THE UNIVERSITY
of EDINBURGH

Emotion and Predictive Processing:

Emotions as perceptions?

J. M. Araya

PhD Philosophy

The University of Edinburgh

2017

Emotion and Predictive Processing:

Emotions as perceptions?

Abstract

In this Thesis, I systematize, clarify, and expand the current theory of emotion based on the principles of predictive processing—the *interoceptive inference view of emotion*—so as to show the following: (1) as it stands, this view is problematic. (2) Once expanded, the view in question can deal with its more pressing problems, and it compares favourably to competing accounts. Thus, the interoceptive inference view of emotion stands out as a plausible theory of emotion.

According to the *predictive processing* (PP) framework, all what the brain does, in all its functions, is to minimize its precision-weighted *prediction error* (PE) (Clark, 2013, 2016; Hohwy, 2013). Roughly, PE consist in the difference between the sensory signals expected (and generated) from the top-down and the actual, incoming sensory signals. Now, in the PP framework, visual percepts are formed by minimizing visual PE in a specific manner: via visual *perceptual inference*. That is, the brain forms visual percepts in a top-down fashion by predicting its incoming lower-level sensory signals from higher-level models of the likely (hidden) causes of those visual signals. Such models can be seen as putting forward content-specifying hypotheses about the object or event responsible for triggering incoming sensory activity. A contentful percept is formed once a certain hypothesis achieves to successfully match, and thus supress, current lower-level sensory signals.

In the *interoceptive inference* approach to interoception (Seth, 2013, 2015), the principles of PP have been extended to account for interoception, i.e., the perception of our homeostatic, physiological condition. Just as perception in the visual domain arises via visual *perceptual inference*, the interoceptive inference approach holds that perception of the inner, physiological milieu arises via *interoceptive perceptual inference*.

Now, what might be called the *interoceptive inference theory of valence* (ITV) holds that the interoceptive inference approach can be used so as to account for subjective feeling states in general, i.e., mental states that feel good or bad—i.e., *valenced* mental states. According to ITV, affective valence arises by way of interoceptive perceptual inference.

On the other hand, what might be called the *interoceptive inference view of emotion* (IIE) holds that the interoceptive inference approach can be used so as to account for *emotions per se* (e.g., fear, anger, joy). More precisely, IIE holds that, in direct analogy to the way in which visual percepts are formed, emotions arise from interoceptive predictions of the causes of current interoceptive afferents. In other words, emotions *per se* amount to interceptive percepts formed via higher-level, content-specifying *emotion* hypotheses.

In this Thesis, I aim to systematize, clarify, and expand the interoceptive inference approach to interoception, in order to show that: (1) contrary to non-sensory theories of affective valence, valence is indeed constituted by interoceptive perceptions, and that interoceptive percepts do arise via interoceptive perceptual inference. Therefore, ITV holds. (2) Considering that IIE exhibits problematic assumptions, it should be amended. In this respect, I will argue that emotions do *not* arise via interoceptive *perceptual inference* (as IIE claims), since this assumes that there must be regularities pertaining to emotion in the physiological domain. I will suggest that emotions arise instead by minimizing interoceptive PE in another fashion. That is, emotions arise via *external interoceptive active inference*: by sampling and modifying the external environment in order to change an already formed interoceptive percept (which has been formed via interoceptive perceptual inference). That is, emotions are specific strategies for regulating affective valence. More precisely, I will defend the view that a certain emotion *E* amounts to a specific strategy for minimizing interoceptive PE by way of a specific set of stored knowledge of the counterfactual relations that obtain between (possible) actions and its prospective interoceptive, sensory consequences ("if I act in this manner, interoceptive signals should evolve in such-and-such way"). An emotion arises when such knowledge is applied in order to regulate valence.

Emotion and Predictive Processing:

Emotions as perceptions?

Lay Summary

In this Thesis, I systematize, clarify, and expand the current theory of emotion based on the principles of predictive processing—the *interoceptive inference view of emotion*—so as to show the following: (1) as it stands, this view is problematic. (2) Once expanded, the view in question can deal with its more pressing problems, and it compares favourably to competing accounts. Thus, the interoceptive inference view of emotion stands out as a plausible theory of emotion.

According to the *predictive processing* (PP) framework, the brain, in all its functions, is attempting to anticipate the information that the senses deliver, and thus reduce as much as possible the difference between the information that it anticipates and the information that it actually receives from the senses. This can be achieved in two ways: via perception or via action. In the PP framework, perception is then understood as emerging from the anticipation of the actual information that a certain sense delivers; while action emerges by changing the world so as to receive the information that the brain is already expecting.

In the *interoceptive inference view of emotion* (IIE), and in the *interoceptive inference theory of valence* (ITV), this basic idea is extended to account for emotion *per se* and valence, respectively. Valence amounts to the positive and negative character that affective states have. For example, joy has a positive character, while fear has a negative character. Emotions *per se* consist in the affective states that we commonly consider emotions to be. For example, fear, anger, joy, and guilt count as emotions in this sense.

ITV holds that valence emerges in a similar fashion as any sensory state emerges. That is, valence is something sensory. However, *non-sensory signal theories of valence* (NSS) have compellingly argued that valence cannot be understood as something

sensory. In this Thesis, I argue that, contrary to NSS, valence can indeed be understood as a sensory phenomenon. Therefore, ITV holds.

