrett, 2019)*

laminated cortical areas and primarily receive prediction error.

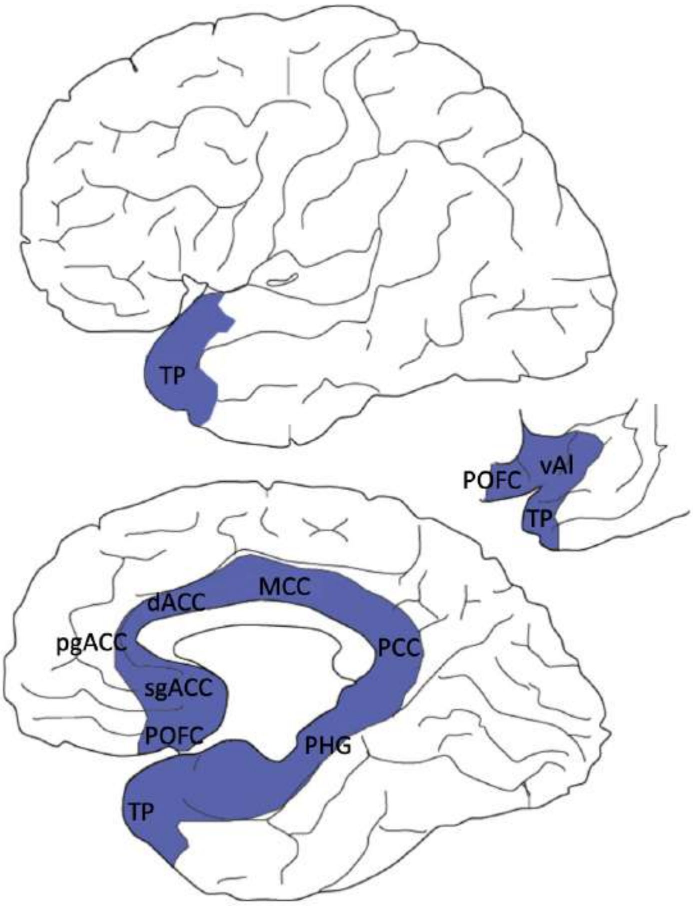


Figure E.4: *Limbic Cortices in the Human Brain. Cortical limbic areas (blue) form a ring around the corpus callosum on the medial wall of each hemisphere, continuing along the temporal cortex and the base of the brain. They are neocortical areas that either lack or have a rudimentary layer IV (i.e., are agranular or dysgranular, respectively). They are located between the simpler allocortex and the better laminated eulaminate cortex. Limbic cortices include the cingulate cortex (subgenual anterior cingulate cortex, **sgACC**; pregenual anterior cingulate cortex, **pgACC**; dorsal anterior cingulate cortex, **dACC**; mid-cingulate cortex, **MCC**; posterior cingulate cortex, **PCC**), the ventral anterior insula, **vAI**, the posterior orbitofrontal cortex,**POFC**, the para-hippocampal gyrus **PHG**, and the temporal pole, **TP**. From Chanes and Barrett (2016)*

Moreover, primary sensory cortices (with the most differentiated laminar structure) receive predictions from less laminated cortical areas and send prediction error. Other cortical areas (with intermediate degrees of laminar differentiation) send both predictions and prediction error depending on the relative laminar differentiation of the receiving cortices.

Based on this hypothesis, the overall organization of the predictive brain can be schematized as in Figure E.5.

As it can be seen from Figure E.5, this organization involves all cortical sensory systems. Indeed, limbic cortices can be identified in visual, auditory, somatosensory

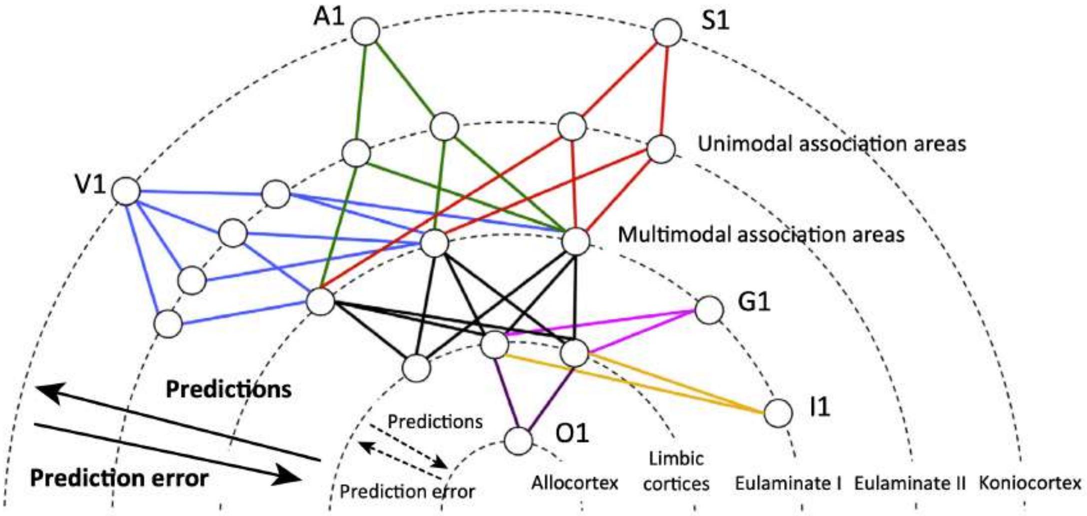


Figure E.5: *The predictive brain according to Chanes and Barrett (2016). Each ring represents a different type of cortex, from greater (exterior circles) to less (interior circles) laminar differentiation. Primary sensory cortices (lower level of each sensory system) are indicated: **A1**, primary auditory cortex; **G1**, primary gustatory cortex; **I1**, primary interoceptive cortex; **O1**, primary olfactory cortex; **S1**, primary somatosensory cortex; **V1**, primary visual cortex. Unimodal association areas include extrastriate areas (V2, V3, V4, MT/V5) for the visual system, superior temporal areas surrounding A1 for the auditory system, and the superior parietal lobule (**SPL**) for the somatosensory system. Multimodal association areas include the dorsolateral prefrontal cortex (**DLPC**), lateral temporal cortex (**LTC**), and posterior parietal cortex (**PPC**). Predictions flow from cortical areas with less laminar differentiation to areas with greater laminar differentiation. Prediction error flows in opposite direction. The number of cortical steps (hierarchical levels) is less in interoceptive, gustatory, and olfactory systems than in exteroceptive visual, auditory, and somatosensory systems.*

and interoceptive systems

For instance, the interoceptive system in charge of the perception and integration of autonomic, hormonal, visceral, and immunological homeostatic signals that collectively describe the physiological state of the body. In this case, as proposed by Barrett and Simmons (2015), visceromotor limbic cortices - notably the anterior and mid-cingulate cortices, (**ACC**, **MCC**) and the ventral anterior insula, **vAI** - send predictions to the primary interoceptive cortex in the mid-to-posterior insula (I1), which is eulaminate in structure. Visceromotor cortical limbic areas also send predictions to subcortical structures that control the autonomic, hormonal, metabolic, and immunological systems (e.g., the amygdala and the hypothalamus).

Clearly, there are differences across systems in the amount of cortical processing. Compared with interoception, information from visual, auditory, and somatosensory modalities is processed more extensively in the cerebral cortex. Predictions and prediction errors are computed across several levels of cortical processing. More precisely, there are several synaptic connections between primary sensory cortices in which representations are more specialized and cortical limbic areas in which they are more integrated. The interoception system entails fewer steps.

Primary interoceptive cortices in mid- and posterior insula (I1) are eulaminate in structure: they have a less developed layer IV than koniocortices of primary visual, auditory, and somatosensory cortices. This difference in degree of laminar differentiation along which predictive signals are coded - smaller in the interoceptive system (eulaminate to limbic) versus larger in the visual, auditory, and somatosensory systems (koniocortex to limbic) - may be one reason why interoceptive perception is less differentiated and lower in dimensionality when compared with exteroceptive perception.

According to this model, limbic cortices, at the top of the predictive hierarchy, create a highly connected, dynamic functional ensemble for information integration and accessibility in the brain (the limbic global workspace). Because of their anatomical position at the top of sensory and motor processing hierarchies, limbic cortices are strongly interconnected, and have strong bidirectional connections with subcortical structures such as the amygdala, the ventral striatum, and the hypothalamus. Therefore, highly integrated neural representations in limbic cortices are easily accessible by virtually the whole brain. In every conscious moment, all modalities are represented in the global workspace, but the type of content that is prioritized may determine whether we categorize the experience as “emotion”, “perception”, or “cognition”. By virtue of their structural and functional properties, they are likely to contribute to create a unified conscious experience.

The model provides novel insights, at the network level, about the flow of information within intrinsic brain networks (Figure E.6).

A brief recap of fundamental large-scale networks is provided in Appendix D

E.2 The SM and intrinsic networks

Each intrinsic network, such as the SN and the DMN (Figure E.7) includes areas with varying degrees of laminar differentiation (including limbic cortices, Paquola et al. (2021)).

Indeed, the functional role of networks such as the DMN may be understood with

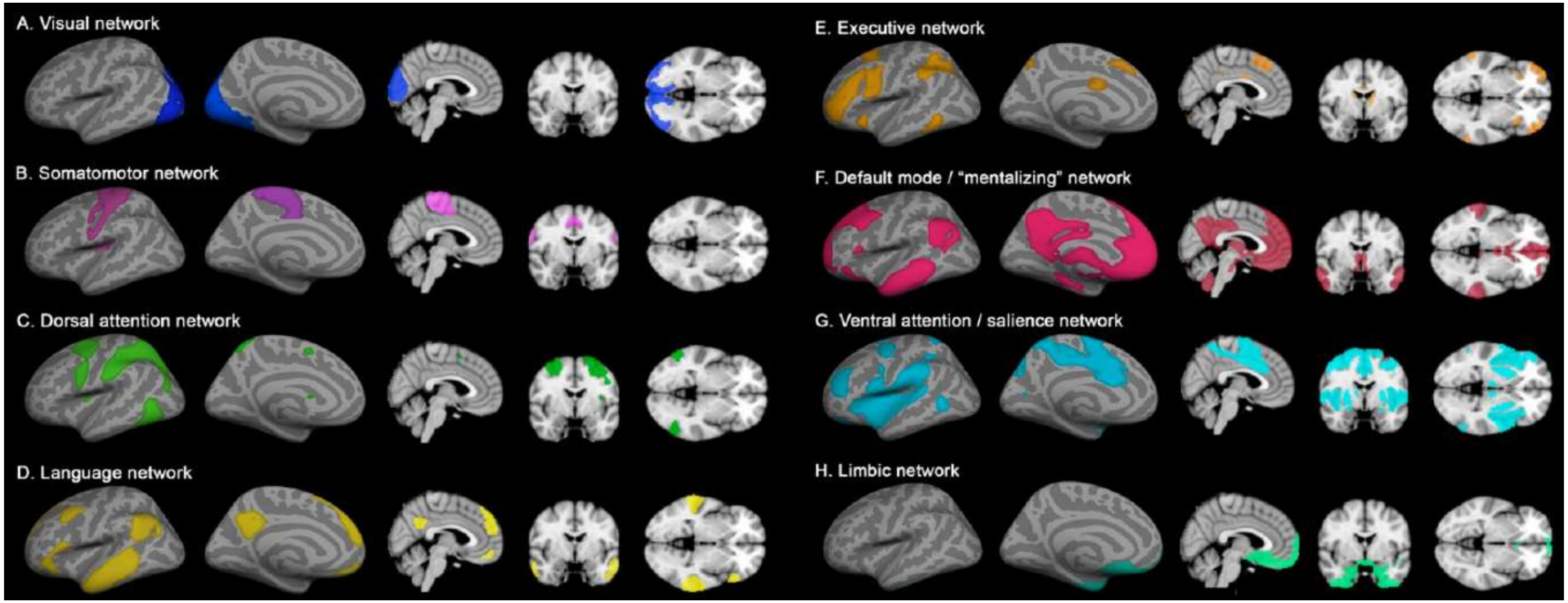

A. Visual network
B. Somatomotor network
C. Dorsal attention network
D. Language network
E. Executive network
F. Default mode / "mentalizing" network
G. Ventral attention / salience network
H. Limbic network

Figure E.6: *Intrinsic networks in the brain. Each network is a population of neurons that fire synchronously so that their firing is strongly related over time. The neurons that make up an intrinsic network coordinate their spontaneous activity via their anatomical connections. The structural connections between neurons do not determine an intrinsic network, but place boundaries on or constrains those networks (whose connectivity is functional). Within these constraints, the intrinsic networks are more variable and have a dynamic, functional repertoire. These networks develop in the first few years of life. Some of them are unique to humans, whereas others can be seen in the brains of other apes, and monkeys, and even in rats, although they may not perform the same function in humans and non-human animals. From* `https://how-emotions-are-made.com/notes/Intrinsic_networks`.

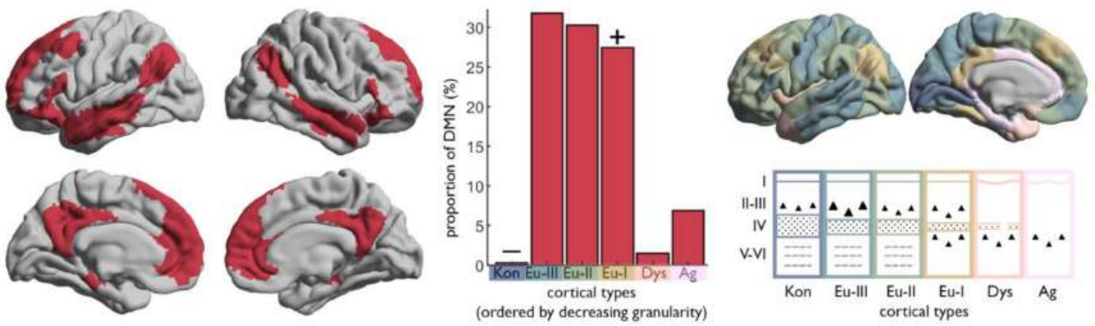


Figure E.7: *Cortical types within the DMN. Left: the canonical DMN. Center: the histogram of cortical types. Right:Cortical types are shown on the surface. The schematic highlights prominent features that vary across cortical types, including the location/size of largest pyramidal neurons (triangles), thickness of layer IV, existence of sublayers in V-VI (grey dashed lines), regularity of layer I/II boundary (straightness of line). Kon=koniocortical. Eul=eulaminate. Dys=dysgranular. Ag=agranular. Adapted from Paquola et al. (2021)*

respect to their neuronal architecture. The clarity of cortical layers as well as the prominence of a granular layer IV (the granularity), decreases with synaptic distance from primary sensory areas. Cytoarchitectural changes present as a gradient running across the cortical surface, which is commonly termed the "sensory-fugal" axis, as it mirrors a shift in receptiveness from the external world to the internal milieu consistent with the architecture discussed in the previous section.

A first dimension of cytoarchitectural changes can be gauged by measuring intracortical variations in cell body staining of the human brain, depth-wise variations in cell density and soma size. This can be can be quantified via diffusion map embedding, a nonlinear manifold learning technique. The first eigenvector (E1) accounts for the principle axis of variation in cytoarchitecture and is distinct to the gradient of laminar elaboration that is captured by the cortical types (Figure E.8). E1 captures changes in the intracortical staining intensity profiles, which capture depth-wise variations in the density and size of cell bodies. Areas with lower E1 values exhibit higher overall staining intensity, with a noticeable peak at mid-depths, whereas areas with higher E1 values show overall lower staining intensity with a flatter profile and more limited differentiation across depths.

A second useful dimension for characterizing the cytoarchitectural gradient is navigation efficiency (Enav), which relates to distance travelled between a seed and a target along the structural connectome.

The same analyses can be performed for all intrinsic networks. From Figure E.10 it can be easily appreciated the uniqueness of the DMN relative to other functional networks. While each functional network harbours multiple cortical types, the distribution of types differs significantly between the DMN and other functional networks, indicating a unique cytoarchitectural makeup of the DMN. The DMN has the most balanced navigation efficiency across cortical types, which is partly attributable to its size and spread.