On the other hand, IIE holds that emotions *per se* arise in direct analogy to the way in which, for example, vision arises. That is, emotions are perceptions. More precisely, emotions are perceptions of our physiological condition (our hear rate increasing, changes in blood circulation, etc.). IIE claims then that an emotion emerges as the brain anticipates the information that the interoceptive senses actually deliver (i.e., the senses that register changes in our physiological condition). In this Thesis, I argue that this view is problematic, simply because there are no emotions configured in our physiology. Thus, the brain cannot learn what interoceptive information to expect in respect to a certain emotion instead of another emotion. This is analogous to the claim that the experience of clouds do not arise from auditive perceptions, simply because different types of clouds (cirrus, cumulus, etc.) are not configured in the 'auditive landscape'. Thus, the brain cannot learn what auditive information to expect in respect to a certain type of cloud (e.g., cirrus) instead of another type of cloud (e.g., cumulus).

It seems then that the PP framework, even though it can account for valence, is without an account of emotions *per se*. In this Thesis, I suggest that the PP framework can indeed account for emotions *per se*. Remember that the brain has two ways achieving its only goal of reducing as much as possible the difference between the information that it anticipates and the information that it receives from the senses: perception and action. Now, common sense tell us that sensing (or perceiving) our bodily, physiological changes is part of our experience of emotion. Then, reducing as much as possible the difference between the interoceptive information that brain anticipates and the information that it receives from the interoceptive senses must be part of the story of how emotions emerge. We saw that it is unlikely that emotions emerge via perception, i.e., by simply perceiving our bodies. Then, if perception is not the way to go in this respect, we are left with action. I suggest that emotions arise via action. Remember that, in the PP framework, action emerges by changing the world so as to receive the information that the brain is already expecting. I suggest then that emotions arise by changing our physiology to its expected state via emotion specific actions in

the external world. For example, fear, as others have suggested, consist in actions relative to the avoidance of a dangerous stimulus. If this view is on track, the PP framework can also account for emotion, the core of our mental life.

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Signed: 

J.M. Araya

Date: 25/08/2017

Table of contents

Acknowledgements

I would like to thank my supervisors, Tillmann Vierkant and Mark Sprevak. As my primary supervisor, T. Vierkant was particularly helpful. Till skilfully kept my non-targeted curiosity on track, without limiting my intellectual autonomy. I would also like to thank my sponsor CONICYT (Becas-Chile).

Introduction

The claim

In this Thesis, I systematize, clarify, and expand the current theory of emotion based on the principles of predictive processing—*the interoceptive inference view of emotion*—so as to show the following: (1) as it stands, this view is problematic. (2) Once expanded, the view in question can deal with its more pressing problems, and it compares favourably to competing accounts. Thus, the interoceptive inference view of emotion stands out as a plausible theory of emotion.

According to the *predictive processing* (PP) framework, all what the brain does, in all its functions, is to minimize its precision-weighted *prediction error* (PE) (Clark, 2013, 2016; Hohwy, 2013). Roughly, PE consist in the difference between the sensory signals expected (and generated) from the top-down and the actual, incoming sensory signals. Now, in the PP framework, visual percepts are formed by minimizing visual PE in a specific manner: via visual *perceptual inference*. That is, the brain forms visual percepts in a top-down fashion by predicting its incoming lower-level sensory signals from higher-level models of the likely (hidden) causes of those visual signals. Such models can be seen as putting forward content-specifying hypotheses about the object or event responsible for triggering incoming sensory activity. A contentful percept is formed once a certain hypothesis achieves to successfully match, and thus supress, current lower-level sensory signals.

In the *interoceptive inference* approach to interoception (Seth, 2013, 2015a), the principles of PP have been extended to account for interoception, i.e., the perception of our homeostatic, physiological condition. Just as perception in the visual domain arises via visual *perceptual inference*, the interoceptive inference approach holds that

perception of the inner, physiological milieu arises via *interoceptive perceptual inference*.

Now, what might be called the *interoceptive inference theory of valence* (ITV) holds that the interoceptive inference approach can be used so as to account for subjective feeling states in general, i.e., mental states that feel good or bad—i.e., *valenced* mental states. According to ITV, affective valence arises by way of *interoceptive perceptual inference*.

On the other hand, what might be called the *interoceptive inference view of emotion* (IIE) holds that the interoceptive inference approach can be used so as to account for *emotions per se* (e.g., fear, anger, joy). More precisely, IIE holds that, in direct analogy to the way in which visual percepts are formed, emotions arise from interoceptive predictions of the causes of current interoceptive afferents. In other words, emotions *per se* amount to interceptive percepts formed via higher-level, content-specifying *emotion* hypotheses.

In this Thesis, I aim to systematize, clarify, and expand the interoceptive inference approach to interoception, in order to show that: (1) contrary to non-sensory theories of affective valence, valence is indeed constituted by interoceptive perceptions, and that interoceptive percepts do arise via interoceptive perceptual inference. Therefore, ITV holds. (2) Considering that IIE exhibits problematic assumptions, it should be amended. In this respect, I will suggest that emotions do *not* arise via interoceptive *perceptual inference* (as IIE claims), but rather they arise by minimizing interoceptive PE in another fashion. That is, emotions arise via *external interoceptive active inference*: by sampling and modifying the external environment in order to change an already formed interoceptive percept (which has been formed via interoceptive perceptual inference). That is, emotions are specific strategies for regulating affective valence. More precisely, I will suggest the view that a certain emotion *E* amounts to a specific strategy for minimizing interoceptive PE by way of a specific set of stored

knowledge of the counterfactual relations that obtain between (possible) actions and its prospective interoceptive, sensory consequences ("if I act in this manner, interoceptive signals should evolve in such-and-such way"). An emotion arises when such knowledge is applied in order to regulate valence. Let me unpack these claims.

Predictive processing: its importance and ambitions

Given its simplicity and enormous unifying and explanatory power, PP is becoming an increasingly attractive way of carrying out theoretical and experimental research in cognitive science. Now, PP is not just another compelling theoretical approach to some cognitive function. PP has the ambitions to constitute itself as an overarching paradigm shift in our understanding of the functionings of the mind/brain. The ambition is high. The principles of PP promise to give us a unifying account of *all* the seemingly disparate variety of mental phenomena, ranging from perception to action (Clark, 2013; Hohwy, 2013).