Together, this shows that structural connectivity of the DMN to other brain regions is organised along E1, particularly for more granular types that have higher navigation

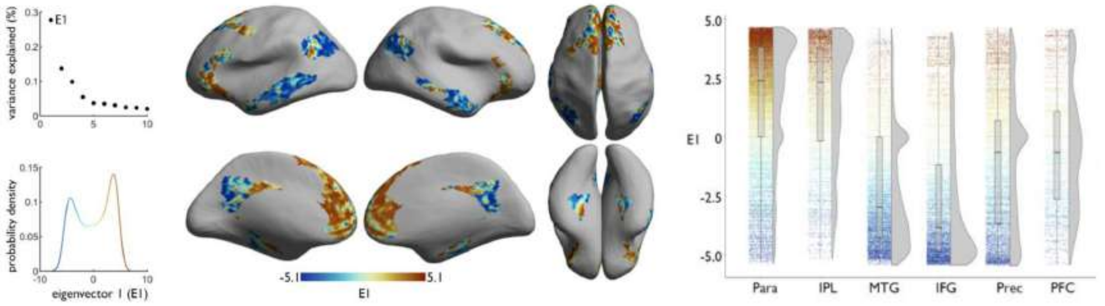


Figure E.8: *Cytoarchitectural gradient and heterogeneity of the DMN. Left: approximate variance explained of affinity matrix by each eigenvector, which reflects variation in cytoarchitecture and the probability density plot shows a bimodal distribution of E1 values. Center: E1 projected onto a 3D rendering of the human brain. Right: Raincloud plot shows the distribution of E1 within circumscribed subregions of the canonical DMN: Para, parahippocampus; IPL, inferior parietal lobule; MTG, middle temporal gyrus; IFG, inferior frontal gyrus; Prec, precuneus; PFC, prefrontal cortex. Adapted from Paquola et al. (2021)*

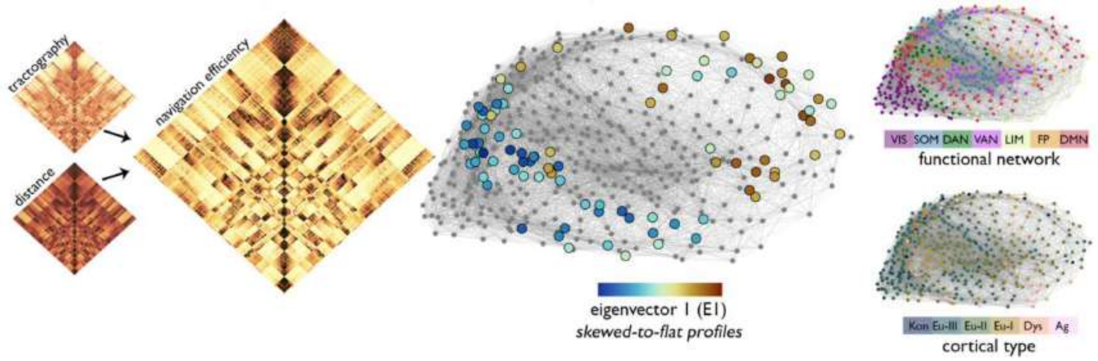


Figure E.9: *Navigation efficiency of the DMN within the structural connectome. Left: Navigation efficiency (Enav) calculated from step-wise progression along the tractography-based connectome, where each step is determined by spatial proximity to the target node. Center: Graph representation of the Enav connectome within the left hemisphere. The matched representations depict E1 (within DMN). Right: the same representation computed for functional networks, and cortical types. Adapted from Paquola et al. (2021)*

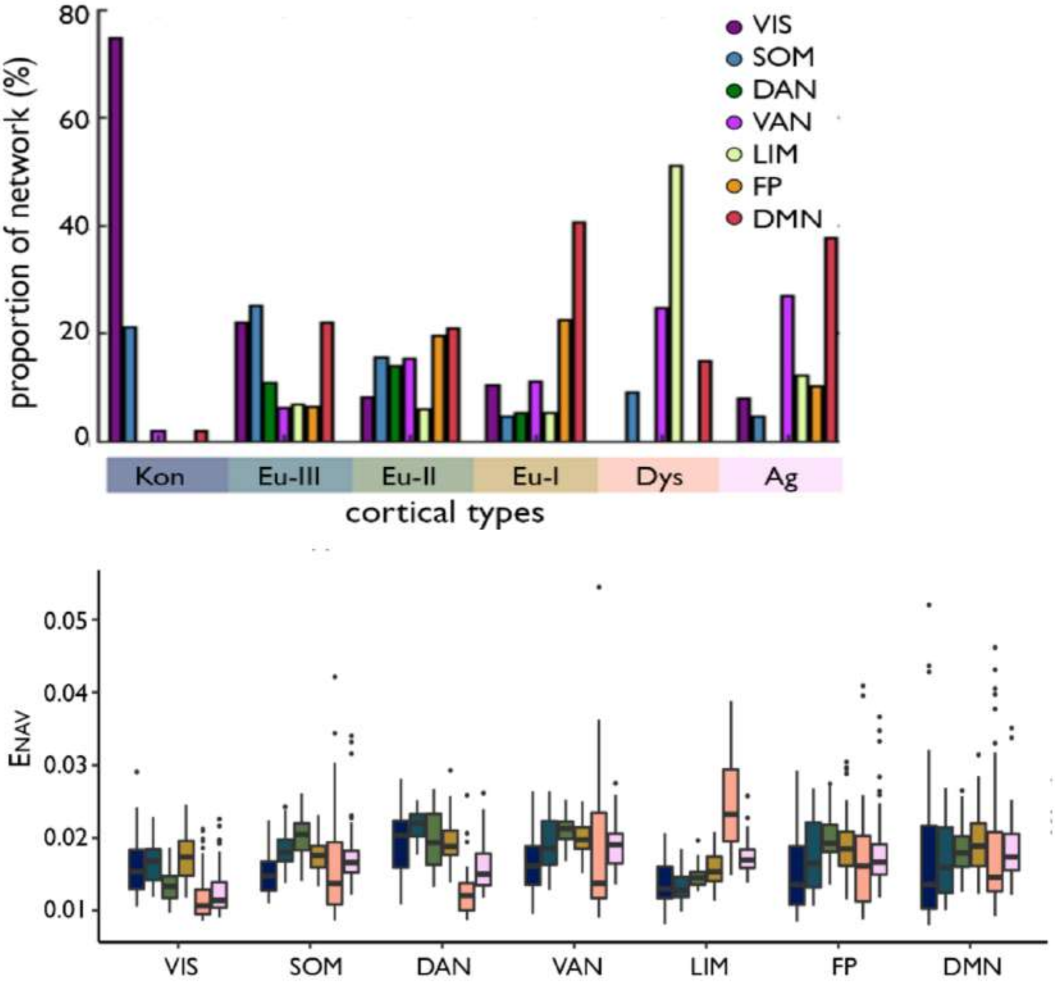


Figure E.10: *Cytoarchitectural heterogenity and navigation efficiency of intrinsic networks. Top: Crosstabulation of functional network by cortical type; frequency is provided relative to total vertices in each type; VIS=visual; SOM=somatomotor; DAN=dorsal attention network; VAN=ventral attention network; LIM=limbic; FP=frontoparietal; DMN=default mode network. Bottom: Boxplots show median connectivity of each parcel to each cortical type stratified by functional network. Adapted from Paquola et al. (2021)*

efficiency to DMN areas with lower E1 positions. The skewed aspect of E1 is thus structurally poised to integrate signals from a large cortical territory.

Extant theories place the DMN as the apex of the sensory-fugal hierarchy or as a parallel network. However, the analyses by Paquola et al. (2021) provide a more nuanced landscape and demonstrate that connectivity is organised along the most prominent cytoarchitectural axis of the DMN, which is not nested within or parallel to the sensory-fugal hierarchy. Instead, the DMN seems to protrude from the sensory-fugal hierarchy, with strong afferent connectivity on one end and insulation on the other. This architecture aligns with a rich club organisation of brain connectivity, in which activation of a single rich club node through feeder connections can ignite meta-stable network dynamics. The areas with convergent afferents, as well as connections within the DMN, may enable recombinations that would not occur within sensory-fugal processing streams. Such topological complexity is thought to be an important trade-off in development and evolution of biological neural networks and illustrates a distinctive role of the DMN in information integration.

By and large, these results seem to provide strong support to the limbic workspace model proposed by Chanes and Barrett (2016) (cfr. Figure E.5) where intrinsic networks are understood as hierarchical systems, with the flow of prediction signals within each network dictated by the structure of the cortical areas involved. In these networks, limbic cortices (e.g., the ventral anterior insula and dorsal anterior cingulate cortex for the salience network and the posterior cingulate cortex and sub/pregenual cingulate cortex for the DMN) issue predictions to better laminated areas in the network. This way, a single network may contain a diverse population of representations across multiple levels of cortical processing.

The limbic cortices that guide allostasis fall within the traditional territory of three intrinsic networks within the brain (Figure E.11).

In a predictive coding perspective, Barrett et al. (2016) argue for the following general functions.

- DMN: generatively uses prior experiences to construct the brain's internal model. If a simulation is an embodied brain state, then the default mode network initiates simulations and represents part of their pattern; its multimodal sensorimotor summaries become more detailed and particularized as they cascade out to primary sensory and motor regions.

- SN: sends predictions that adjust the internal model to the conditions of the sensory periphery, again in the service of allostasis. The SN tunes the internal model by anticipating which prediction errors are likely to be allostatically relevant and therefore worth the metabolic cost of encoding and consolidation, and then modulating the gain on those errors accordingly. These predictions are called precision signals. Precision signals optimize the sampling of the sensory periphery for allostasis. Via their core position in the brain's rich club, and their role in multisensory integration, the SN's precision signals apply attention to every sensory system in the brain (this is sometimes called affective attention). Precision signals directly alter the gain on neurons as they compute prediction error from incoming sensory input. Unexpected sensory inputs that are anticipated to have allostatic implications (because they are likely to impact survival, offering reward

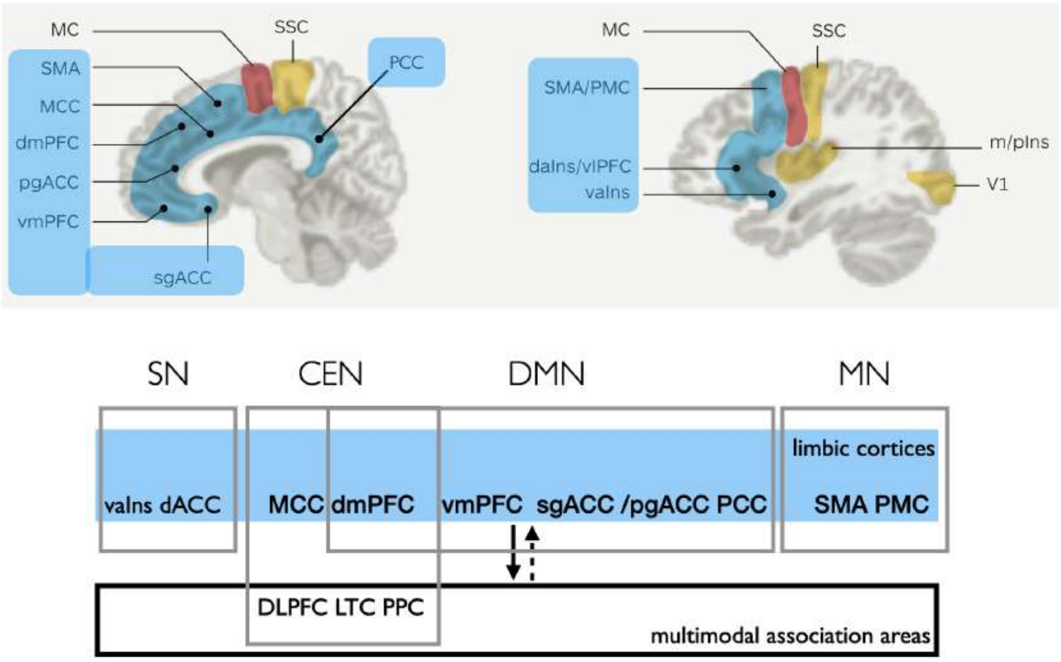


Figure E.11: *Mapping of limbic areas considered in Figure E.4 and E.5 to functional networks: DMN, SN, CEN and motor Network. sgACC, subgenual anterior cingulate cortex; vmPFC, ventromedial prefrontal cortex; pgACC, pregenual anterior cingulate cortex; dmPFC, dorsomedial prefrontal cortex; MCC, midcingulate cortex, is ventral to dmPFC and SMA; vaIns, ventral anterior insula; daIns, dorsal anterior insula; vlPFC, ventrolateral prefrontal cortex; SMA, supplementary motor area; PMC, premotor cortex; m/pIns, mid/ posterior insula (primary interoceptive cortex); SSC, somatosensory cortex; V1, primary visual cortex; and MC, motor cortexAdapted from Barrett et al. (2016)*

or threat, or are of uncertain value) will be treated as signal and learned to better predict energy needs in the future. Importantly, the SN helps accomplish multimodal integration (its spatial topography strongly overlaps with the multimodal integration network). Moreover, primary interoceptive cortex (in the dorsal mid to posterior insula) is a component of the SN, ensuring that every mental event (not just emotions) is infused with interoception, which is made available to consciousness as affect. This state of affairs provides the recipe for affective realism, where people experience supposed facts about the world that are created in part by interoception and the associated affective feelings.

- CEN: neurons within the frontoparietal control network sculpt and maintain simulations for longer than the several hundred milliseconds it takes to process imminent prediction errors. They apply attention to adjust the degree of confidence in sensory predictions (i.e., adjusting priors) and they may also help to suppress or inhibit simulations whose priors are very low. It thus support the need to learn on a single trial, without recurring to statistical regularities in the world, in a quickly changing environment or when the prediction error was large. Indeed, as a prediction generator, the brain is constructing simulations across many different timescales (i.e. it is integrating information across the few moments that constitute an event, but also across longer time frames at various scales). The CEN (which contains key limbic rich-club hubs in the mid cingulate cortex and anterior insula) also may have a role to play in managing sensory prediction errors, by applying attention to select those body movements that will generate the expected sensory input. inputs, presumably with help from cerebellar and striatal prediction errors. These movements then generate the sensory inputs that reduce prediction error and confirm an existing prediction.

In its bare essentials, the prediction process dynamics triggered by limbic regions, as generally outlined in Figure E.5 is summarised in Figure E.12. Note that for simplicity sake, gustatory olfactory and intermediate multimodal areas are not included in the scheme.

The primary task of a brain is to implement allostasis in the service of efficient metabolism and energy regulation. For example, allostasis describes the brain's capacity both to predict that to start running requires more oxygen in the body's striate muscles, as well as to mobilize the needed resources by increasing cardiac output, redistributing blood flow from organs that can spare oxygen. These predictions cause changes in the body's internal systems (the immune, endocrine and autonomic nervous systems) and the sensations that arise from those changes are called interoception. The interoceptive prediction signals are represented as a change in affect (i.e. the expected sensory consequences within the body). The skeletomotor prediction signals prepare the body for movement and the extrapersonal sensory prediction signals prepare upcoming perceptions, that is exteroception.

To such purpose, the brain hosts an internal model of the world from the perspective of its body's physiological needs, following the well-known cybernetics principle that anything which regulates a system must contain an internal model of that system. Limbic cortices initiate allostatic predictions to the hypothalamus and brainstem nuclei (e.g. periaqueductal grey, para-brachial nucleus, nucleus of the solitary tract) to regulate the autonomic, neuroendocrine and immune systems. The incoming sensory inputs from the internal milieu of the body are carried along the vagus nerve and small diameter C and Ad fibres to limbic regions (dotted lines). Comparisons between prediction signals and ascending sensory input results in prediction error that is available to update the brain's internal model. In this way, prediction errors are learning signals and can adjust subsequent predictions. Efferent copies of allostatic predictions are sent to motor cortex as motor predictions (solid lines) and prediction errors are sent from motor cortex to limbic cortices (dotted lines). Sensory cortices receive sensory predictions from several sources. They receive efferent copies of allostatic predictions and efferent copies of motor predictions. Sensory cortices with less well-developed lamination (e.g. primary interoceptive cortex) also send sensory predictions to sensory cortices that are more well developed (e.g. somatosensory and primary visual cortices). The cerebellum models sensory prediction errors from the periphery and relays them to cortex to rapidly modify motor predictions (i.e. it is hypothesized to predict the sensory consequences of a motor command much faster than actual sensory prediction errors can be received, and helps the cortex reduce the sensory consequences caused by one's own movements); it may have the same role to play for allostatic predictions given the connectivity between the cerebellum and cingulate cortices, hypothalamus and the amygdala. It is worth noting that, different from traditional "faculty"-based

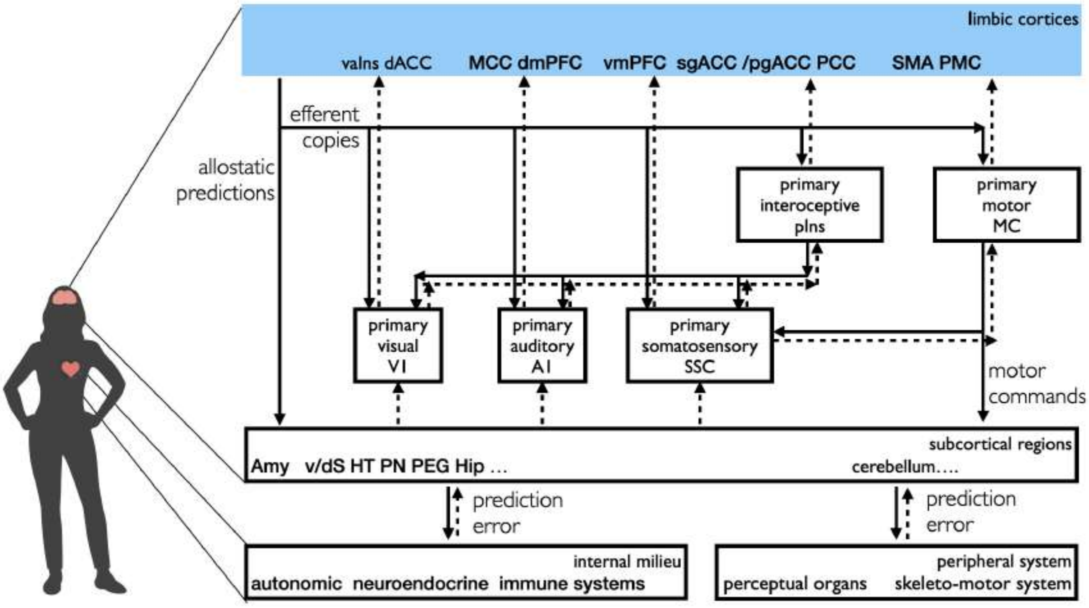


Figure E.12: *Predictive coding in the human brain. Key limbic cortices (in blue) provide cortical control of the body's internal milieu and peripheral systems. Downward arrows represent predictions; upwards dashed arrows, prediction errors. sgACC, subgenual anterior cingulate cortex; vmPFC, ventromedial prefrontal cortex; pgACC, pregenual anterior cingulate cortex; dmPFC, dorsomedial prefrontal cortex; MCC, midcingulate cortex, is ventral to dmPFC and SMA; vaIns, ventral anterior insula; daIns, dorsal anterior insula; vlPFC, ventrolateral prefrontal cortex; SMA, supplementary motor area; PMC, premotor cortex; m/pIns, mid/ posterior insula (primary interoceptive cortex); SSC, somatosensory cortex; V1, primary visual cortex; and MC, motor cortex. Adapted from Barrett et al. (2016)*

views, information flowing from the amygdala to the cortex is not emotional *per se*, but signals uncertainty about the predicted sensory input (via the basolateral complex) and helps to adjust physiological functions in support of allostasis (via the central nucleus). The arousal signals that are associated with increases in amygdala activity can be considered as learning signals. Similarly, prediction errors from the ventral striatum to the cortex (referred to as reward prediction errors) convey information about sensory inputs that impact allostasis more than expected (i.e. indicating that this information should be encoded and consolidated in the cortex, and acted upon immediately). Dopamine is hypothesized to support vigorous action and learning that is necessary to secure the rewards that maintain efficient allostasis (or restore it in the event of disruption), rather than playing a necessary or sufficient role in rewards themselves. Other neuromodulators, such as opioids, may be more intrinsically rewarding.