The very basics of key PP notions

Roughly, in the PP framework, the traditional approach to perceptual processing is turned upside down. Traditionally, perception is considered to be a bottom-up driven process, in which sensory signals or features are accumulatively processed from the bottom-up until a coherent percept is finally formed. Contrary to the bottom-up approach, as I mentioned above, Bayesian PP accounts of perception hold that the mind/brain forms percepts in a top-down fashion, by predicting its incoming lower-level sensory signals from higher-level models/hypotheses of the likely (hidden) external causes of those signals (see Clark, 2013; Hohwy, 2013).

Crucially, as top-down signals produced by generative models provide predictions of sensory effects, bottom-up signals provide PE, as defined above. PE is critical in the PP framework, not only because, from a more global perspective, PP holds that all what the brain does is to minimize its PE. The latter is also crucial in the PP framework since during online processing, PE is used as feedback for updating and improving models or hypotheses and, in the long term, as a learning signal that improves models' predictions.

Importantly, in order to minimize PE, the system needs to determine the reliability of PE across all levels of the perceptual hierarchy. That is, the *precision* of incoming signals also needs to be inferred. Precisions determine, in a context-sensitive manner, which aspects of the signal are ignored (as they are deemed to be unreliable, or too variable), and which aspects of the signal are amplified (as they are deemed to be reliable). Thus, in contexts where the input is deemed to be unreliable, top-down hypotheses will be assigned more weight than usual. For example, think of a case in which the auditive stimulus amounts to white noise. In this sort of context, audition will then be driven almost completely by stored expectations, at the expense of what the world has to say via auditive PE: you begin hearing that song you heard this morning, and perhaps a conversation. In this kind of case, the incoming auditive PE is estimated to have too little precision, and thus top-down expectations are assigned significantly more weight. Hallucinations thus take place. In other contexts, the input will be deemed reliable. Think of the situation in which someone tries to read a very small footnote in good lightning conditions. In this case, those aspect of the input which encode information relative to the letters will be given more weight. In these latter type of cases, PE will be taken to be informative, and it will then be used to guide the current hypothesis-selection process and, in the long term, learning.

Now, as Hohwy (2013) remarks, as the process of model/hypothesis selection and revision in light of precision weighted PE unfolds, and learning thus takes place, generative models manage to extract the causal regularities from the world. In other words, priors are learned from experience, and over time they recapitulate the

regularities that configure the structure of the world. This is what allows the system to issue successful predictions of the worldly causes of incoming signals, and thus minimize PE.

Lower levels of the cortical hierarchy encode regularities that operate at fast time-scales, which capture variant aspects of experience. On the other hand, higher levels encode increasingly more complex regularities that operate at slow time-scales, which capture relatively more invariant aspects of experience.

Importantly for this Thesis, in the PP framework, *action* plays a key role during percept formation, and it operates under the same imperative toward PE minimization. Action takes here the form of *active inference*. Active inference consists in changing the environment so as to obtain sensory data that fits considered predictions or hypotheses. As Seth (2015a) remarks, active inference can take place not only so as to conform to current expectations, as it occurs during motor behaviour by making proprioceptive data fit proprioceptive expectations (by significantly increasing the *precision* of proprioceptive predictions), so that 'desired' (or expected) movement occurs. Active inference can also take place during percept formation (i.e., during *perceptual inference*), in order to confirm, disconfirm and disambiguate perceptual hypotheses. As Seth (2015a) notes, this requires storing representations of the counterfactual relations that obtain between (possible) actions and its prospective sensory consequences. This is what Seth calls 'knowledge of sensorimotor contingencies'.

The latter aspect of active inference will be key in this Thesis, as I will argue that emotions arise via 'external interoceptive active inference'. More precisely, as I already mentioned, I will defend the view that a certain emotion *E* amounts to a specific strategy for minimizing interoceptive PE by way of a specific set of stored representations of interoceptive sensorimotor contingencies—"if I act in this manner, interoceptive signals should evolve in such-and-such way". (However, even though they operate under the same principles of PE minimization, active inference and

perceptual inference differ in an obvious functional respect: they exhibit different direction of fit. While perceptual hypotheses have mind-to-world direction of fit—hypotheses attempt to match incoming signals—active inference has world-to-mind direction of fit—the system attempts so find signals that fit predicted signals.)

Rationale and importance of extending PP to emotions and affective valence

PP is already doing explanatory work in a wide variety of psychological domains. However, PP was conceived and developed as an account (and re-conceptualization) of *perceptual* processes. That is why its principles have been mainly applied in the explanation of mental phenomena that, in some way or another, can be readily understood as perceptual in nature – e.g., visual perception, binocular rivalry, illusions and delusions, etc. (for a review, see, e.g., Clark, 2013; Friston, 2005, 2009).

Now, according to the Jamesian view of emotion (James, 1884), emotions can be understood as *perceptions* of bodily, interoceptive changes. Considering that the Jamesian view that emotions can be understood as a perceptual process has recently seen a resurgence of interest in emotion research (e.g., Prinz, 2004), an obvious next step for PP's explanatory ambitions is to apply its principles in accounting for emotion.

For PP's explanatory ambitions, developing a successful PP account of emotion is not only interesting for the simple fact that emotions drive our lives: they are the mental phenomena that we personally care most about. But developing a successful PP account of emotion is also interesting for more theoretical reasons. In this respect, developing a successful PP account of emotion can play a pivotal role in accounting for other higher-level mental phenomena. For example, it is widely agreed that moral judgments are dependent on emotional processes (e.g., Greene, 2001; Haidt, 2001; Nichols, 2004; Pizarro, 2000; Prinz, 2007). If that is the case, a successful PP account of emotion could inform a major part (if not all) of a PP account of moral judgment.

The same can be said of other mental functions dependent on emotion, such as, for example, decision-making, social cognition, motivation, and self-control. It could also inform explanations of certain mental dysfunctions dependent on emotional processes, such as, for example, borderline personality disorder, psychopathy, and alexithymia-related disorders. More interestingly, mental life is affective to the core. Therefore, developing an account of emotion is mandatory for PP's ambitions of constituting itself as an overarching, unifying framework in cognitive science.

To date there is no fully developed PP account of emotion on offer. Nonetheless, A. Seth (Seth, 2013; Seth et al, 2012; Seth & Critchley, 2013) and J. Hohwy (2013, pp. 242-244) have recently offered a first sketch of how such extension might go. Taking into account the fact that PP mainly works as an account of perception, and that perceptual views of emotion have been recently compellingly defended, those first sketches have suggested a PP version of the perceptual, interoceptive view of emotion. According to this kind of view, emotions are *perceptions* of distinct bodily changes. Resting on the plausibility of some of the commitments of this kind of view, those first PP accounts of emotion see emotion as arising from predictive *interoceptive inferences*—more precisely, from interoceptive *perceptual inference*.

Seth's (and Hohwy's) proposal: the interoceptive inference view of emotion

According to the *interoceptive inference view of emotion* (IIE), emotions arise from interoceptive hypotheses about the causes of incoming interoceptive signals. In this view, in direct analogy to the way in which visual percepts are formed (Seth, 2015a), for an emotion to arise, emotion models/hypotheses need to supress from the top-down incoming interoceptive inputs. Mismatches between predicted interoceptive signals and actual inputs result in interoceptive PE signals that cause to replace the considered perceptual interoceptive hypothesis. Analogously to vision, even though finding a fitting interoceptive hypothesis requires that the whole system constantly contributes to that task across all levels of the interoceptive hierarchy, as Hohwy remarks, in IIE,

emotions are "reduced to basic interoceptive states" (Hohwy, 2013, p.243) and our perception of them: "emotion arises as a kind of *perceptual inference* on our own internal states." (*Ibid*. italics are mine). Emotions result then via interoceptive perceptual inference *simpliciter*, just as visual percepts result via visual perceptual inference *simpliciter*. In other words, according to IIE, emotions are interoceptive percepts which are formed guided by an emotion-hypothesis. The claim is that, once interoceptive afferents are triggered by some external event, the winning emotion perceptual hypothesis 'explains away' that interoceptive activity *initially* triggered by such an external event. In this manner an interoceptive percept is formed, which exhibits the content of such a wining perceptual emotion-hypothesis.

IIE is problematic

In this Thesis, I argue that IIE is problematic. Perceptual, interoceptive theories of emotion exhibit problematic assumptions. Being IIE one of such theories, it is also misguided. More specifically, remember that in the PP framework, the generative models from which hypotheses are put forward extract the regularities that configure the structure of the world. This, in order to be able to issue successful predictions about worldly (hidden) causes of incoming signals. In other words, priors are learned from experience (via model/hypothesis selection and revision in light of precision-weighted PE), and over time they manage to recapitulate the regularities of the world (Hohwy, 2013). According to IIE, emotions result from interoceptive predictions. Where do interoceptive priors come from? From the causal regularities that obtain in the inner world, i.e., in the physiological landscape. Thus, IIE is committed to the assumption that there must be causal regularities pertaining to emotion in the physiological domain. However, I will show that this is not the case: there are no bodily, physiological regularities relative anger, fear, joy, sadness, etc. There are no emotions configured in the physiological landscape. Then, it is unlikely that emotion models get to encode interoceptive expectations in the way required by IIE. Therefore, IIE should be amended.

Interoceptive inference and the interoceptive inference theory of affective valence

Even though Seth and Hohwy, in their discussion of 'interoceptive inference', refer to typical instances of emotion as the explanandum (e.g., anger, fear, joy), the interoceptive inference approach to interoception can also be considered as an account *not* of emotions *per se* (anger, fear, joy), but of affective states more generally. That is, of subjective feeling states: mental states that feel good or bad. In other words, the interoceptive inference approach to interoception can also be considered as an account of *valence*. This is clear from the fact that, in this respect, Seth quotes the work of Damasio (1994) and Gu & Fitzgerald (2014) on decision making, which relies on 'subjective feeling states' more generally, rather than on emotion *per se* (and the latter work is indeed based on the interoceptive inference approach). In fact, outside philosophy, 'emotion' is usually used to refer to that more general sort of mental state (i.e., affect) (Barrett & Bliss-Moreau, 2009). Seth and Hohwy would certainly agree then that their treatment of 'interoceptive inference' is not meant to be an account of emotion *per se* exclusively, in the restricted sense in which the term 'emotion' is used in philosophy. Their treatment of 'interoceptive inference' is meant to apply also to affective states more generally, such as, for example, hunger and thirst (in fact, they do mention that kind of mental states in their treatment of 'interoceptive inference'). Then, the interoceptive inference approach can also be taken to be an account of affective states in general. Or more precisely, of the affective aspect of 'subjective feeling states', of mental states that feel good or bad. Now, the essential component part of all affective states is *valence* (or affective valence). Valence is what makes affective mental states, such as hunger, pain and joy, states that feel good or bad, positive or negative. That is, valence is what makes certain mental states *affective* in the first place. Thus, the interoceptive inference approach also counts as an account of valence, or affect in general. In this respect, the claim put forward by the interoceptive inference approach is that the affective aspect of subjective feeling states—i.e., affective valence—arise by minimizing interoceptive PE, in direct analogy to the way

vision operates. That is, affective valence results from a perceptual, interoceptive process. Let's call this the *interoceptive inference theory of valence* (ITV).