Figure E.13 outlines an expanded view of the predictive dynamics shown in Figure E.12

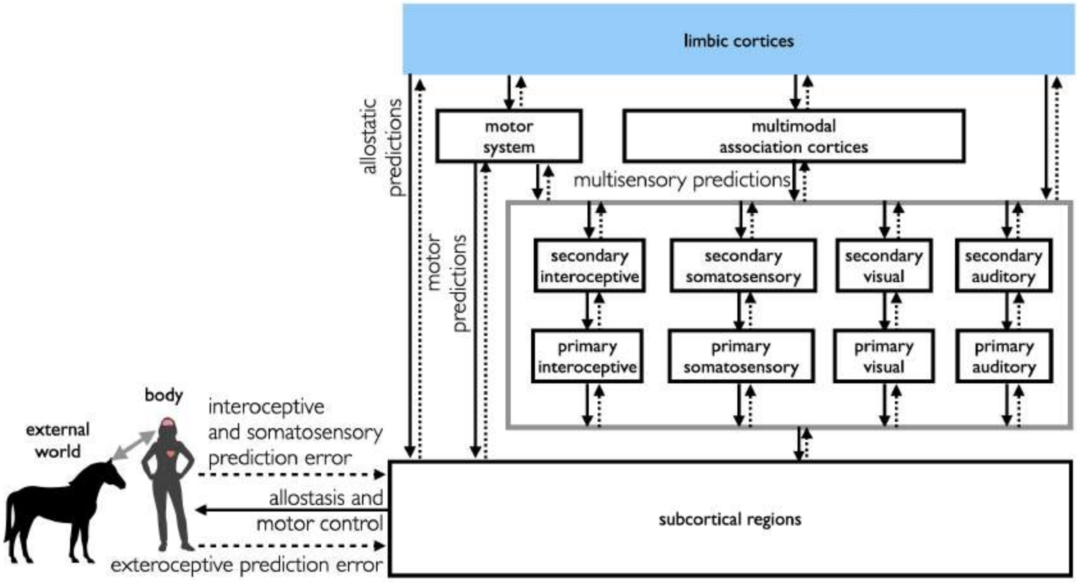


Figure E.13: *Predictive coding in the human brain: expanded from Figure E.12 Downward arrows represent predictions; upwards dashed arrows, prediction errors.*

Figure E.14 summarizes Figure E.12 and maps at the different processing levels - conceptual, perceptual and corporeal - the main intrinsinc networks accounting for the computations at the conceptual level relying on the multimodal integration of lower level sensory predictions (perceptual level).

In the above schemes, the sub-cortical level, which is at the interface between cortical processing and peripheral/body physiology has been blurred. Albeit being a level of extraordinary complexity, overall it can still be characterised in terms of predictive activity Smith et al. (2017). Figure E.15 shows more detailed characterization of the multi-level control architecture that allows for adaptively coordinated cognitive, affective, autonomic, and behavioral responses that also takes into account, to some detail, the subcortical level, with specific reference to heart-rate (HR) control (Smith et al., 2017)

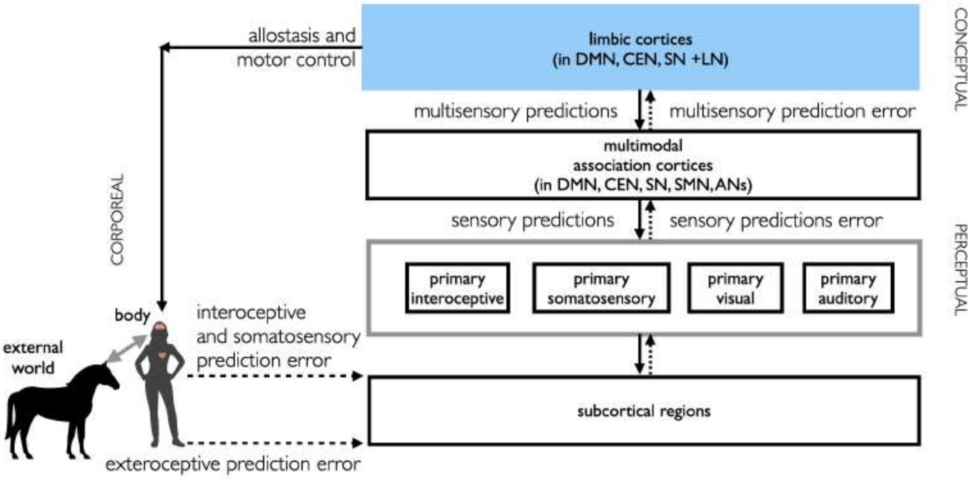


Figure E.14: *Predictive coding in the human brain. The different processing levels - conceptual, perceptual and corporeal - are higlighted. An indicative mapping on the key intrinsic network is provided. DMN, default mode network; CEN, central executive network; SN, saliency network; LN, limbic network; SMN, somatomotor network; ANs, attention networks (dorsal and ventral). Motor control is not shown for simplicity*

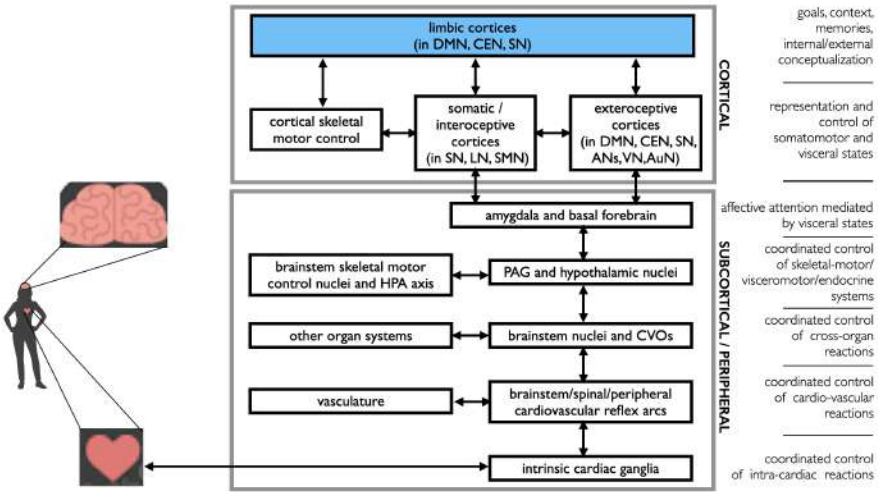


Figure E.15: *The scheme details the coordination of the processing levels from the highest conceptual level (cortical), accounted for by fundamental intrinsic networks down to the lowest (subcortical and peripheral) intra-cardiac and cardiovascular control levels. Double arrows summarise both forward (bottom-up, from periphery to cortex) and backward (top-down, feedback) signals. The right-most side of the scheme highlights the main computational goals at the different levels. DMN, default mode network; CEN, central executive network; SN, saliency network; LN limbic network; ANs, attention networks; SMN, somatomotor network; VN, AuN, visual and auditory networks; PAG, periaqueductal grey; HPA hypothalamic- adrenal-pituitary axis; CVOs, circumventricular organs. Adapted from Smith et al. (2017); Zhang et al. (2019); Chanes and Barrett (2016)*

APPENDIX *F*

The neurobiology of language

Language is a quintessentially human ability. Research has long probed the functional architecture of language in the mind and brain using diverse neuroimaging, behavioral, and computational modeling approaches. However, adequate neurally-mechanistic accounts of how meaning might be extracted from language are sorely lacking. In particular, the debate spins around a fundamental question (Fedorenko and Thompson-Schill, 2014): are some computations unique to human language or can language be solved by more general-purpose mental operations? On one side it has been argued that there is a high degree of functional specificity in the brain regions that support language. On the other, hypotheses have been advanced about putative language regions that are, instead, grounded in domain-general terms. Here, we summarise some results so far achieved.

F.1 The classical view

For more than a century the neurobiological model that has dominated the field was the Wernicke–Lichtheim–Geschwind (WLG) model (Hagoort, 2014, for an introduction) . In this model, the human language faculty was situated in the left perisylvian cortex, with a strict division of labor between the frontal and temporal regions. Wernicke's area in left temporal cortex was assumed to subserve the comprehension of speech, whereas Broca's area in left inferior frontal cortex was claimed to subserve language production. The arcuate fasciculus connected these two areas.

Although Broca's area, Wernicke's area and adjacent cortex are core nodes in the language network, the distribution of labor between these regions is different than was claimed in the WLG model. Lesions in Broca's region are known to impair not only language production but also language comprehension, whereas lesions in Wernicke's region also affect language production. Recently, functional neuroimaging studies pro-

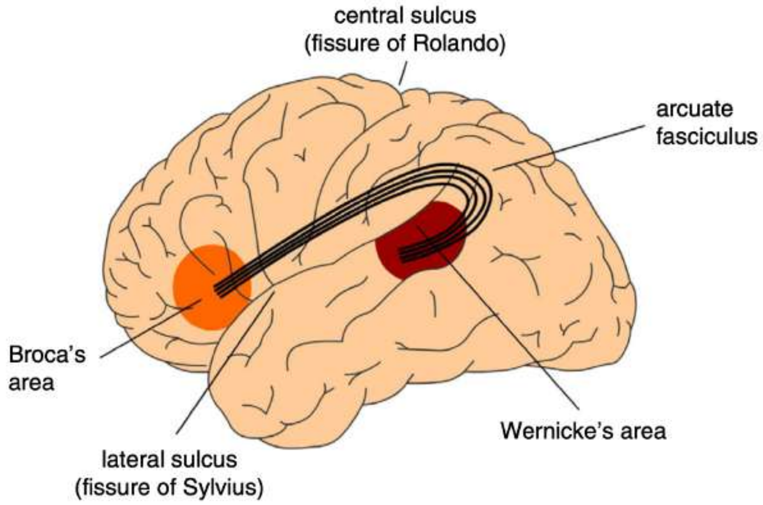


Figure F.1: *The classical Wernicke–Lichtheim–Geschwind model of the neurobiology of language. In this model Broca's area is crucial for language production, Wernicke's area subserves language comprehension, and the necessary information exchange between these areas (such as in reading aloud) is done via the arcuate fasciculus, a major fiber bundle connecting the language areas in temporal cortex (Wernicke's area) and frontal cortex (Broca's area). The language areas are bordering one of the major fissures in the brain, the so-called Sylvian fissure. Collectively, this part of the brain is referred to as perisylvian cortex. From Hagoort (2014)*

vided further evidence that the WLG model is no longer tenable. For example, central aspects of language production and comprehension are subserved by shared neural circuitry. Moreover, the classical model focused on single word processing, whereas a neurobiological account of language should go on beyond production and comprehension of single words

F.2 Departures from the classical view

Sentence processing crucially differentiates three linguistic processing phases after an initial phase of acoustic-phonological analysis.

1. Sentence-level processing: the local phrase structure is built on the basis of word category information.

2. Syntactic and semantic processing of relations in the sentence: these involve the computation of the relations between the verb and its arguments, thereby leading to the assignment of thematic roles (i.e., the analysis of who is doing what to whom); once both semantic and syntactic information lead to the compatible interpretation, comprehension can easily take place.

3. Integration: for sentences in which semantic and syntactic information do not easily map, a final consideration and integration of the different information types is achieved, possibly including the context or world knowledge.

During auditory sentence processing, these three different phases interact with linguistic prosody providing, for example, information about phrase boundaries relevant for syntactic processes. Linguistic prosody can also signal what is in the thematic focus of a sentence and whether an utterance is a declarative sentence or a question. This information is either essential or modulatory to the syntactic and semantic processes in a given sentence.

Different brain regions in the left and right hemisphere have been identified to support particular language functions (see Figure F.2). At the most general level:

- networks involving the temporal cortex and the inferior frontal cortex with a clear left lateralization were shown to support syntactic processes;
- less lateralized temporo-frontal networks subserve semantic processes.

Within dual stream models, the ventral pathway has been taken to support sound-to-meaning mapping, whereas the dorsal pathway connecting the posterior dorsal-most aspect of the temporal lobe and the posterior frontal lobe has been suggested to support auditory-motor integration.

As to the latter, it has been argued (Friederici, 2011) that projections from sensory to the premotor cortex (via dorsal pathway I) could support bottom-up information processes, whereas projections from Broca's area to the temporal context (via dorsal pathway II) could subserve top-down processes drawing prediction about the incoming information, thereby easing its integration.

More specifically one might distinguish the following processes and involved areas (for details, see Friederici, 2011, and for areas refer to Figure F.2.

- **Acoustic-phonological analysis**. This process is the by the auditory cortex and adjacent areas such as the Heschls gyrus (HG). A primary step is to differentiate speech from non-speech acoustic signals, This primary auditory analysis is computed in HG. The planum temporale (PT) has been proposed as the region for the segregation and matching of spectrotemporal patterns and as serving as a hub gating the information to higher-order cortical areas. Speech perception of phonemes (consonants) was found to activate a region anterolateral to HG in the STG/STS.
- **Initial syntactic processes**. it has been suggested that the frontal operculum (FOP) together with the anterior STG supports local structure building. More generally, this net- work could be viewed as the system that supports rule-based combinatorics of adjacent elements. However, Studies investigating sentence processing under less proficient processing conditions as in language development and second language learning show that processing phrase structure violations involves the IFG, in particular Broca's area, and not just the FOP. This suggests that there may be a shift in the recruitment of necessary parts of the ventral prefrontal cortex for local syntactic structure building as a function of language proficiency.
- **Computation of semantic and syntactic relations**. Many of the neuroimaging studies on language comprehension report activation in the anterior and posterior temporal lobe. While some studies concluded that the anterior and posterior temporal regions react specifically to semantic or syntactic aspects, others challenged

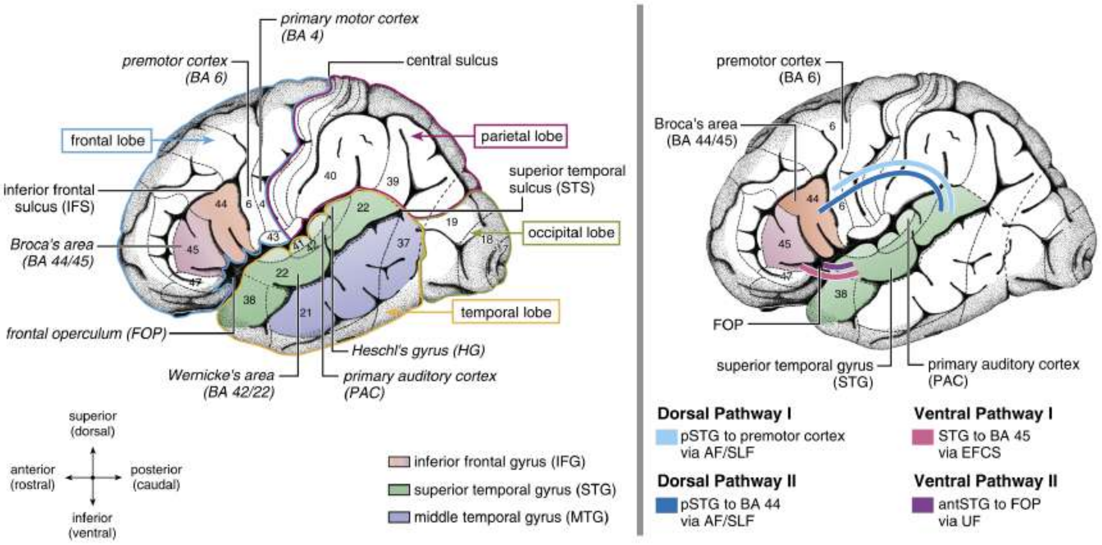


Figure F.2: *Neurobiology of language.* ***Left panel****: Anatomical and cytoarchitectonic details of the left hemisphere. The different lobes (frontal, temporal, parietal, occipital) are marked by colored borders. Major language relevant gyri (IFG, STG, MTG) are color coded. Numbers indicate language-relevant Brodmann Areas (BA). The coordinate labels superior/inferior indicate the position of the gyrus within a lobe (e.g., superior temporal gyrus) or within a BA (e.g., superior BA 44; the superior/ inferior dimension is also labeled dorsal/ventral). The coordinate labels anterior/posterior indicate the position within a gyrus (e.g., anterior superior temporal gyrus; the anterior/posterior dimension is also labeled rostral/caudal). Broca's area consists of the pars opercularis (BA 44) and the pars triangularis (BA 45). Located anterior to Broca's area is the pars orbitalis (BA 47). The frontal operculum (FOP) is located ventrally and more medially to BA 44, BA 45. The premotor cortex is located in BA 6. Wernicke's area is defined as BA 42 and BA 22. The primary auditory cortex (PAC) and Heschl's gyrus (HG) are located in a lateral to medial orientation.* ***Right panel****: Structural connectivities between the language cortices. Schematic view of two dorsal pathways and two ventral pathways. Dorsal pathway I connects the superior temporal gyrus (STG) to the premotor cortex via the arcuate fascile (AF) and the superior longitudinal fascicle (SLF). Dorsal pathway II connects the STG to BA 44 via the AF/SLF. Ventral pathway I connects BA 45 and the temporal cortex via the extreme fiber capsule system (EFCS). Ventral pathway II connects the frontal operculum (FOP) and the anterior temporal STG/STS via the uncinate fascile (UF). Adapted from Friederici (2011)*

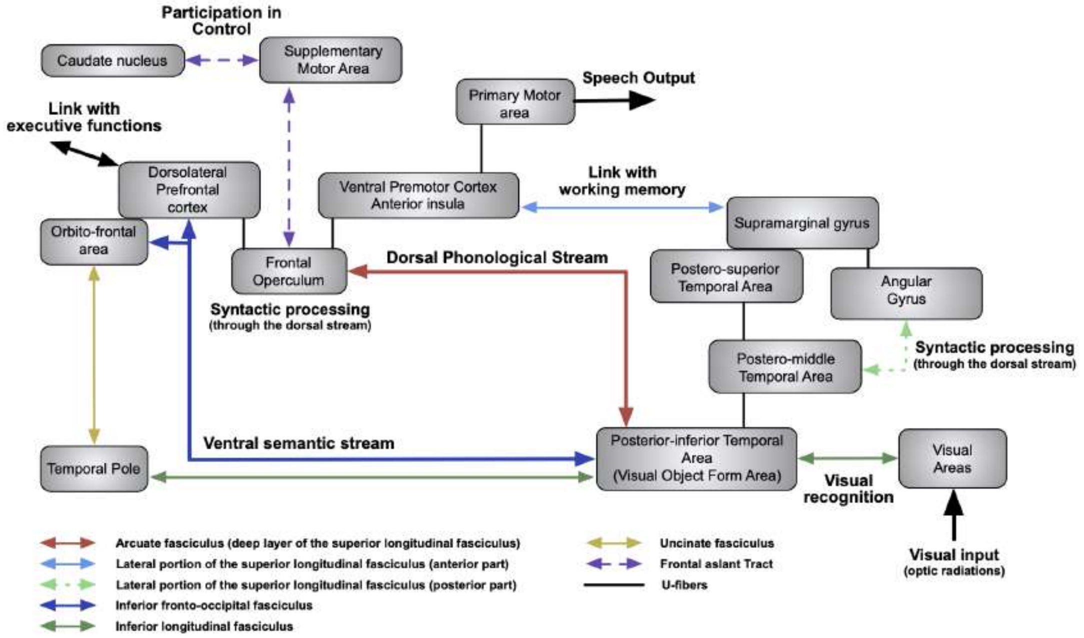

Participation in Control
Caudate nucleus
Supplementary Motor Area
Speech Output
Primary Motor area
Link with executive functions
Dorsolateral Prefrontal cortex
Ventral Premotor Cortex Anterior insula
Link with working memory
Supramarginal gyrus
Orbito-frontal area
Frontal Operculum
Dorsal Phonological Stream
Postero-superior Temporal Area
Angular Gyrus
Syntactic processing (through the dorsal stream)
Postero-middle Temporal Area
Syntactic processing (through the dorsal stream)
Ventral semantic stream
Posterior-inferior Temporal Area (Visual Object Form Area)
Visual Areas
Temporal Pole
Visual recognition
Visual input (optic radiations)
Arcuate fasciculus (deep layer of the superior longitudinal fasciculus)
Lateral portion of the superior longitudinal fasciculus (anterior part)
Lateral portion of the superior longitudinal fasciculus (posterior part)
Inferior fronto-occipital fasciculus
Inferior longitudinal fasciculus
Uncinate fasciculus
Frontal aslant Tract
U-fibers

Figure F.3: *The dual-route model of language processing. This model is based on double anatomo-functional dissociations obtained with cortico-subcortical electrostimulation during a naming task (visual input). Adapted from Herbet and Duffau (2020)*

this view by arguing either that the anterior temporal lobe or the posterior temporal lobe is not domain specific. The anterior STG is systematically involved whenever syntactic structure has to be processed (sentences versus word lists). the posterior temporal cortex (posterior STG/STS) is clearly involved in language processing, and its function appears to be primarily to integrate different types of information. For sentence processing, this might mean the integration of semantic and syntactic information. The IFG, in particular Broca's area, has long been known to support language production and comprehension processes. Subregions of Broca's area have been allocated to different aspects of language processing, either seeing BA 44 as supporting syntactic structure building, BA 44/45 as supporting thematic role assignment and BA 45/47 supporting semantic processes (67), or specifying Broca's area (BA 44/45) as the region supporting the computation of syntactic movement (96), or defining Broca's region (BA 44/45/47) as the space for the unification of different aspects in language. In general, when semantic processing demands increase due to task or stimulus configurations, more anterior portions of the IFG are recruited. When perceptual processing conditions induce increased demands during syntactic processes, more posterior/superior regions of the IFG towards the IFS are recruited. The data thus point towards a language processing system which allocates different subregions in the perisylvian default language network as needed. The role of Broca's area as a central region for syntactic processes has also been demonstrated in the context of artificial grammar learning.

- **Integration**. This is matter of debate. Some researchers (assume that the final integration of syntactic and semantic information takes place in the left posterior STG, whereas others assume that unification of different language-relevant information types is located in the left IFG. In general, IFG's role as a region of combining semantic and syntactic information may be restricted to its more anterior parts.

- **Prosody**. Two types of prosodic information are usually distinguished: emotional prosody and linguistic prosody. As to linguistic prosody, is mainly encoded in the intonational contour, which signals the separation of constituents (syntactic phrases) in a spoken sentence and the accentuation of (thematically) relevant words in a speech stream. Studies (e.g., Ref. 200) suggest a relative involvement of the RH (right emisphere). The less segmental information there is available, the more dominant the RH. For example, processing of pitch information (intonational contour) is correlated with an activation increase in the RH (but can be modulated by task demands)

Speech (and reading) is a temporal process and, beyond the above specific area/-function characterisation, it is of interest to understand how speech comprehension evolves over time. Unfortunately, the data available from the fMRI studies on language processing do not provide the sufficient time resolution to capture this crucial aspect. The cognitive description of the comprehension process itself consists of several subprocesses that take place in a serial cascading and partly parallel fashion. Three linguistic processing phases have been assumed, and these correlate with functionally distinct components identified in the electrophysiological signal.

In the last decades, different language-relevant event-related brain potential (ERP) components have been identified: an early left anterior negativity (ELAN) between 120 and 200 ms, taken to reflect initial syntactic structure building processes; a centroparietal negativity between 300 and 500 ms (N400), reflecting semantic processes; and a late centroparietal positivity (P600), taken to reflect late syntactic processes. Moreover, in the time window be- tween 300 and 500 ms, a left anterior negativity (LAN) was observed to syntactic features that mark the grammatical relation between arguments and verb, and this was taken to reflect the assignment of thematic relations (who did what to whom) (see Figure F.4). This led to the formulation of the so-called three-phase model of language comprehension allocating different components in the event-related brain potential to different processes in the comprehension process. Recently, the process has been revised, but the different ERP components are still observed during language processes.

The neurotemporal dynamics of language comprehension can be described as follows, from a feedforward perspective.

- **Phase 1**: An initial phrase structure on the basis of word category information is built. This process is highly automatic, independent of semantic and verb argument information, and independent of task demands. The process involves a portion of the left STG immediately anterior to the primary auditory cortex, possibly connecting to the FOP located ventrally to Broca's area.

- **Phase 2**:the relation between the verb and its arguments is computed to assign the thematic roles in a sentence. Morphosyntactic information (subject-verb agree-

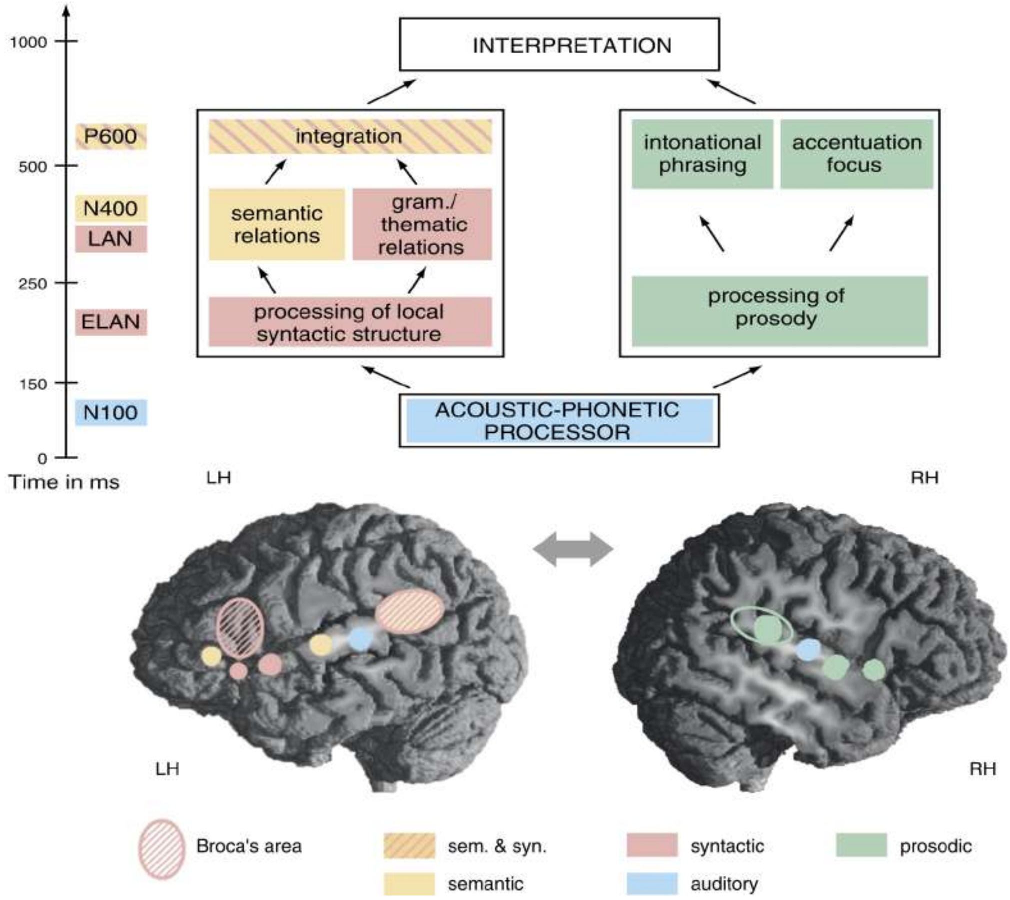


Figure F.4: *The time course of language comprehension. Top panel shows the different stages of the events and processes involved: acoustic-phonological processes (N100); initial syntactic Processes (ELAN); computation of syntactic and semantic Relations (LAN/N400); integration and interpretation (P600). Th bottom panel shows main regions involved in the left hemisphere (LH, language) and the right hemisphere (RH, prosody). Adapted from Friederici (2011)*

ment, LAN), case information (LAN or N400, depending on the particular language), and lexical selectional restriction in- formation (N400) are taken into consideration to achieve assignment of the relation between the different elements in a sentence. The on-line assignment of semantic relations mainly appears to involve the mid and posterior portion of the temporal cortex. Processes of subject-verb agreement have not been clearly localized, but the distribution of the LAN suggests an involvement of the left frontal cortex.

- **Phase 3**: the final interpretation takes place, with semantic and syntactic information being taken into account and mapped onto world knowledge. At the linguistic level, the difficulty of integrating syntactic and semantic information and the need for reanalysis is reflected in a P600. The difficulty of mapping linguistic information onto world knowledge also appears to elicit a P600 effect. At the moment it remains open whether these two P600 effects are members of the same family of ERP components or not.

F.3 In search of a language network

Prior investigations of functional specialization have focused on the response profiles of particular brain regions. Beyond the controversial idea of localizing specific functions within specific areas, a cogent problem, in our perspective, is whether a "language network" can be defined, in the vein of other functional networks we have discussed so far.

As pointed out by Fedorenko and Thompson-Schill (2014), one might object that questions about the language network are ill posed, because language is not a single thing. Indeed, when talking about whether language relies on domain-specific versus domain- general machinery (or some combination of the two), researchers are often referring to different mental processes that language encompasses and there is no agreement on the right ontology of these processes. Such ontologies in human cognitive neuroscience are typically inspired by theoretical and experimental behavioral work in psychology and cognitive science, although they often lag behind. At present, based on differences in functional profiles and some neuropsychological patient evidence, we can at least distin- guish between: (i) the sensory language regions (in the auditory and visual cortices); (ii) the speech articulation regions; and (iii) the 'higher-level' language-processing regions (Figure F.5).

It should be clear that the right criterion (or set of criteria) for the language network is subject to debate. If we focus on the properties of the individual nodes of the network, one could argue that either: (i) the language network is not functionally specialized for language because not all of its nodes are functionally specialized for language; or (ii) the language network is functionally specialized, if the presence of some specialized nodes is sufficient. If we focus on the edges, the language network would qualify as functionally specialized because the specialized regions would be engaged only during language-processing tasks and thus, by definition, the combination of brain regions (and presumably the connections among them) engaged during language processing would be unique. Indeed, researchers have argued that there is a high degree of functional specificity in the brain regions that support language. Others have advanced hypotheses

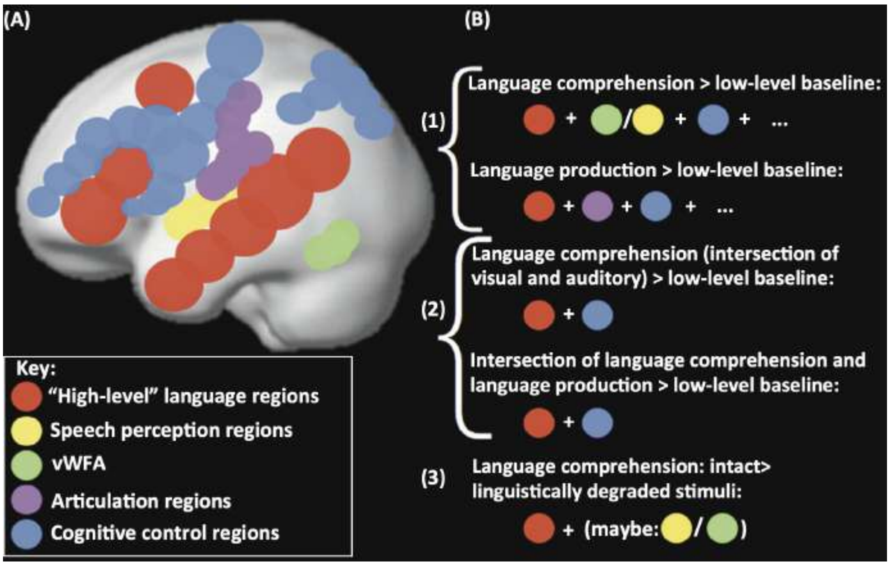

(A)
(B)
Key:
🔴 "High-level" language regions
🟡 Speech perception regions
🟢 vWFA
🟣 Articulation regions
🔵 Cognitive control regions
(1)
Language comprehension > low-level baseline:
🔴 + 🟢/🟡 + 🔵 + ...
Language production > low-level baseline:
🔴 + 🟣 + 🔵 + ...
(2)
Language comprehension (intersection of visual and auditory) > low-level baseline:
🔴 + 🔵
Intersection of language comprehension and language production > low-level baseline:
🔴 + 🔵
(3)
Language comprehension: intact> linguistically degraded stimuli:
🔴 + (maybe: 🟡/🟢)

Figure F.5: *The language network under different definitions. (A) A schematic depiction of five sets of brain regions that are sometimes included in the language network: red, the classic high-level language-processing regions; yellow, speech perception regions; green, visual word-form area; purple, speech articulation regions; and blue, cognitive control regions. (B) A schematic illustration of possible definitions of the language network, ranging from very liberal (1) to more conservative (2 and 3). Adapted from Fedorenko and Thompson-Schill (2014)*

about putative language regions that are, instead, grounded in domain-general terms (Fedorenko and Thompson-Schill, 2014).

To sum up, a language network plausibly includes a functionally specialized "core" (brain regions that coactivate with each other during language processing) and a domain-general "periphery", namely a set of brain regions that may coactivate with the language core regions at some times but with other specialized systems at other times, depending on task demands.

Figure F.6) outlines this extended network. The core areas within the temporal lobe are constrained to the regions 2 (auditory word forms), 3b and 3c (phonological processing), 4b (elementary lexical semantics), and 4b (syntax). As the "margins" of these core functions, three functionally different regions are addressed, referring to (1) parietal and posterior temporal regions contributing to sensorimotor processing, processing of language in context, and theory of mind, (2) inferior temporal-occipital regions that are predominantly linked to visual object representations and their association with language processing, and (3) the temporal pole as a language interface toward the processing of emotions, valence, and social cognition. Such extensions may not be required for very simple language tasks, but they are largely relevant for natural language communication and for understanding sentences in context. They may even be active at the stage of single-word processing, based on the nature of word and concept representations (Hertrich et al., 2020).

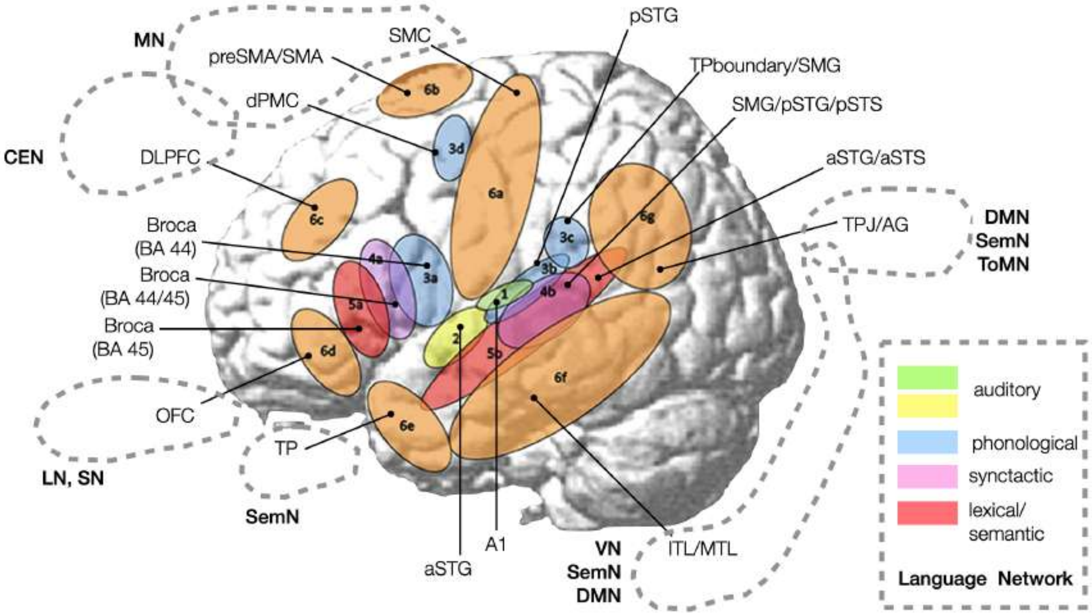


Figure F.6: *Left-hemisphere schematic display of the auditory cortex (region 1), the core language network (2–5, coloured regions as in the legend), and its margins (6, brown regions). Dashed regions show possible connections to main intrinsic networks. (1) auditory cortex A1 as the primary input structure for verbal communication, (2) auditory word form area as a perceptual core region of the language modality, (3) phonological areas linking an auditory-phonetic to an articulatory language code (4) syntax processing, manipulating and detecting structures above word level, (5) lexical-semantic core areas linking phonological codes to lexical meanings, (6a) Sensorimotor cortex, (6b) Supplementary motor area (SMA) and pre-SMA, (6c) dorsolateral prefrontal cortex, (6d) orbitofrontal cortex, (6e) temporal pole, (6f) middle and inferior temporal regions, (6g) parietal and temporoparietal regions. Not shown in this figure, but also relevant for language processing, are right-hemispheric areas homologues of left-dominant language areas, subcortical regions including the basal ganglia, cerebellum, and thalamus, and inner regions of the cortex comprising insula and cingulate cortex. Adapted from Hertrich et al. (2020)*

F.4 Neurobiology of semantics

The extended network shown in Figure F.6 has been proposed as the basis of the brain semantic network.