The perceptual hypotheses in question, insofar as they are *interoceptive* perceptual hypotheses, predict sensory, interoceptive activity. Thus, in ITV, valence properties count as sensory, interoceptive representations. In ITV, affective valence must be seen then as a sensory, perceptual phenomenon—also, considering that everything in the PP machinery seems to be sensory, in PP valence properties should also be constituted by sensory processing.

This poses a major challenge to ITV. What might be called *non-sensory signal theories of valence* (NSS) (Prinz, 2004, 2010; Carruthers, 2011) have compellingly showed that valence cannot be understood as a perceptual phenomenon. So any account that models valence on perception seems to be doomed to fail. Contrary to ITV, valence seems not to be a sensory representation. ITV seems not to be tenable.

Roughly, NSS holds that valence amounts to a motivating inner signal that "marks" mental states as good or bad, welcome or unwelcome. These views also share the claim that valence is *not* a sensory/perceptual state, and, being just a signal, it neither amounts to an amodal (or multimodal) representation of any kind (e.g., a concept). In other words, the valenced aspect of a sensory experience is regarded as something that "attaches" to the sensory experience itself, *the latter being something distinct from the former*. That is, valence and bodily perception are independent phenomena. Therefore, contrary to ITV, what makes mental states feel good or bad—i.e., valence—is not a sensory/perceptual phenomenon of any kind. In Prinz's version of NSS, valence amounts to inner-reinforcer signals that command the maintenance or cessation of certain (perceptual representations of) patterns of bodily changes. Similarly, according to Carruthers, valence amounts to a non-sensory signal that combines with perceptual representations of stimuli in different modalities, making the latter attractive or repellent for the agent. Thus, in both versions of NSS, interoceptive perception is

something distinct, separate from affective valence, the latter being a *non-sensory* component in the furniture of the affective mind. Affective valence is not then grounded in the interoceptive system. Now, considering that NSS can accommodate more desiderata than current competing accounts on the nature of affective valence (such as the *approach/avoidance* and the *evaluative view*), and it can explain the intuitiveness of the latter, NSS is quite explanatorily powerful. ITV faces a pressing problem: valence seems to be independent from interoceptive perception.

This worry that affective valence cannot be constituted by sensory processing (which includes the minimization interoceptive PE via interoceptive perceptual inference), as the interoceptive inference approach must claim, has not yet been identified and addressed by proponents of the interoceptive inference approach to affective phenomena (Seth and Hohwy). As there seems to be a gap between perceptual processing and valence, the yet emerging view that affective valence arises via interoceptive processing must begin to deal with the worry in question.

The dilemma faced by the interoceptive inference hypothesis (and troubles for PP's ambitions)

Then, the PP framework faces a dilemma. On the one hand, IIE cannot account for emotion *per se*, since it counts as a perceptual, interoceptive theory of emotion. This kind of theories exhibit deeply problematic assumptions. Being an interoceptive theory of emotion, those assumptions are also shared by IIE. On the other hand, the interoceptive inference approach can also be taken as an account of the affective aspect of 'subjective feeling states', i.e., of valence. This is what I called above ITV. As such, the interoceptive inference approach counts as a perceptual, interoceptive theory of valence. In other words, ITV is committed to the view that valence is a sensory, interoceptive phenomenon. However, this kind of theories have been compellingly discredited by NSS (Prinz, 2004, 2010; Carruthers, 2011). According to NSS, valence is not a sensory/perceptual phenomenon of any kind, but rather a signal that only

"attaches" to sensory representations, the former being independent from the latter. Thus, insofar as ITV counts as a perceptual, interoceptive theory of valence, it cannot either account for valence, the defining property of affective mental states. Therefore, the PP framework seems to lack the resources to account for emotion, and also to account for the essence of affect (i.e., valence). Mental life is affective to the core. If PP cannot account for that aspect of mentality, it simply fails as an overarching, unifying framework in cognitive science. This is a major drawback for PP ambitions.

The proposal: thesis in a nutshell

There is a way out of this dilemma. I aim to systematize, clarify, and expand the interoceptive inference approach to interoceptive processing proposed by Seth (and Hohwy) in order to defend two intertwined claims. Firstly, I will show that, contrary to NSS's tenets, valence is indeed constituted by interoceptive perceptions, and that such process results from PE minimization. ITV holds.

In this respect, I argue that Prinz's and Carruthers's arguments against the view that valence is a perceptual phenomenon are misguided, and that NSS exhibits problematic assumptions on its own. Very roughly, and leaving secondary arguments for later, firstly, contrary to Carruthers's tenets, I show that the evidence that he presents does not suggest that felt arousal—which, as he rightly maintains, is uncontroversially grounded in interoception—needs to be considered as a separate construct than valence. Nothing in Carruthers's arguments precludes then that valence is grounded in interoception.

Secondly, the main worry that Prinz puts forward against the view that valence is a perceptual phenomenon—which, as such, can be felt—is that the wide variation in the way in which positive (or negative) feelings feel indicates that they do not share a

common perceived felt aspect, namely, positive (or negative) valence. Thus, valence is not something perceptual, i.e., something that can be felt.

I show that the PP framework offers a nice way out of this problem. As I mentioned above, in the PP framework, the assessment of the *precision* of PE, which is identified with attention, occupies a central role in the whole PP inferential machine. Remember that precisions determine which aspects of the signal are ignored (as they are deemed to be unreliable), and which aspects of the signal are amplified (as they are deemed to be reliable). Thus, in contexts where the input is deemed to be unreliable, top-down hypotheses will be assigned more weight than usual. In other contexts, the input will be deemed reliable. In these latter type of cases, PE will be taken to be informative, and it will then be used to guide the current hypothesis-selection process and, in the long term, learning.