While the core language network as defined here is largely restricted to phonological and elementary lexical-semantic functions, semantic processing, as a whole, comprises a huge network that is deeply embodied in various ways. It includes all kinds of world knowledge and comprises multiple areas in the brain such as modality-specific representations, sensorimotor regions, and emotion systems Hertrich et al. (2020). Furthermore, convergence zones toward more generalizing and abstract categories in temporal and inferior parietal regions play an important role for semantic processing as well as dorsomedial and inferior prefrontal cortices, controlling the goal-directed activation and selection of semantic information. Figure F.6 also highlights the possible connections to main intrinsic networks.

Based on structural connectivity analyses, three major subcomponents of the semantic system have been outlined comprising (1) a large-distance orbitofrontal-temporal-occipital network assembling object properties, (2) a middle and inferior frontal-subcortical module serving executive control of semantic processing, and (3) a medial temporal module as an interface to episodic memory Hertrich et al. (2020). Regarding object representations, the semantic system is organized in a system of gradients in cortical features from sensory and sensorimotor to transmodal areas. The medial temporal module of the semantic system overlaps with the hippocampal-cortical memory system as a general interface for memory storage, management, and retrieval. Regarding memory content, there seems to be a lateral-medial gradient in semantic representations where lateral regions relate to external knowledge and processes while medial regions relate to self-processing and autobiographic episodic memory.

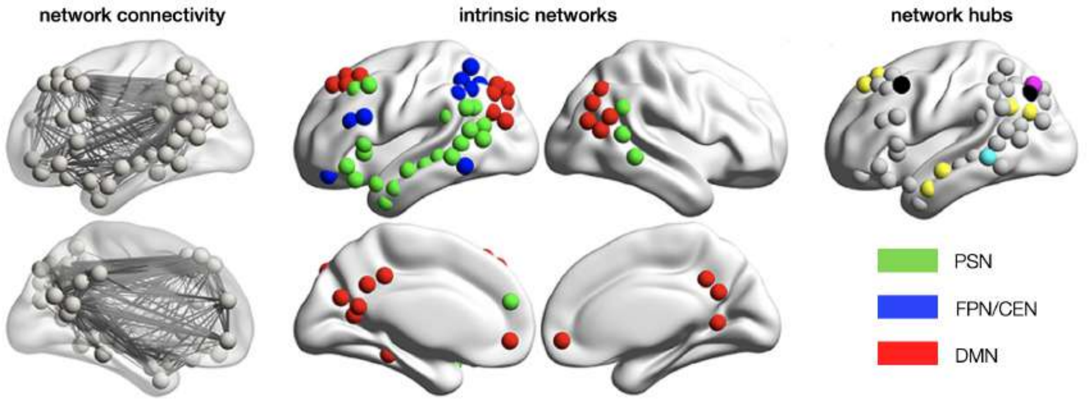


Figure F.7: *The organization of the intrinsic functional network of semantic processing. Left: the semantic network showing nodes and edges, with nodes defined as the regions consistently activated during semantic processing (obtained from a meta-analysis), and edges defined as the resting-state functional connectivity strength; Middle: the components of the semantic network obtained by applying a graph-theoretic approach to the underlying connection patterns. Right: The connector hubs linking the three modules. Adapted from Xu et al. (2017)*

A semantic model has been proposed by Xu et al. (2017) grounded in three functional networks as the basis of semantic processing, comprising (cfr., Figure F.8) (1) the perisylvian "language-supported system" (partially overlapping with the core language network), (2) the "multimodal experiential system" also addressed as the DMN, integrating experience-based knowledge across multiple modalities (see below), and (3) the left-dominant frontoparietal CEN as a semantic control system.

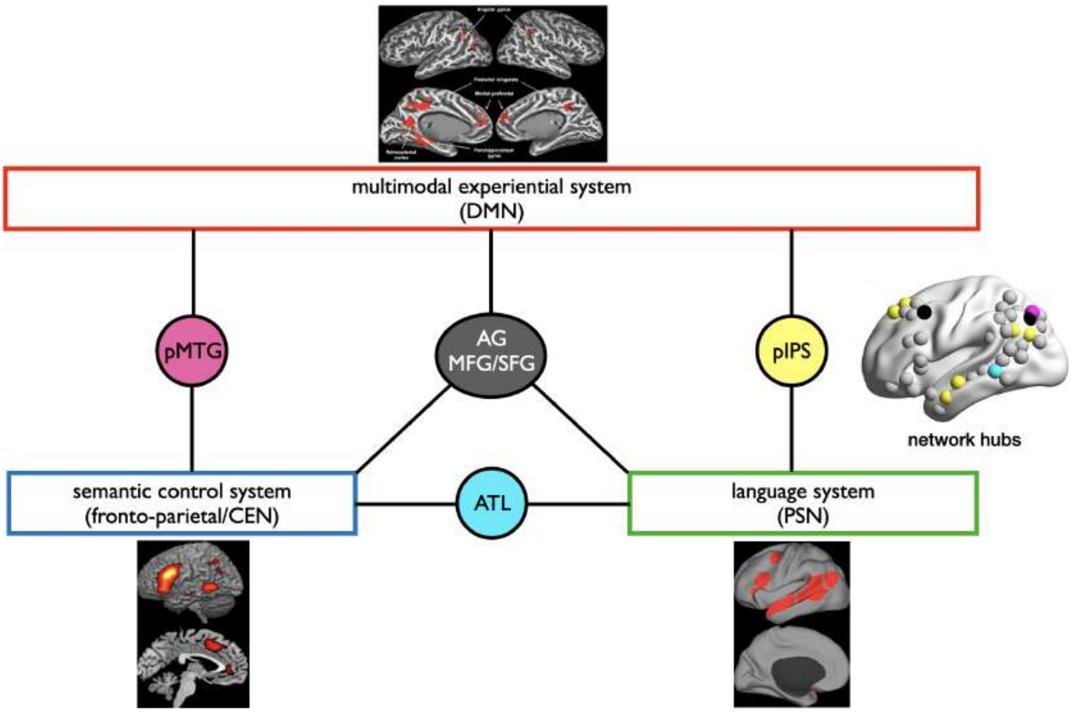


Figure F.8: *The schematic presentation of the tri-network neurocognitive model of semantic processing proposed by Xu et al. (2017). Left frontoparietal central executive network, CEN; DMN, default mode network; PSN, perisylvian network; pMTG, posterior middle temporal gyrus; ATL, anterior temporal lobe; pIPS, posterior intraparietal sulcus; AG, angular gyrus; SFG/MFG, superior and middle frontal gyrus. Adapted from Xu et al. (2017)*

These three networks are linked together in hub regions, comprising the anterior temporal lobe, posterior middle temporal gyrus, posterior intraparietal sulcus, angular gyrus, and parts of superior and middle frontal gyrus (Figure F.8). In general, depending on task demands, for semantic processing various memory systems may be recruited and temporarily linked together, or single subsystems can locally get expanded or diminished.

The distinction between concrete and abstract concepts seems to be of particular interest, being closely related to the nature of language and its sensory embodiment or disembodiment. So the embodiment of abstract in comparison to concrete concepts is more complex, abstract items are more related to emotional processing, they have a stronger representation in left inferior frontal gyrus and left temporoparietal cortex, and they are characterized by longer processing time and different electrophysiological responses in the N400 domain and later potentials Hertrich et al. (2020).

A related tri-model, $L \cup M$ (standing for Language/union/Memory) has been recently proposed by Roger et al. (2022). This tripartite model is organized around three latent variables/dimensions. The Receiver-Transmitter (RT System) – dimension mainly encompasses aspects related to speech perception, phonology, articulation, and syntax. In addition, and even if to a lesser extent, working memory and comprehension saturate this dimension. This suggests that RT comprises processes related to the externalization of verbalizable outputs, implying spell-out and sensory inputs influencing all cognitive processes and the outputs. RT may involve perceptuo-motor information processing operations.

The second dimension, labeled Controller-Manager (CM System) is represented primarily by verbal working memory and comprehension. This component is also more broadly related to articulation, phonology, syntax, associative memory, and lexical access/retrieval. Thus, it could refer to the controlled assembly of elementary operations allowing to transform a verbal input actively into an elaborated and appropriate verbal output (i.e., the accurate mapping between meaning and sounds or, conversely, between sounds and meaning; between word and signification or between sentences/ discourse and meaning, depending on the level of processing). Concretely, incremental binding, monitoring, evaluation, or (error-) prediction operations can be engaged as active inference algorithms (i.e., predicting future states according to the trajectory defined by a given policy). Thus, CM would engage operations common to language production and comprehension.

The third dimension covers neurocognitive aspects related to language comprehension, associative memory, lexical access/retrieval, verbal semantic, episodic, and working memories. In a simplistic way, it can be described as a "Transformer-Associative" computational component (TA System) as it includes computations to build and maintain mental, conceptual, and multimodal representations. The op- erations underlying TA encompass, for instance, abstraction/dimension- ality reduction, multimodal/relational binding, pattern separation/ completion, and replay.

All together, the three dimensions support semantic encoding (Language $\rightarrow$ Memory) and decoding (Memory $\rightarrow$ Language).

Neurally, the SN presents the functional properties to support the (RT System), accompanied by information from the sensory-motor network (SMN, including mainly motor and auditory-perceptive networks).

The Controller-Manager dimension primary function is the organization, development, and maintenance of verbal representations; as such it would be underpinned by a top-down controlled network, the CEN. More precisely, the fronto-parietal control network (FPN), especially lateralized in the left hemisphere.

Processes related to the Transformer-Associative dimension mainly involve the highly integrative DMN.

Continuous interactions between the three networks, performed in parallel and executed in a more or less controlled manner, support the global $L \cup M$ workspace. The a "networks ballet" (Roger et al., 2022), happening through SN- CEN-DMN transitions and dynamical synchronizations, can actively and synergistically support the $L \cup M$ cognitive states and the "common ground" (individuals communicate by relying on the shared set of beliefs, ideas, and knowledge while also making assumptions about the interlocutors' perspectives).

At the same time, sites, located at the crossroads of the leading networks, present

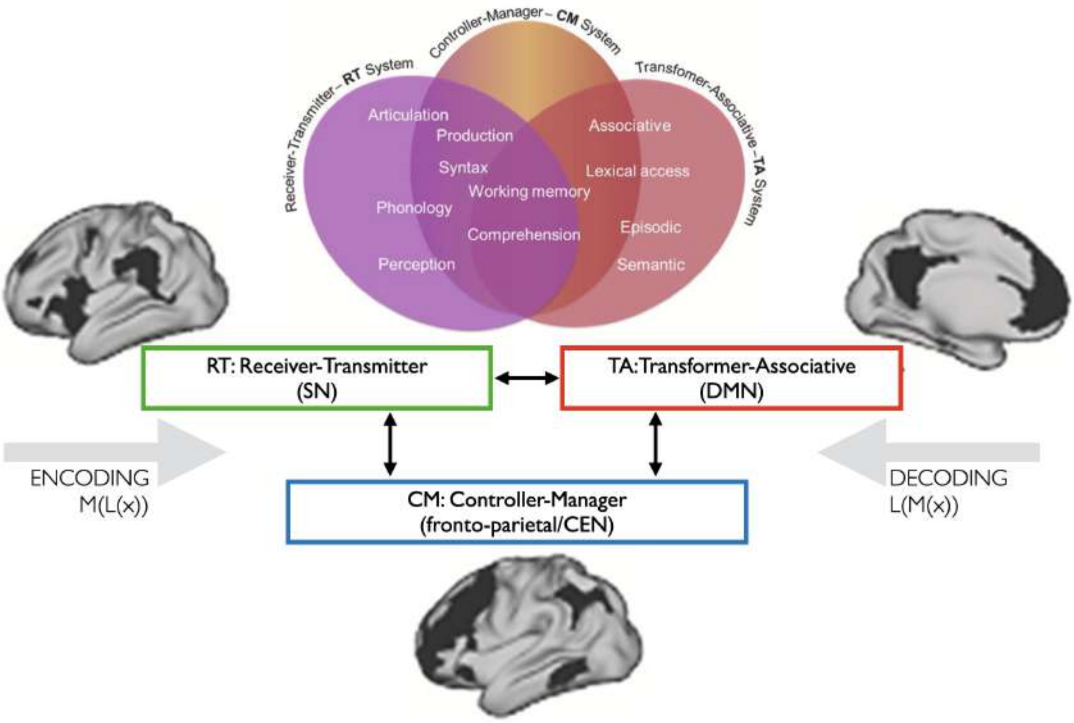

Controller-Manager – CM System
Receiver-Transmitter – RT System
Transfomer-Associative – TA System
Articulation
Production
Syntax
Working memory
Phonology
Comprehension
Perception
Associative
Lexical access
Episodic
Semantic
RT: Receiver-Transmitter (SN)
TA: Transformer-Associative (DMN)
CM: Controller-Manager (fronto-parietal/CEN)
ENCODING M(L(x))
DECODING L(M(x))

Figure F.9: *The $L \cup M$ model. Top panel: the neurocognitive overlap between language and memory according to the main latent dimensions in the form of a Venn diagram. The diagram is composed of three subsets that are both distinct and interrelated . The encapsulation of these modules forms the union of language-memory behaviors, while the overlaps form the language-memory intersection. The three dimensions have been labeled: "Receiver-Transmitter (RT System) - Controller-Manager (CM) - Transformer-Associative (TA)". Bottom panel: Latent dimensions (RT-CM-TA) are individually associated with specific brain networks (SN-CEN-DMN, respectively). In terms of behavior, internal encoding implied in verbal comprehension, for example, consists of encoding declarative inputs (engaging the TA System) via more or less attentive listening of verbal indications (involving the RT and CM dimensions). Here, language feeds memory ($M(L(x))$). Decoding or externalization, leads to the production of language involving a mapping of internal verbal representations and thoughts (TA System) with the corresponding ordered output forms (thus involving manipulation of Systems CM and RT up to verbal evocation). Here, memory feeds language ($L(M(x))$). Adapted from Roger et al. (2022)*

essential properties to act as connector hubs that are core regions able to integrate information from the different networks locally. Among these convergence areas, the inferior frontal gyrus (IFG complex) follows a SN-FPN-DMN gradient during the transition from pars opercularis to pars orbitalis (Figure F.10. The IFG complex could functionally and gradually integrate phonological, syntactic, and semantic representations (Figure F.10). Similar local gradients exist in the insula, the supramarginal and angular gyrus, the posterior upper/mid temporal gyrus, the supplementary motor area, the dorso-lateral prefrontal cortex, the cerebellum, and the basal ganglia. These local integrators could serve as interfaces to interconnect the different $L \cup M$ dimensions by manipulating external information and internal mental representations. Their role could be particularly crucial when the demand for inter-network connection is reinforced, e.g., during online activity

Finally, the role of peripheral hubs that strengthen intra-network connections is also central. One of these peripheral hubs is the hippocampus linking information from the anterior-posterior DMN regions at rest.

To sum up, the model posits clearly that the spectrum of observable behaviors depends on an embedding of local (regional) and global (states) brain dynamics that support specialized operations (in this case language)

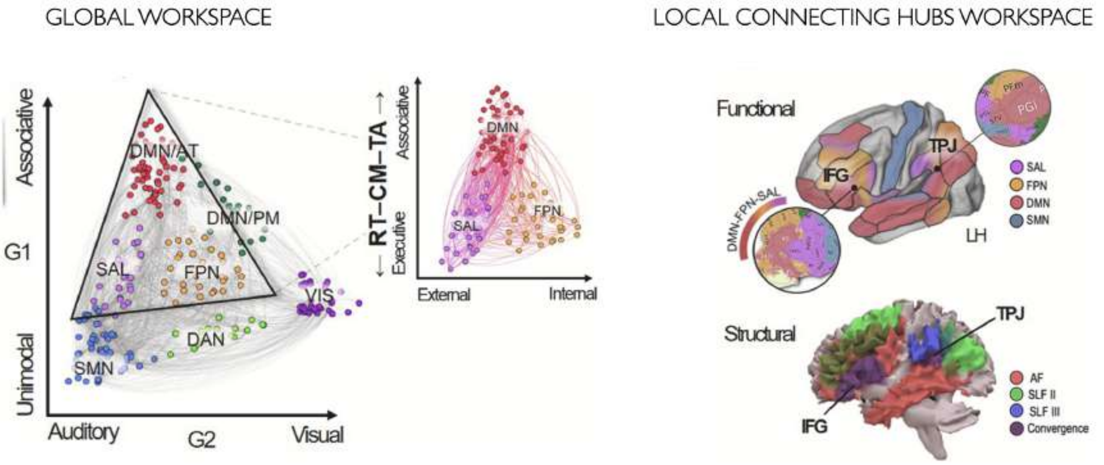


Figure F.10: *Neural workspace of the $L \cup M$ model. Left panel: global functional topography of the links between brain regions belonging to different networks and projected in a reduced space (n = 48 healthy controls, at rest). This global topology corroborates the dimensions and interactions proposed in the $L \cup M$ framework . Right panel: example of functional local SN-FPN-DMN continuums (connector hubs). These functional convergence zones correspond to structural convergence zones where the terminations of traditionally described language and/or memory bundles are intertwined (Arcuate fascicle: AF; and branches II and III of the superior longitudinal fascicle: SLF II-III). Adapted from Roger et al. (2022)*

As a general conclusion, we can say that many current proposals of the neural architecture of language continue to endorse a view whereby certain brain regions selectively support syntactic/combinatorial processing, although the locus of such "syntactic hub", and its nature, vary across proposals. Linguistic theorizing, empirical evidence from language acquisition and processing, and computational modeling have jointly painted a picture whereby lexico-semantic and syntactic processing are deeply inter-connected and perhaps not separable.

In a recent work, Fedorenko et al. (2020) searched for selectivity for syntactic over lexico-semantic processing using a powerful individual-subjects fMRI approach across three sentence comprehension paradigms that have been used in prior work to argue for such selectivity: responses to lexico-semantic vs. morpho-syntactic violations; recovery from neural suppression across pairs of sentences differing in only lexical items vs. only syntactic structure; and same/different meaning judgments on such sentence pairs. Across experiments, both lexico-semantic and syntactic conditions elicited robust responses throughout the left fronto-temporal language network. Critically, however, no regions were more strongly engaged by syntactic than lexico-semantic processing, although some regions showed the opposite pattern. Thus, contra many current proposals of the neural architecture of language, syntactic/combinatorial processing is not separable from lexico-semantic processing at the level of brain regions—or even voxel subsets—within the language network, in line with strong integration between these two processes that has been consistently observed in behavioral and computational language research. These results further suggest that the language network may be generally more strongly concerned with meaning than syntactic form, in line with the primary function of language: to share meanings across minds (Fedorenko et al., 2020).

F.5 Neurobiology of pragmatics

Neurobiological models of language not only need to address the circuitry that is crucial for encoding/decoding the content of an utterance, but they also need to specify the neural infrastructure for inferring what the speaker intended to communicate by uttering a sentence. This is what is broadly referred to as neuropragmatics.

How do brains represent (and share) beliefs, knowledge and components of context in order to infer speaker's meanings and to engage in successful communication? What cognitive functions do pragmatic abilities rely upon? How do they express themselves over time? And what is the cognitive architecture of pragmatics as a system (if a single system can be assumed)?

Language use is an instance of social interaction. Human interactions predominantly involve the dissemination of true or false knowledge for good or for ill (Frith and Frith, 2005). In everyday speech we frequently explain behavior in terms of mental states. The ability to acquire knowledge about other peoples' beliefs and desires is called "mentalizing" or "mind reading". Thus, having a theory of mind enables many important human interactions (Frith and Frith, 2005).

Theory of Mind (ToM) refers to the ability to understand the minds of others, comprising cognitive and affective components.

Through having a ToM we can recognize that another person's knowledge is different from our own. I know what's behind the rock, but he doesn't, because, from where he is, he cannot see that there is a scorpion. Having a theory of mind allows us to manipulate other people's behavior by manipulating their beliefs. If he is my friend I can warn him about the scorpion. If he is my enemy I can tell him it is safe. This latter is called tactical deception or Machiavellianism (Frith and Frith, 2005).

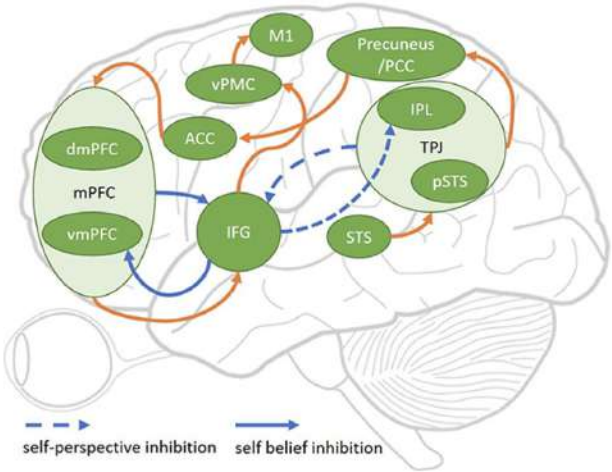
There is currently much interest in identifying a social brain, a circumscribed network of brain regions specialized for the social domain. Mentalizing is one of a number of problems confronting this social brain.

When brain activity is measured during the performance of a wide range of tasks engaging ToM, two regions have been consistently identified: a medial prefrontal region (paracingulate cortex) and the temporo-parietal junction (TPJ) in the superior temporal sulcus (see, Figure F.11).

Specifically, it's a largely bilateral network embedding the TPJ, medial parts of the temporal lobe, the temporal pole, parts of medial frontal cortex, and the precuneus. The medial frontal region is also engaged when subjects reflect upon their own mental states, as well as those of others with the more inferior orbital region responding especially to emotional states. The TPJ seems to have a special role in using perceptual cues to recognize the actions and intentions of biological agents.

Furthermore, various cerebro-cerebellar circuits seem to play a major role for ToM processing, representing a cerebro-cerebellar mentalizing network.

Regarding language functions, the ToM system is primarily engaged in pragmatic processing when individual- or situation-specific meanings must be derived, or when inferences have to be made such as required for understanding indirect requests (Xu et al., 2017). Although the ToM and the language network can be considered as distinct networks, they can get synchronized during language comprehension. ToM processing is also related to the DMN (Xu et al., 2017).



Brain area	Function (simplified)
STS	visually encodes biological motion
TPJ	distinguishes self and others store relevant stimuli
IPL	stores self-relevant stimuli
pSTS	stores other-relevant stimuli
Precuneus	infers visual access
PCC	sends information
ACC	belief reasoning
mPFC	stores beliefs
vmPFC	stores self-belief
dmPFC	stores other's belief
IFG	self-perspective inhibition self-belief inhibition
vPMC	encodes kinematics
M1	motion execution

Figure F.11: *The ToM network. Left: major functional brain areas, pathways, and their interactions. Right: a simplified indicative mapping of major functions that can be related to the components of the network. Adapted from Zeng et al. (2020)*

Among others, an important hub for different network connections for social signalling processes is the TPJ. The TPJ is a variably defined region located roughly where the IPL meets the superior temporal lobe, and is not associated with any objective landmarks. Most investigators would probably define the TPJ as a small region that overlaps only the most ventral part of the IPL at the true intersection of the AG, SMG and posterior superior temporal lobe. However, many other labels are used to describe activations around this region (e.g. IPL, ventral parietal cortex, lateral parietal cortex, AG, SMG, and posterior STS). The ubiquitous use of the term TPJ likely includes the often coactivated posterior superior temporal regions; thus, the compound term “IPL/TPJ” is in many case preferred; even though the IPL and TPJ overlap, even with the most conservative definition of the TPJ, they are not synonymous with each other. As such, three cognitive networks at least overlap the IPL/TPJ: the frontoparietal CEN, the DMN and the cingulo-opercular network (CON). In particular, it appears that the DMN has two distinct network nodes in the IPL/TPJ, one located at the intersection between the AG, SMG and STG, and one located posterior to the previous in the AG. IPL/TPJ nodes have been reported to activate in self-perception, introspection and memory, social cognition and its dorsal component may play a general role across a broad range of task domains.

The tight connection between the ToM and the DMN networks speaks to the general role of the latter. The DMN is, by its very seminal characterisation, sensitive to intrinsic information (long-term memories, conditional responses, beliefs, emotions and so on). But more recently it has been shown that it is an active and dynamic “sense-making” network that integrates incoming extrinsic information with prior intrinsic information over long timescales to form rich, context-dependent, idiosyncratic models of the situation as it unfolds over time.

As the activity in the DMN is shaped by our unique history, it is by nature idiosyncratic. However, at the same time, our knowledge, memories and beliefs are shaped by the people we are connected to and the world in which we are immersed.

It has been argued that the DMN provides a space for social “others” (the extrinsic people we interact with, thus relying on the ToM) to shape the self (our set of intrinsic memories and beliefs), which in turn can enable us to shape the memories and beliefs of others. This unique interplay between the extrinsic and intrinsic forces provides a mechanism for negotiating a shared neural code to facilitate learning and communication via shared common ground. This view is informed by numerous studies using naturalistic stimuli and inter-subject analyses that map shared neural responses across participants. These studies speak for an extensive overlap between the DMN and the “social brain” - the brain regions involved in social cognition (e.g. ToM).

Most intriguing findings across these studies is the discovery of neural patterns in the DMN that are shared across participants and aligned to the abstract structure and interpretation of the external events (Yeshurun et al., 2021). The results suggest that participants who understand the situation in the same way will have similar neural patterns in nodes of the DMN, irrespective of considerable differences in the low-level stimulus properties of the sensory input. At the same time, subtle differences in DMN neural response patterns across participants seem to correlate with subtle differences in their interpretation. Together, these studies suggest that, through the interaction between the intrinsic self and the extrinsic world, the DMN develops a shared neural code (Yeshurun et al., 2021).

Most important here, brain–brain coupling between speakers and listeners has been clearly shown. The responses in the speaker’s brain recorded while telling a personal story in the fMRI scanner are clearly correlated with the listeners’ brain responses while listening to the story. The speaker’s neural responses were coupled (that is, correlated, with or without a short time lag; see Figure F.12) to the listeners’ neural responses in brain regions at various levels of the timescale processing hierarchy. Speaker– listener coupling in early auditory regions reflected shared processing of low-level acoustic properties of the stimulus, such as the audio amplitude envelope. By contrast, in language areas and the DMN, the responses in the listeners' brains lagged behind the responses in the speaker's brain, suggesting that responses in the speaker's brain causally shaped the responses in the listeners' brains. Furthermore, speaker–listener neural coupling in higher-order areas, including the DMN, reflected communication and shared understanding of the narrative. The same situation can be described using spoken words, written text or abstract animated shapes (Yeshurun et al., 2021).

Overall, the communication cycle along a dyadic social interaction can be summarised as in Figure F.13.

Shared neural activity at the sensory level naturally arises from the tendency of these areas to align with the low-level perceptual properties of the external stimuli. Shared neural activity at the top of the processing hierarchy, in the DMN, naturally arises from the tendency of social brains to align thoughts and actions.

As Yeshurun et al. (2021) put it, the DMN is "default' not because it is engaged when we are looking inward, nor because it is shaped by others. The DMN is "default" because it is central for integrating external and internal information, allowing for shared communication and alignment tools, shared meanings, shared narratives and, above all, shared communities and social networks Yeshurun et al. (2021).

F.6 Glossary

- Language: classically defined as a natural, intrinsic, and universal ability of human beings to construct communication systems using codes (speech sounds or written symbols) and to use these codes. Language cognitively involves a semantic system (vocabulary and lexical access), specialized sensory-motor capacities of perception and production (phonology), as well as capacities for decoding, manipulating (grammar/ syntax), and understanding these codes (shared symbolism; comprehension).

- Language production: the physical signal used to transmit language and share thoughts. Speech production would require, among other capacities, syntactic and articulation processes.

- Syntax: rules for organizing elements - word segments, words, sentences into a grammatical discourse - to generate combinatorial and hierarchical structures.

- Verbal comprehension: encompasses various processes helping to construct understandable and meaningful speech productions (expressive language skills up to the pragmatics of language, for instance) and to understand verbal productions (receptive language skills echoing vocabulary or semantics).

- Memory: the ability to maintain information or representations of past experience or knowledge, arguing to be based on mental processes of encoding, retention and retrieval, or reactivation. Several forms of memory have been proposed depending, for example, on the degree of consciousness or attention given to the process (implicit versus explicit memorization) and/or the duration of retention (short-term versus long-term memorization).

- Working memory: positioned between short and long-term memory and concerns the ability to explicitly maintain and manipulate (re-) instantiated information to

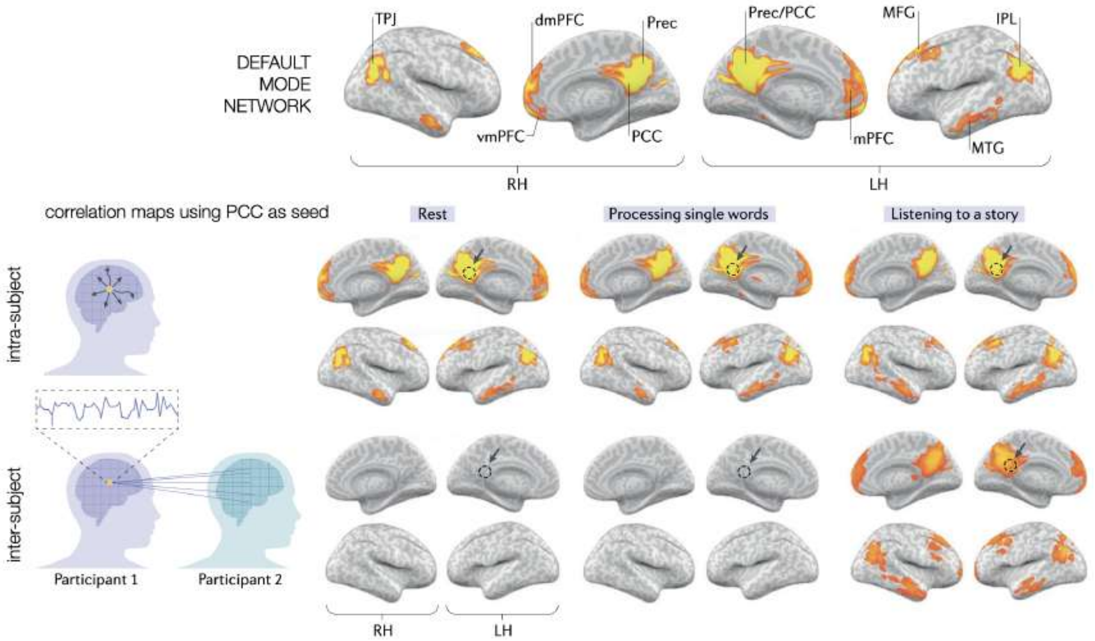


Figure F.12: *Functional correlation within and between subjects at rest or when processing words or listening to a story. Top: Regions of the DMN defined by functional connectivity analysis. These regions include the posterior cingulate cortex (PCC) and precuneus (Prec), the ventromedial prefrontal cortex (vmPFC) and dorsomedial prefrontal cortex (dmPFC), and the bilateral temporoparietal junction (TPJ). IPL, inferior parietal lobule; LH, left hemisphere; RH, right hemisphere; MFG, middle frontal gyrus; MTG, middle temporal gyrus. Center: Within-participant functional correlation maps between the posterior cingulate cortex (PCC) seed (yellow voxel in the schematic; dashed circles in the brain maps) and the whole-brain neural activity. The functional correlation analysis delineates nodes of the default mode network (DMN) in which the activity fluctuates together (co-varies) in a given participant, owing to the direct or indirect anatomical connections during rest (left panel), processing of single words (middle panel) and listening to a coherent story (right panel). Bottom: Inter-subject functional correlation maps between the PCC seed and the whole-brain neural activity observed in other participants. This analysis can filter out spontaneous intrinsic neural facilitation, and as such reveals no substantial stimulus-locked correlations in the DMN during rest (left panel) or during the processing of single words (middle panel). By contrast, however, inter-subject functional correlation exposed stimulus-locked shared responses across participants in the DMN as subjects listen to and process a spoken story minutes long (right panel). LH, left hemisphere; RH, right hemisphere. Adapted from Yeshurun et al. (2021)*

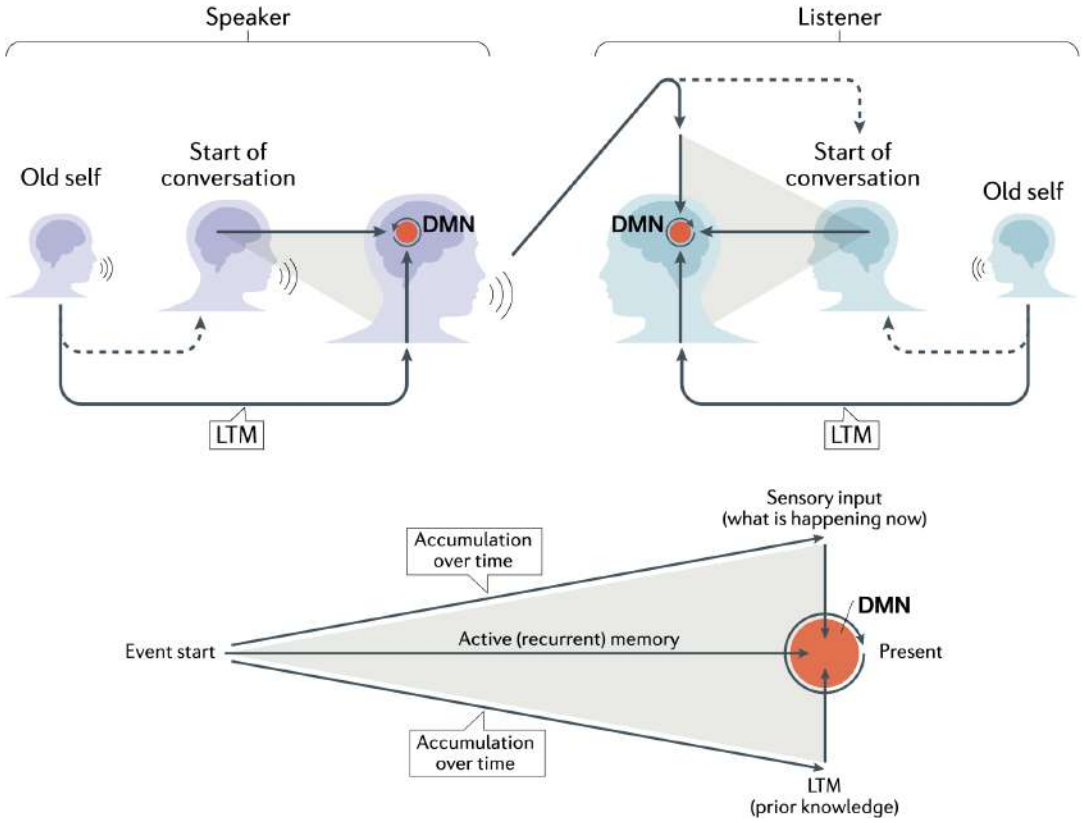


Figure F.13: *Speaker and listener coupled through their default mode network (DMN) activity. Top panel: the interaction unfolding in time; our thoughts, feelings and actions are constantly being shaped by the actions, memories and stories of others. Bottom panel: Activity in the DMN is modulated by incoming external information (top arrow), which is actively accumulated (grey expanding triangle) and integrated (red circle) over hundreds of seconds (horizontal arrow) with our intrinsic information (long-term memories (LTMs), conditional responses, beliefs and so on, represented by the bottom arrow) to form a rich, context-dependent, dynamic model of the unfolding situation. At the same time, our LTMs shape the way we process the external input. This unique interplay between the extrinsic and intrinsic forces provides a space for negotiating a shared neural code necessary for establishing shared meanings, shared communication tools, shared narratives and, importantly, shared communities and social networks. Adapted from Yeshurun et al. (2021)*

perform complex cognitive tasks of learning, reasoning or comprehension. Working memory is generally considered to be part of executive functioning (or central executive system), covering concepts such as planning, inhibition, and mental flexibility.