Precisions, together with *accuracy*, play a key role in determining which contents become phenomenally conscious. Roughly, models from which perceptual hypotheses are put forward exhibit more *accuracy* in case they better represent the causal structure of the world, as the more PE is minimized. Importantly, as Hohwy (2012) suggests, perceptual states become conscious in case they exhibit a relatively significant degree of both, accuracy and precision. Now, our expectations of precision depend on context. Among other things, this implies not only that, during percept formation, precisions in different modalities are differentially inferred, but that, depending on context, precisions differ within the same modality for different sensory features.

The key is to consider precisions as key for dealing with Prinz's worry. ITV can straightforwardly explain the rich variation in our experience of positive (and negative) feelings by appealing to the context-dependent variation of precision-weighting for different sensory attributes during interoceptive percept formation. Thus, depending on context, certain interoceptive attributes can be amplified, while others ignored during interoceptive perceptual inference.

On the other hand, the experience of emotion is accompanied in a phenomenologically unified way by experiences across many modalities. Then, ITV can also explain the rich variation in question by appealing to the context-dependent variation of precision weighting across the different modalities that play a role during the valenced experience of emotion. That is, precisions determine which aspects of the whole multimodal stream of sensory input are to be ignored, and which aspects of the signal will be amplified. Moreover, precisions can be differentially assigned at each level of the perceptual hierarchy. Thus, the same bounded interoceptive features that compose an interoceptive percept can be, depending on context, differentially attended. In this manner, bodily experience can exhibit some variation.

Finally, not only the arguments of the defenders of NSS against the view that valence is a perceptual phenomenon are misguided, NSS is also problematic on its own. While Carruthers's version is committed to an ill-motivated dissociation between valence and arousal; Prinz's version implies an implausible neural basis for valence, and assumes an implausible dissociation between motivation and interoception. This leaves the door open for the view that interoceptive percepts formed via interoceptive perceptual inference can indeed account for affect in general (i.e., it can account for valence). In fact, several strands of empirical evidence (and some theoretical considerations) suggest that this is the case. The proposed view can reply to objections, and satisfies agreed desiderata for a theory of valence.

However, the second horn of the dilemma presented above turns out to be more difficult to deal with. IIE seems not to be able to account for emotions *per se*. Interoceptive, perceptual theories of emotion are indeed doomed to fail—as there are no significant regularities pertaining to emotion in the physiological domain. There are no emotions configured in the physiological landscape. So a PP account of emotion should not model emotion as interoceptive perceptions. Contrary to IIE's claim, predicting interoceptive signals during perceptual inference cannot be then what is

primary in emotion generation. Thus, Seth's (and Hohwy's) IIE lacks a way to account for emotions *per se*.

This might sound rather puzzling. On the one hand, forming an interoceptive percept by predicting interoceptive signals cannot be what is primary in emotion generation. On the other hand, common sense (and also experimental research) tells us that every time we experience an emotion, however, this experience is accompanied by interoceptive, bodily feelings. Of course, this is no real puzzle. This simply suggests that, even though having an emotion does not consist in perceiving interoceptive changes, having an emotion does involves some type of process that must be intertwined with interoception. In other words, something more that interoceptive percept formation is needed so as to account for emotion, and it must be something closely intertwined with interoceptive, bodily perception. It is at this juncture that the second claim proposed in this Thesis comes forward.

The second claim I will defend is that 'interoceptive inferences' can indeed account for emotion. However, this demands amending Seth's proposal in a key respect. I agree with IIE's claim that emotions arise by minimizing interoceptive PE—after all, common sense (and also experimental research) indicates that interoception is part of emotion generation. I think that this more general claim is on the right track. However, I will propose that, contrary to IIE, emotions do *not* arise by minimizing interoceptive PE in the *specific* way proposed by IIE. I will propose that emotions, instead of arising via interoceptive *perceptual inference*, arise via interoceptive *active inference*. In other words, emotions are not about forming an interoceptive percept of a certain sort. Rather, emotions are strategies for changing an interoceptive percept that has already been formed (via interoceptive perceptual inference). That is, emotions are specific strategies for regulating valence (affect). Now, interoceptive percepts (i.e., valence) inform about our homeostatic condition. Then, emotions are better seen as specific strategies for regulating homeostasis.

More precisely, I will suggest that emotions arise by minimizing interoceptive PE via what I will call *external interoceptive actions.* Here the task consists in minimizing the discrepancy between an *already formed interoceptive percept* that informs about current homeostatic condition and the hard-wired goal (or expectation) of stable homeostasis. Such discrepancy amounts to high-level interoceptive PE. *External interoceptive actions* consists in changing the external environment in order to change such an interoceptive percept. Insofar as external interoceptive actions are a form of active inference, they require *representations of 'sensorimotor contingencies'*. Roughly, counterfactual knowledge of the way in which interoceptive signals will evolve, if certain actions ensue. Insofar as external interoceptive actions require representations of 'sensorimotor contingencies', emotion models need to store such kind of knowledge. I will defend the view that emotions are specific forms of 'external interoceptive actions': A certain emotion *E* amounts to a specific strategy for minimizing interoceptive PE by way of a specific set of stored knowledge of the counterfactual relations that obtain between (possible) actions and its prospective interoceptive, sensory consequences ("if I act in this manner, interoceptive signals should evolve in such-and-such way"). An emotion arises when such knowledge is applied in order to regulate valence. In this sense, emotions are specific strategies for regulating affect by way of specific forms of action-guiding stored knowledge. As long as emotions are not identified here with interoceptive perceptions, this view does not require there to be regularities pertaining to emotion in the inner milieu. Therefore, if the proposed view holds, PP's ambitions are safe: interoceptive PE minimization can account for affect and emotion, the core of our mental life.