- Verbal working memory: involves a system for programming the utterance, scheduling verbal items at several levels (words, phonemes and articulatory gestures), and maintaining what needs to be produced (phonological loop and rehearsal); and appears conceptually close to the definition of syntax.
- Declarative memory: involved in maintaining information about facts/knowledge or events for a significant period of time (long-term memory) and consciously recalling information. Declarative memory is classically divided into two subtypes: semantic memory and episodic memory.
- Semantic memory: general and factual knowledge about the world and abstract concepts (noetic consciousness). It allows individuals to make sense of information and/or to engage in cognitive processes such as object recognition or appropriate language use.
- Episodic memory: evokes the memory of personally experienced events associated with a particular time and place (spatiotemporal context), involving a sense of self-awareness (or autonoetic consciousness). In addition to the conscious recall of past events, episodic memory implies a "mental journey through time" (mental time travel, i.e., a projection into the past and/or future).
- Associative memory: retrieval or activation of memories (stimulus, behaviors, facts, events...) conceptually or contextually associated.

Note: Definitions are extracted primarily from the dictionary of the American Psychological Association `https://dictionary.apa.org`.

Bibliography

Abbott, L. and Kepler, T. B. (1990). Model neurons: from hodgkin-huxley to hopfield. In *Statistical mechanics of neural networks*, pages 5–18. Springer.

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.

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.

Alexander, D. M., Trengove, C., Johnston, P., Cooper, T., August, J., and Gordon, E. (2005). Separating individual skin conductance responses in a short interstimulus-interval paradigm. *Journal of neuroscience methods*, 146(1):116–123.

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.

Alves, P. N., Foulon, C., Karolis, V., Bzdok, D., Margulies, D. S., Volle, E., and de Schotten, M. T. (2019). An improved neuroanatomical model of the default-mode network reconciles previous neuroimaging and neuropathological findings. *Communications biology*, 2(1):1–14.

Anderson, D. J. and Adolphs, R. (2014). A framework for studying emotions across species. *Cell*, 157(1):187–200.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Andreas, J. and Klein, D. (2016). Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182.

Angela, J. Y. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692.

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-taylor, J. S. (2008). Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems*, pages 17–24.

Arnold, M. B. (1960). *Emotion and personality.* Columbia University Press.

Atzil, S., Gao, W., Fradkin, I., and Barrett, L. F. (2018). Growing a social brain. *Nature Human Behaviour*, 2(9):624–636.

Augustine, S. (1876). *The confessions.* Clark.

Austin, J. L. (1962). *How to do things with words.* Harvard University Press, Cambridge, MA.

Averill, J. R. (1980). A constructivist view of emotion. In Plutchik, R. and Kellerman, H., editors, *Theories of Emotion*, chapter 12, pages 305–339. Academic Press.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE.

Barbas, H. and García-Cabezas, M. Á. (2016). How the prefrontal executive got its stripes. *Current Opinion in Neurobiology*, 40:125–134.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind.* MIT press.

Barrett, L. F. (2009). The future of psychology: Connecting mind to brain. *Perspectives on psychological science*, 4(4):326–339.

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.* Houghton Mifflin Harcourt, New York, NY.

Barrett, L. F. (2017c). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23.

Barrett, L. F. (2020). *Seven and a half lessons about the brain.* Pan Macmillan, London, UK.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68.

Barrett, L. F. and Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in experimental social psychology*, 41:167–218.

Barrett, L. F. and Quigley, K. S. (2021). Interoception: The secret ingredient. In *Cerebrum: the Dana Forum on Brain Science*, volume 2021. Dana Foundation.

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

Barrett, L. F. and Russell, J. A. (2014). *The psychological construction of emotion.* Guilford Publications.

Barrett, L. F. and Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Current opinion in neurobiology*, 23(3):361–372.

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. and Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews. Neuroscience*, 16(7):419.

Barrett, L. F., Wilson-Mendenhall, C. D., and Barsalou, L. W. (2015). The conceptual act theory: A roadmap. In Barrett, L. F. and Russell, J. A., editors, *The psychological construction of emotion*, pages 83–110. The Guilford Press.

Barros, P. and Wermter, S. (2016). Developing crossmodal expression recognition based on a deep neural model. *Adaptive behavior*, 24(5):373–396.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & cognition*, 11(3):211–227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of experimental psychology: learning, memory, and cognition*, 11(4):629.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, 22(4):577–660.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.

Barsalou, L. W. (2014). On the indistinguishability of exemplar memory and abstraction in category representation. *Advances in social cognition*, 3:61–88.

Baruah, M. and Banerjee, B. (2020). A multimodal predictive agent model for human interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1022–1023.

Beal, M. and Ghahramani, Z. (2003). The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, volume 7, pages 453–464. Oxford University Press.

Bechara, A. and Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and economic behavior*, 52(2):336–372.

Bechara, A., Damasio, H., and Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3):295–307.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155.

Berridge, K. C. and Robinson, T. E. (2003). Parsing reward. *Trends in neurosciences*, 26(9):507–513.

Billard, A. G., Calinon, S., and Dillmann, R. (2016). Learning from humans. In Siciliano, B. and Khatib, O., editors, *Handbook of Robotics*, chapter 74, pages 1995–2014. Springer, Secaucus, NJ, USA, 2nd edition edition.

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. *The Journal of Machine Learning Research*, 20(1):973–978.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Boccignone, G., Conte, D., Cuculo, V., D'Amelio, A., Grossi, G., and Lanzarotti, R. (2018a). Deep construction of an affective latent space via multimodal enactment. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):865–880.

Boccignone, G., Conte, D., Cuculo, V., D'Amelio, A., Grossi, G., and Lanzarotti, R. (2018b). 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. (2007). Bayesian models and simulations in cognitive science. In *Models and Simulations 2*, pages 1–16. Tilburg Center for Logic and Philosophy of Science.

Boccignone, G. and Cordeschi, R. (2015). Coping with levels of explanation in the behavioral sciences. *Frontiers in Psychology*, 6:213.

Bohn, M. and Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, 1:223–249.

Braunwald, E., Fauci, A., Isselbacher, K., et al. (1998). *Harrison's Principles of Internal Medicine*. McGraw-Hill Companies, USA.

Breazeal, C. (2003). Toward sociable robots. *Robotics and autonomous systems*, 42(3):167–175.

Breazeal, C., Edsinger, A., Fitzpatrick, P., and Scassellati, B. (2001). Active vision for sociable robots. *IEEE Transactions on systems, man, and cybernetics-part A: Systems and Humans*, 31(5):443–453.

Breazeal, C. and Scassellati, B. (1999). How to build robots that make friends and influence people. In *Proceedings IEEE/RSJ International Conference onIntelligent Robots and Systems, IROS'99*, volume 2, pages 858–863. IEEE.

Brito, B., Zhu, H., Pan, W., and Alonso-Mora, J. (2020). Social-vrnn: one-shot multi-modal trajectory prediction for interacting pedestrians. *arXiv preprint arXiv:2010.09056*.

Broekens, J., Degroot, D., and Kosters, W. A. (2008). Formal models of appraisal: Theory, specification, and computational model. *Cognitive Systems Research*, 9(3):173–197.

Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159.

Brooks, R. A., Breazeal, C., Marjanovic, M., Scassellati, B., and Williamson, M. M. (1999). The cog project: Building a humanoid robot. *Lecture Notes in Computer Science*, pages 52–87.

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.

Bunt, H. and Black, B. (2000). The abc of computational pragmatics. *Abduction, Belief, and Context in Dialogue: Studies in Computational Pragmatics*, 1:1.

Cagli, R. C., Coraggio, P., Napoletano, P., and Boccignone, G. (2008). What the draughtsman's hand tells the draughtsman's eye: A sensorimotor account of drawing. *International Journal of Pattern Recognition and Artificial Intelligence*, 22(05):1015–1029.

Cangelosi, A. and Stramandinoli, F. (2018). A review of abstract concept learning in embodied agents and robots. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752):20170131.

Cavallo, F., Semeraro, F., Fiorini, L., Magyar, G., Sinčák, P., and Dario, P. (2018). Emotion modelling for social robotics applications: a review. *Journal of Bionic Engineering*, 15(2):185–203.

Chanes, L. and Barrett, L. F. (2016). Redefining the role of limbic areas in cortical processing. *Trends in cognitive sciences*, 20(2):96–106.

Chater, N., Tenenbaum, J., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291.

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrá, P., and Sanborn, A. (2020). Probabilistic biases meet the bayesian brain. *Current Directions in Psychological Science*, 29(5):506–512.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT press, Cambridge, MA.

Christensen, T. C., Barrett, L. F., Bliss-Moreau, E., Lebo, K., and Kaschub, C. (2003). A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, 4(1):53–78.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. (2015). A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28:2980–2988.

Churamani, N., Kerzel, M., Strahl, E., Barros, P., and Wermter, S. (2017). Teaching emotion expressions to a human companion robot using deep neural architectures. In *International Joint Conference on Neural Networks (IJCNN), 2017*, pages 627–634. IEEE.

Ciria, A., Schillaci, G., Pezzulo, G., Hafner, V. V., and Lara, B. (2021). Predictive processing in cognitive robotics: a review. *Neural Computation*, 33(5):1402–1432.

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.

Clore, G. L. and Ortony, A. (2008). Appraisal theories: How cognition shapes affect into emotion. In Lewis, M., Haviland-Jones, J. M., and Barrett, L. F., editors, *The Handbook of emotions*, pages 628–642. Guilford Press, New York, NY.

Clore, G. L. and Ortony, A. (2013). Psychological construction in the occ model of emotion. *Emotion Review*, 5(4):335–343.

Coen-Cagli, R., Coraggio, P., Napoletano, P., Schwartz, O., Ferraro, M., and Boccignone, G. (2009). Visuomotor characterization of eye movements in a drawing task. *Vision research*, 49(8):810–818.

Cohn-Gordon, R., Goodman, N., and Potts, C. (2018). Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443.

Colombo, M. and Seriès, P. (2020). Bayes in the brain — on bayesian modelling in neuroscience. *The British journal for the philosophy of science*.

Cordeschi, R. (2002). *The Discovery of the Artificial: Behavior, Mind and Machines Before and Beyond Cybernetics*. Kluwer Academic Publishers.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. Wiley and Sons, New York, N.Y.

Crivelli, C. and Fridlund, A. J. (2019). Inside-out: From basic emotions theory to the behavioral ecology view. *Journal of Nonverbal Behavior*, 43(2):161–194.

Cuculo, V. (2018). *A probabilistic approach to the construction of a multimodal affect space*. PhD thesis, Dipartimento di Matematica – Università degli Studi di Milano "La Statale", Milano, Italy.

Cuculo, V. and D'Amelio, A. (2019). Openfacs: an open source facs-based 3d face animation system. In *International Conference on Image and Graphics*, pages 232–242. Springer.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt.

Daunizeau, J., Friston, K. J., and Kiebel, S. J. (2009). Variational bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D: Nonlinear Phenomena*, 238(21):2089–2118.

Davidoff, J., Davies, I., and Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature*, 398(6724):203–204.

De Ridder, D., Vanneste, S., and Freeman, W. (2014). The bayesian brain: phantom percepts resolve sensory uncertainty. *Neuroscience & Biobehavioral Reviews*, 44:4–15.

De Waal, F., Macedo, S. E., and Ober, J. E. (2006). *Primates and philosophers: How morality evolved.* Princeton University Press.

Demiris, Y., Aziz-Zadeh, L., and Bonaiuto, J. (2014). Information processing in the mirror neuron system in primates and machines. *Neuroinformatics*, 12(1):63–91.

Dennett, D. (1987). *The Intentional Stance*. MIT Press, Cambridge, MA.

Dias, J., Mascarenhas, S., and Paiva, A. (2014). Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Emotion modeling*, pages 44–56. Springer.

Dimitrievska, V. and Ackovska, N. (2020). Behavior models of emotion-featured robots: A survey. *Journal of Intelligent & Robotic Systems*, 100(3):1031–1053.

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

D'Mello, S., Kappas, A., and Gratch, J. (2018). The affective computing approach to affect measurement. *Emotion Review*, 10(2):174–183.

Eckman, P. (1972). Universal and cultural differences in facial expression of emotion. In *Nebraska symposium on motivation*, volume 19, pages 207–284. University of Nebraska Press.

Eco, U. (2000). *Kant and the platypus: Essays on language and cognition*. Vintage Books, London,UK.

Edelberg, R. (1993). Electrodermal mechanisms: A critique of the two-effector hypothesis and a proposed replacement. *Progress in electrodermal research*, pages 7–29.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48(4):384.

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

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Ekman, P. and Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Ekman, P. and Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press.

Ekman, P., Sorenson, E. R., and Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88.

Electrophysiology, T. F. o. t. E. S. o. C. t. N. A. S. (1996). Heart rate variability : Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065.

Ellsworth, P. C. and Scherer, K. R. (2003). Appraisal processes in emotion. *Handbook of affective sciences*, 572:V595.

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.

Erk, K. (2021). The probabilistic turn in semantics and pragmatics. *Annual Review of Linguistics*, 8.

Fedorenko, E., Blank, I. A., Siegelman, M., and Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203:104348.

Fedorenko, E. and Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in cognitive sciences*, 18(3):120–126.

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116(4):752.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Fridlund, A. and Russell, J. (2021). Evolution, emotion and facial behavior a 21st century view. In Al-Shawaf, L. and Shackelford, T., editors, *The Oxford Handbook of Evolution and the Emotions*. Oxford University Press, Oxford, UK.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392.

Frijda, N. H. (1986). *The emotions*. Cambridge University Press, New York, NY.

Friston, K. (2008). Hierarchical models in the brain. *PLoS computational biology*, 4(11):e1000211.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301.

Friston, K. and Stephan, K. (2007). Free-energy and the brain. *Synthese*, 159:417–458. Published online.

Frith, C. and Frith, U. (2005). Theory of mind. *Current biology*, 15(17):R644–R645.

Gallese, V. (2007). Embodied simulation: from mirror neuron systems to interpersonal relations. In *Novartis Found Symp*, volume 278, pages 3–12.

García-Cabezas, M. Á., Zikopoulos, B., and Barbas, H. (2019). The structural model: a theory linking connections, plasticity, pathology, development and evolution of the cerebral cortex. *Brain Structure and Function*, 224(3):985–1008.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459.

Giere, R. (1999). Using models to represent reality. In Magnani, L., Nersessian, N., and Thagard, P., editors, *Model-Based Reasoning in Scientific Discovery*, pages 41 – 57, New York. Kluwer/Plenum.

Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., and Alameda-Pineda, X. (2021). Dynamical variational autoencoders: A comprehensive review. *Foundations and Trends in Machine Learning*, 15(1-2):1–175.

Gluz, J. and Jaques, P. A. (2017). A probabilistic formalization of the appraisal for the occ event-based emotions. *Journal of Artificial Intelligence Research*, 58:627–664.

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.

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

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.

Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer.

Hagoort, P. (2014). Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current opinion in Neurobiology*, 28:136–141.

Hall, J. E. (2010). *Guyton and Hall Textbook of Medical Physiology, 12th Edition*. Saunders, 12 edition.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Harnad, S. (2003). Categorical perception. In *Encyclopedia of Cognitive Science*, volume 67. MacMillan: Nature Publishing Group.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Herbet, G. and Duffau, H. (2020). Revisiting the functional anatomy of the human brain: toward a meta-networking theory of cerebral functions. *Physiological reviews*, 100(3):1181–1228.

Hertrich, I., Dietrich, S., and Ackermann, H. (2020). The margins of the language network in the brain. *Frontiers in Communication*, 5:93.

Hinton, G. E. et al. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA.

Hoemann, K., Khan, Z., Feldman, M. J., Nielson, C., Devlin, M., Dy, J., Barrett, L. F., Wormwood, J. B., and Quigley, K. S. (2020a). Context-aware experience sampling reveals the scale of variation in affective experience. *Scientific reports*, 10(1):1–16.

Hoemann, K., Wu, R., LoBue, V., Oakes, L. M., Xu, F., and Barrett, L. F. (2020b). Developing an understanding of emotion categories: Lessons from objects. *Trends in cognitive sciences*, 24(1):39–51.

Hoemann, K., Xu, F., and Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental psychology*, 55(9):1830.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57.

Horii, T., Nagai, Y., and Asada, M. (2016). Imitation of human expressions based on emotion estimation by mental simulation. *Paladyn, Journal of Behavioral Robotics*, 7(1).

Hortensius, R., Hekele, F., and Cross, E. S. (2018). The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):852–864.

Hosseini, M. and Maida, A. (2020). Hierarchical predictive coding models in a deep-learning framework. *arXiv preprint arXiv:2005.03230*.

Hutchinson, J. B. and Barrett, L. F. (2019). The power of predictions: An emerging paradigm for psychological research. *Current directions in psychological science*, 28(3):280–291.

Izard, C. E. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological review*, 100(1):68.

Jackendoff, R. and Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

James, W. (1884). What is an emotion? *Mind*, pages 188–205.

James, W. (1890). *The principles of psychology*. Henry Holt.

Jannach, D., Manzoor, A., Cai, W., and Chen, L. (2021). A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

Jordan, M. I. (1998). *Learning in graphical models*, volume 89. Springer Science & Business Media.

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.

Kao, J. T. and Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In *CogSci*.

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.

Katz, J. J. and Fodor, J. A. (1963). The structure of a semantic theory. *language*, 39(2):170–210.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current opinion in neurobiology*, 9(6):718–727.

Keltner, D., Sauter, D., Tracy, J., and Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior*, pages 1–28.

Khan, I. and Cañamero, L. (2018). Modelling adaptation through social allostasis: Modulating the effects of social touch with oxytocin in embodied agents. *Multimodal Technologies and Interaction*, 2(4):67.

Kidger, P., Foster, J., Li, X., and Lyons, T. J. (2021). Neural sdes as infinite-dimensional gans. In *International Conference on Machine Learning*, pages 5453–5463. PMLR.

Kilner, J. M., Friston, K. J., and Frith, C. D. (2007). The mirror-neuron system: a bayesian perspective. *Neuroreport*, 18(6):619–623.

Kim, Y., Lee, H., and Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *Proc.IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3687–3691. IEEE.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kleckner, I. R., Zhang, J., Touroutoglou, A., Chanes, L., Xia, C., Simmons, W. K., Quigley, K. S., Dickerson, B. C., and Barrett, L. F. (2017). Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature human behaviour*, 1(5):1–14.

Knill, D., Kersten, D., and Yuille, A. (1996). *Introduction: A Bayesian formulation of visual perception*, pages 1–21. Cambridge University Press.

Koban, L., Gianaros, P. J., Kober, H., and Wager, T. D. (2021). The self in context: brain systems linking mental and physical health. *Nature Reviews Neuroscience*, 22(5):309–322.

Koch, C. (1999). *Biophysics of Computation - Information Processing in Single Neurons*. Oxford University Press, New York.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press, Cambridge, MA.

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological psychology*, 84(3):394–421.

Kronrod, Y., Coppess, E., and Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic bulletin & review*, 23(6):1681–1712.

Kuppens, P., Oravecz, Z., and Tuerlinckx, F. (2010). Feelings change: accounting for individual differences in the temporal dynamics of affect. *Journal of personality and social psychology*, 99(6):1042.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Lakoff, G. (1971). *On generative semantics*, volume 232, page 296. Cambridge University Press, Cambridge, MA.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lang, P. J., Greenwald, M. K., Bradley, M. M., and Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273.

Lassiter, D. and Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 194(10):3801–3836.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press on Demand.

LeDoux, J. E. and Hofmann, S. G. (2018). The subjective experience of emotion: a fearful view. *Current Opinion in Behavioral Sciences*, 19:67–72.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Levenson, R. W. (1988). Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. In *Social Psychophysiology and Emotion: Theory and Clinical Applications*. John Wiley & Sons.

Leys, R. (2017). *The ascent of affect.* University of Chicago Press, Chicago, IL.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Lim, A. and Okuno, H. G. (2014). The mei robot: towards using motherese to develop multimodal emotional intelligence. *IEEE Transactions on Autonomous Mental Development*, 6(2):126–138.

Lindquist, K. A. (2013). Emotions emerge from more basic psychological ingredients: A modern psychological constructionist model. *Emotion Review*, 5(4):356–368.

Lindquist, K. A., MacCormack, J. K., and Shablack, H. (2015a). The role of language in emotion: Predictions from psychological constructionism. *Frontiers in psychology*, 6:444.

Lindquist, K. A., Satpute, A. B., and Gendron, M. (2015b). Does language do more than communicate emotion? *Current directions in psychological science*, 24(2):99–108.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., and Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *The Behavioral and brain sciences*, 35(3):121.

Ma, H. and Yarosh, S. (2021). A review of affective computing research based on function-component-representation framework. *IEEE Transactions on Affective Computing*, pages 1–1.

MacKay, D. (2004). *Information Theory, inference and Learning Algorithms.* Cambridge University Press, Cambridge, UK.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., and Chen, S. (2021). Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior research methods*, 53(4):1689–1696.

Mancuso, L., Cavuoti-Cabanillas, S., Liloia, D., Manuello, J., Buzi, G., Duca, S., Cauda, F., and Costa, T. (2021). Default mode network spatial configuration varies across task domains. *bioRxiv*.

Mandler, G. et al. (1975). *Mind and emotion.* John Wiley & Sons.

Marino, J. (2020). Predictive coding, variational autoencoders, and biological connections. *arXiv preprint arXiv:2011.07464*.

Marino, J. (2021). Predictive coding, variational autoencoders, and biological connections. *Neural Computation*, 34(1):1–44.

Marino, J., Yue, Y., and Mandt, S. (2018). Iterative amortized inference. In *International Conference on Machine Learning*, pages 3403–3412. PMLR.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H. Freeman, New York.

Marsella, S. C. and Gratch, J. (2009). Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90.

Massaroli, S., Poli, M., Park, J., Yamashita, A., and Asama, H. (2020). Dissecting neural odes. *Advances in Neural Information Processing Systems*, 33:3952–3963.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, 85(3):207.

Menon, V. (2015). Salience network. In Toga, A. W., editor, *Brain mapping: an encyclopedic reference*, volume 2, pages 597–611. Academic Press, Elsevier, Cambridge, MA.

Menon, V. and D'Esposito, M. (2022). The role of pfc networks in cognitive control and executive function. *Neuropsychopharmacology*, 47(1):90–103.

Mesquita, B., Boiger, M., and De Leersnyder, J. (2016). The cultural construction of emotions. *Current opinion in psychology*, 8:31–36.

Metta, G., Sandini, G., Natale, L., Craighero, L., and Fadiga, L. (2006). Understanding mirror neurons: a bio-robotic approach. *Interaction studies*, 7(2):197–232.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Moerland, T. M., Broekens, J., and Jonker, C. M. (2018). Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, 107(2):443–480.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Montague, R. (2019). *English as a formal language*, pages 94–121. De Gruyter Mouton.

Moors, A. (2013). On the causal role of appraisal in emotion. *Emotion Review*, 5(2):132–140.

Moors, A. (2020). Appraisal theory of emotion. *Encyclopedia of personality and individual differences*, pages 232–240.

Murphy, G. (2004). *The big book of concepts*. MIT press, Cambridge, MA.

Nakasone, A., Prendinger, H., and Ishizuka, M. (2005). Emotion Recognition from Electromyography and Skin Conductance. *The 5th International Workshop on Biosignal Interpretation*, pages 219–222.

Naseem, U., Razzak, I., Khan, S. K., and Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35.

Natale, L., Nori, F., Metta, G., Fumagalli, M., Ivaldi, S., Pattacini, U., Randazzo, M., Schmitz, A., and Sandini, G. (2013). The icub platform: a tool for studying intrinsically motivated learning. In *Intrinsically motivated learning in natural and artificial systems*, pages 433–458. Springer.

Neale, S. (1992). Paul grice and the philosophy of language. *Linguistics and philosophy*, pages 509–559.

Newell, A. (1980). Physical symbol systems. *Cognitive science*, 4(2):135–183.

Nussbaum, M. C. (2003). *Upheavals of thought: The intelligence of emotions*. Cambridge University Press, Cambridge, UK.

Ojha, S., Vitale, J., and Williams, M.-A. (2021). Computational emotion models: a thematic review. *International Journal of Social Robotics*, 13(6):1253–1279.

Ong, D. C., Wu, Z., Tan, Z.-X., Reddan, M., Kahhale, I., Mattek, A., and Zaki, J. (2019a). Modeling emotion in complex stories: the stanford emotional narratives dataset. *IEEE Transactions on Affective Computing*, 12(3):579–594.

Ong, D. C., Zaki, J., and Goodman, N. D. (2019b). Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, 11(2):338–357.

Oravecz, Z., Tuerlinckx, F., and Vandekerckhove, J. (2011). A hierarchical latent stochastic differential equation model for affective dynamics. *Psychological methods*, 16(4):468.

Ortony, A. and Clore, G. (2015). Can an appraisal model be compatible with psychological constructionism? In Barrett, L. F. and Russell, J. A., editors, *The psychological construction of emotion*, pages 305–333. The Guilford Press, New York, NY.

Ortony, A., Clore, G., and Collins, A. (1988). The cognitive structure of emotions. *CBO9780511571299*.

Ortony, A., Clore, G. L., and Collins, A. (1990). *The cognitive structure of emotions*. Cambridge university press.

Oztop, E., Kawato, M., and Arbib, M. (2006). Mirror neurons and imitation: A computationally guided review. *Neural Networks*, 19(3):254–271.

Oztop, E., Kawato, M., and Arbib, M. A. (2013). Mirror neurons: functions, mechanisms and models. *Neuroscience letters*, 540:43–55.

Palminteri, S., Wyart, V., and Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6):425–433.

Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, New York, NY.

Paquola, C., Garber, M., Frässle, S., Royer, J., Tavakol, S., Cruces, R., Jefferies, E., Smallwood, J., and Bernhardt, B. (2021). The unique cytoarchitecture and wiring of the human default mode network. *bioRxiv*.

Peirce, C. S. (1991). *Peirce on signs: Writings on semiotic*. UNC Press Books.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):510.

Picard, R. W. (1997). *Affective computing*, volume 252. MIT press Cambridge.

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.

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Pylyshyn, Z. (1984). *Computation and Cognition*. MIT Press, Cambridge, MA.

Quigley, K. S., Lindquist, K. A., and Barrett, L. F. (2013). Inducing and measuring emotion and affect: Tips, tricks, and secrets. In Reis, H. T. and Judd, C. M., editors, *Handbook of research methods in social and personality psychology*, pages 220–252. Cambridge University Press, New York, NY, USA.

Rai, S. and Chakraverty, S. (2020). A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682.

Rao, R. P. and Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural computation*, 9(4):721–763.

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.

Reisenzein, R., Hudlicka, E., Dastani, M., Gratch, J., Hindriks, K., Lorini, E., and Meyer, J.-J. C. (2013). Computational modeling of emotion: Toward improving the inter-and intradisciplinary exchange. *IEEE Transactions on Affective Computing*, 4(3):246–266.

Repplinger, M., Beinborn, L., and Zuidema, W. (2018). Vector-space models of words and sentences. *Nieuw Archief voor Wiskunde*, 19(3):167–174.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR.

Richardson, S. (2020). Affective computing in the modern workplace. *Business Information Review*, 37(2):78–85.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Proc. 10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition*, pages 1–8. IEEE.

Rizzolatti, G. and Sinigaglia, C. (2016). The mirror mechanism: a basic principle of brain function. *Nature Reviews Neuroscience*, 17(12):757–765.

Roger, E., Banjac, S., Thiebaut de Schotten, M., and Baciu, M. (2022). Missing links: The functional unification of language and memory ($l \cup m$). *Neuroscience & Biobehavioral Reviews*, 133:104489.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. *Review of personality & social psychology*.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Russell, S. J. and Norvig, P. (2022). *Artificial intelligence: a modern approach*. Pearson Education Limited, Harlow, UK, 4 edition.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55.

Sanborn, A. N. and Chater, N. (2016). Bayesian brains without probabilities. *Trends in cognitive sciences*, 20(12):883–893.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, 117(4):1144.

Sánchez-López, Y. and Cerezo, E. (2019). Designing emotional bdi agents: good practices and open questions. *The Knowledge Engineering Review*, 34.

Santini, S. and Dumitrescu, A. (2008). Context as a non-ontological determinant of semantics. In *International Conference on Semantic and Digital Media Technologies*, pages 121–136. Springer.

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.

Scarantino, A. (2017). How to do things with emotional expressions: The theory of affective pragmatics. *Psychological Inquiry*, 28(2-3):165–185.

Scassellati, B. (2002). Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24.

Schachter, S. and Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5):379.

Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion*, 7(3-4):325–355.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, 92(120):57.

Scherer, K. R. (2009). Emotion theories and concepts (psychological perspectives). In Sander, D. and Scherer, s. K. R., editors, *Oxford companion to emotion and the affective sciences*, pages 145–149. Oxford University Press, Oxford, UK.

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

Schuller, B. W., Picard, R., André, E., Gratch, J., and Tao, J. (2021). Intelligent signal processing for affective computing [from the guest editors]. *IEEE Signal Processing Magazine*, 38(6):9–11.

Schuller, D. and Schuller, B. W. (2018). The age of artificial emotional intelligence. *Computer*, 51(9):38–46.

Searle, J. R., Willis, S., et al. (1995). *The construction of social reality*. Simon and Schuster.

Seeley, W. W. (2019). The salience network: a neural system for perceiving and responding to homeostatic demands. *Journal of Neuroscience*, 39(50):9878–9882.

Selvaraj, N., Jaryal, a., Santhosh, J., Deepak, K. K., and Anand, S. (2008). Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography. *Journal of medical engineering & technology*, 32(6):479–484.

Senft, G. (2014). *Understanding pragmatics*. Routledge, New York, NY.

Sennesh, E., Theriault, J., Brooks, D., van de Meent, J.-W., Barrett, L. F., and Quigley, K. S. (2022). Interoception as modeling, allostasis as control. *Biological Psychology*, 167:108242.

Seuren, P. A. (2009). *Language from within: Vol. 1. Language in cognition*. Oxford University Press.

Sheridan, T. B. (2020). A review of recent research in social robotics. *Current opinion in psychology*, 36:7–12.

Sikström, S. and Garcia, D. (2020). *Statistical semantics: Methods and applications*. Springer Nature, Cham, Switzerland.

Smallwood, J., Bernhardt, B. C., Leech, R., Bzdok, D., Jefferies, E., and Margulies, D. S. (2021). The default mode network in cognition: a topographical perspective. *Nature reviews neuroscience*, 22(8):503–513.

Smith, C. A. and Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813.

Smith, R., Thayer, J. F., Khalsa, S. S., and Lane, R. D. (2017). The hierarchical basis of neurovisceral integration. *Neuroscience & biobehavioral reviews*, 75:274–296.

Solomon, R. C. (2007). *Not passion's slave: emotions and choice*. Oxford University Press.

Solomon, R. C. (2008). The philosophy of emotions. In Lewis, M., Haviland-Jones, J. M., and Barrett, L. F., editors, *The Handbook of emotions*, pages 3–15. Guilford Press, New York, NY.

Speaks, J. (2021). Theories of Meaning. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition.

Sperber, D. and Wilson, D. (1986). *Relevance: Communication and cognition*. Harvard University Press, Cambridge, MA.

Sporns, O., Kötter, R., and Friston, K. J. (2004). Motifs in brain networks. *PLoS biology*, 2(11):e369.

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

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.

Stuart, A. M. (2010). Inverse problems: a bayesian perspective. *Acta Numerica*, 19:451–559.

Stuhlmüller, A. and Goodman, N. D. (2014). Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*, 28:80–99.

Susanto, Y., Cambria, E., Ng, B. C., and Hussain, A. (2021). Ten years of sentic computing. *Cognitive Computation*, pages 1–19.

Tani, J., Ito, M., and Sugita, Y. (2004). Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using rnnpb. *Neural Networks*, 17(8):1273–1289.

Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T., and Asoh, H. (2016). Symbol emergence in robotics: a survey. *Advanced Robotics*, 30(11-12):706–728.

Tatham, M. and Morton, K. (2006). *Speech production and perception*. Springer.

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.

Tomasello, M. (2010). *Origins of human communication*. MIT press, Boston, MA.

Tomkins, S. S. (1962). *Affect, imagery, consciousness*. Springer, New York.

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.

Torre, J. B. and Lieberman, M. D. (2018). Putting feelings into words: Affect labeling as implicit emotion regulation. *Emotion Review*, 10(2):116–124.

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*.

Tsuji, S., Cristia, A., and Dupoux, E. (2021). Scala: A blueprint for computational models of language acquisition in social context. *Cognition*, page 104779.

Tzen, B. and Raginsky, M. (2019a). Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*.

Tzen, B. and Raginsky, M. (2019b). Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, pages 3084–3114. PMLR.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of the brownian motion. *Physical review*, 36(5):823.

van de Meent, J.-W., Paige, B., Yang, H., and Wood, F. (2018). An introduction to probabilistic programming. *arXiv preprint arXiv:1809.10756*.

Van Kampen, N. (2011). *Stochastic Processes in Physics and Chemistry*. Elsevier.

van Tonder, G. J. and Ejima, Y. (2000). Bottom–up clues in target finding: Why a dalmatian may be mistaken for an elephant. *Perception*, 29(2):149–157.

Vilares, I. and Kording, K. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Annals of the New York Academy of Sciences*, 1224(1):22.

Wang, B. and Titterington, D. (2004). Variational bayesian inference for partially observed diffusions. Technical report, Technical Report 04-4, University of Glasgow. http://www. stats. gla. ac. uk/Research/-TechRep2003/04-4. pdf.

Wang, Z., Ho, S.-B., and Cambria, E. (2020). A review of emotion sensing: categorization models and algorithms. *Multimedia Tools and Applications*, 79(47):35553–35582.

Whorf, B. (1964). *Language, thought, and reality: Selected writings*. MIT Press, Cambridge, MA.

Wiese, E., Metta, G., and Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, 8:1663.

Wiriyathammabhum, P., Summers-Stay, D., Fermüller, C., and Aloimonos, Y. (2016). Computer vision and natural language processing: recent approaches in multimedia and robotics. *ACM Computing Surveys (CSUR)*, 49(4):1–44.

Wittgenstein, L. (2009). *Philosophical Investigations*. John Wiley & Sons.

Wood, A., Rychlowska, M., Korb, S., and Niedenthal, P. (2016). Fashioning the face: sensorimotor simulation contributes to facial expression recognition. *Trends in cognitive sciences*, 20(3):227–240.

Wood, L. A. and Harré, R. (1986). *The Social Construction of Emotions*. Blackwell.

Xu, Y., He, Y., and Bi, Y. (2017). A tri-network model of human semantic processing. *Frontiers in Psychology*, 8:1538.

Yan, F., Iliyasu, A. M., and Hirota, K. (2021). Emotion space modelling for social robots. *Engineering Applications of Artificial Intelligence*, 100:104178.

Yang, Y., Saleemi, I., and Shah, M. (2013). Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1635–1648.

Yeshurun, Y., Nguyen, M., and Hasson, U. (2021). The default mode network: where the idiosyncratic self meets the shared social world. *Nature Reviews Neuroscience*, 22(3):181–192.

Yon, D. and Frith, C. D. (2021). Precision and the bayesian brain. *Current Biology*, 31(17):R1026–R1032.

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.

Yu, S., Yu, K., Tresp, V., Kriegel, H.-P., and Wu, M. (2006). Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 464–473.

Yule, G. (2020). *The study of language*. Cambridge university press, London,UK.

Zeng, Y., Zhao, Y., Zhang, T., Zhao, D., Zhao, F., and Lu, E. (2020). A brain-inspired model of theory of mind. *Frontiers in Neurorobotics*, 14:60.

Zhang, J., Abiose, O., Katsumi, Y., Touroutoglou, A., Dickerson, B. C., and Barrett, L. F. (2019). Intrinsic functional connectivity is organized as three interdependent gradients. *Scientific reports*, 9(1):1–14.

Zhao, S., Yao, X., Yang, J., Jia, G., Ding, G., Chua, T.-S., Schuller, B. W., and Keutzer, K. (2021). Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.