The suggested view turns out to be a promising view, insofar as it avoids the problems of the main kinds of theories of emotion, and it satisfies the desiderata for theories of emotion. Then, the suggested view sets the basis for an especially dedicated, focused treatment of the view that emotions arise via external interoceptive active inference, together with the development of its philosophical implications.

Global argumentative steps: Chapter's structure

In order to delimitate the kinds of affective phenomena of main interest in this Thesis, in **Chapter 1**, after discussing the construct of *affect* (Section 1.1.), I distinguish between different kinds of affective states (Section 1.2.). Each of them constitute distinct possible explanatory targets. The main distinction drawn here is that between *affect in general* and *emotions per se*. In Section 1.1., I defend the claim that valence, rather than arousal, stands out as the paramount defining construct in *affect*. Valence is the essence of affect, to put it that way. Affective mental states are mental states that are positive or negative, and valence is precisely the construct that endues mental states with such a polarity. Now, there are several kinds of affective mental states, among them: moods, drives, hedonic states, and emotions. *Emotions per se* are affective states directed at *core relational themes*. It is also important to have in mind the distinction between *emotion generation* and *emotion regulation*. Roughly, the latter phenomenon consists in changing an emotion once it has already been formed; while the former phenomenon consists in forming an emotion in the first place. A theory of emotion is generally taken to have as its explanandum the generation of emotion rather than its regulation. Then, as any other emotion theory, a theory of emotion based on the PP framework must also have emotion generation, rather than emotion regulation, as its explanatory target. The *generation* of emotion episodes is then the focus of this thesis.

In this Thesis I argue that IIE can be amended in such a way so as to become a promising view, which compares favourably to the main theories available on the nature of emotion. Then, it is relevant to discuss those aspects in which competing theories turn out to be unsatisfactory, and those aspect in which they result satisfactory. A theory of emotion that has the resources to avoid those aspects in which other theories fail, and to accommodate their advantages, should be taken to be on a promising track. So, in Section 1.3., I briefly present the main families of emotion theories, and discuss their problems and merits. This will give us a sense of those aspects that should be taken into account when developing a satisfactory view on the nature of emotion.

There are three main families of emotion theories, namely, perceptual theories, cognitive theories, and action or agential theories. Perceptual theories account for the bodily feelings characteristic of affective states quite well. However, they struggle with accounting for the intentionality of emotion. Moreover, insofar as there are no patterns of bodily changes able to configure emotions in the physiological landscape (Barret, 2006b; Quigley & Barrett, 2014), it is unlikely that emotions could arise by simply perceiving bodily changes, among other reasons (Section, 1.3.1.). On the other hand, cognitive theories straightforwardly account for the intentionality of emotion. However, cognitive theories can hardly account for the phenomenology and motivational aspect of emotion. Finally, action or agential theories also struggle with accounting for the intentionality of emotion. However, agential theories directly account for its motivational aspect. Even though currents versions of the agential theory exhibit some issues, I think that they are, in general, on the right track. The PP view on emotion to be suggested in Chapter 7 draws significantly on current versions of the agential theory. However, this variation over IIE does also integrate the insights of the main families of emotion theories mentioned above. As I will argue, the expansion of Seth's view suggested in Chapter 7 turns out to be satisfactory in that sense: A predictive processing view of emotion has the resources to integrate the insights, and avoid the problems, of the different approaches to emotion.

IIE is the first attempt of accounting for emotion in PP terms. IIE accounts for emotion generation by extending, in a direct fashion, the process by which visual percepts are formed to the case of interoceptive percept formation. That is, according to IIE, emotions are interoceptive percepts of a certain sort. Then, in **Chapter 2**, in order to better grasp IIE, I begin by presenting the PP account of visual (exteroceptive) percept formation. (Section 2.1.). Visual percept formation (and exteroceptive percept formation) is the most paradigmatic case of the workings of PP. I emphasize in this Section those aspects of predictive processing that make it especially interesting as an account of perceptual processing. Particularly, (a) the role of *precisions* (Section 2.4.), (b) the importance of cross-tall between modalities across all levels of the perceptual

hierarchy, and the constant influence of high-level multimodal and amodal knowledge (Section 2.6.), (c) the role of stored knowledge about 'sensorimotor contingencies' during active inference (Section 2.5.2.), and (d) the fact that, via learning, the cortical hierarchy recapitulates regularities in the world (Section 2.7.).

In **Chapter 3**, I present, systematize, and clarify Seth's (and Hohwy's) proposal that interoception operates under principles of PP. Importantly, for expository and systematization purposes, I think that a distinction should be made. This is the distinction between the *interoceptive inference approach to interoception*, the *interoceptive inference theory of valence* (ITV), and the *interoceptive inference view of emotion* (IIE). In the *interoceptive inference approach to interoception*, the principles of PP have been extended to account for interoceptive processing *simpliciter*, i.e., the perception of our homeostatic, physiological condition. Now, what might be called the *interoceptive inference theory of valence* (ITV) holds that the interoceptive inference approach can be used so as to account for affective states more generally. That is, it can be used to account for subjective feeling states: mental states that feel good or bad. Remember that valence is what makes affective states *affective* in the first place. *Valence* is the essence of affect. ITV holds then that the interoceptive inference approach can be used so as to account for valence: Valence arises via interoceptive inference. On the other hand, what might be called the *interoceptive inference view of emotion* (IIE) holds that the interoceptive inference approach can be used so as to account for *emotions per se* (e.g., fear, anger, joy). In this Chapter, I limit myself to presenting the *interoceptive inference approach* to interoception: How do the principles of PP apply to interoceptive processing? This will set the basis to better grasp the main commitments of IIE (and ITV)—in the Chapter that follows (Chapter 4), I present and problematize IIE.

However, I begin this Chapter by discussing to some extent the notions of 'interoception', 'interoceptive percepts', and 'homeostasis'. Discussing these notions is key for my purposes. In the first place, the interoceptive inference approach to interoception, ITV, and IIE are articulated in terms of these notions. Then, discussing

them is helpful for fully grasping the underlying commitments of ITV and IIE, and also their implications. In the second place, these notions play a major role in the claims defended in this Thesis, so they will be recurring themes in the Sections and Chapters to come.

Several trends of evidence and theoretical considerations point to the claim that the principles of PP readily apply to interoceptive processing. The interoceptive inference approach to interoception is thus well motivated. After motivating this claim in Section 3.2., I present, systematize, and clarify the interoceptive inference approach to interoception. In this respect, I highlight three inferential strategies by which interoceptive PE can be minimized: *interoceptive perceptual inference*, *internal interoceptive active inference* (or 'internal interoceptive action'), and *external interoceptive active inference* (or 'external interoceptive action'). Distinguishing between these kinds of strategies is essential for the view proposed in this Thesis. I will defend the view that valence (affect) results from *interoceptive perceptual inference*; while emotions result from *external interoceptive actions*. Importantly, the latter require stored *representations of 'sensorimotor contingencies'*. Let me briefly unpack the main aspects of these strategies.

Interoceptive perceptual inference amounts to issuing and updating hypotheses from the top-down, so that the generated interoceptive predictions fit the incoming interoceptive signal. Here the task consists in forming an interoceptive percept that informs about the homeostatic, physiological condition of the organism. This is achieved by suppressing from the top-down incoming interoceptive signals, and thus minimizing interoceptive PE: the difference between the incoming interoceptive signals and the interoceptive signals expected and generated from the top-down. Here the content of the percept is determined by the content of the hypothesis that manages to successfully fit incoming signals. Just as in the case of visual percept formation via perceptual inference, in the interoceptive case, once an interoceptive percept is formed, experience arises—in this case, as I will defend, a *valenced* feeling state arises (e.g., thirst).

Just as in the case of vision, in the case of interoception, PE can be also minimized by changing the input so as to fit the hypothesis (i.e., changing the inner world—and thus incoming interoceptive signals—so as to fit the model), instead of changing the hypothesis to fit the input (i.e., changing the model to fit the inner world), as in the above strategy. That is, interoceptive PE can be also minimized via *active inference*. Contrary to the above strategy, in this sort of strategy, the direction of fit is world-to-mind, i.e., changing the (inner) world so as to fit the distal, hard-wired goal (or high-level expectation) of maintaining homeostasis. Note that this presupposes that an interoceptive percept that informs about a deviation from homeostatic balance, or a certain physiological change, has already being formed. As I mentioned above, this takes place via interoceptive perceptual inference. When such a homeostatic imbalance is thus detected, this triggers, what might be called high-level *interoceptive PE*: the difference now between the expected "goal state" of homeostatic balance and the current interoceptive percept that informs of a homeostatic deviation. Now the organism needs to change its inner world so as to fit the hard-wired goal (or expectation) of maintaining homeostasis. The main task of the interoceptive system is *not* now forming a percept, but rather bringing physiological variables to their goal state by minimizing high-level interoceptive PE. This calls for interoceptive action.

As I mentioned above, in the case of interoception there are two kinds of strategies of active inference, namely, *interoceptive internal actions*, and *interoceptive external actions*.

Interoceptive internal action consist then in changing physiological inputs so as to fit an interoceptive goal state (or expected state of physiological balance). In order to achieve this, "physiological policies" (autonomic reflexes) are engaged. They consist in making use of resources that are already available within the organism, such as secreting vasopressin (this requires transiently assigning low precision to interoceptive PE). However, internal interoceptive actions rarely can rectify homeostatic imbalances

by themselves (Craig, 2015). Here is when *external interoceptive actions* come into play.

External interoceptive actions consist in modifying the external environment, and your situation in it, in order to rectify homeostasis. Interestingly, for 'external interoceptive action' to occur, interoceptive predictive models at higher levels require to encode, as Seth (2015a) notes, counterfactual knowledge of the sensory consequences of action. More precisely, knowledge of the counterfactual relations between, on the one hand, particular exteroceptive states, motors states, and changes to the environment, and, on the other hand, the interoceptive activity that they would ensue. Such stored links between exteroceptive and interoceptive signals allow the system to know which interoceptive states can be obtained via which exteroceptive (and proprioceptive) states. This permits that an action can be found that achieves to cause the desired interoceptive goal-state that homeostasis demands.

After systematizing and refining the manner by which the principles of PP apply to the case of interoceptive processing *simpliciter*, in **Chapter 4** I present and critically discuss IIE. That is, the claim that *emotions per se* arise from successfully minimizing interoceptive PE via interoceptive perceptual inference. In other words, the claim that emotions amount to interoceptive feelings, or interoceptive, bodily perceptions of a certain sort. To date, even though still first sketches, the most developed interoceptive inference theories of emotion are due to A. Seth (Seth, 2013; Seth et al., 2012; Seth & Critchley, 2013) and J. Hohwy (2011, 2013). In this Chapter, I present Seth's IIE and Hohwy's IIE, and critically discuss their proposals. As I mentioned above, I argue that IIE is problematic as there are no causal regularities pertaining to emotion in the physiological domain. Therefore, IIE should be amended. Interoceptive perceptual inferences are not in the driver's seat when it comes to *emotion* generation.

However, as I mentioned above, the interoceptive inference approach can also be taken to be an account of affective valence. In this respect, the claim put forward by the

interoceptive inference approach is that affective *valence* arises by minimizing interoceptive PE via interoceptive *perceptual* inference. This is what I called *the interoceptive inference theory of valence* (ITV). Contrary to non-sensory signal theories of valence (NSS), I think this view is on the right track.

Then, in **Chapter 5**, in order to show that affective valence is likely to be a perceptual phenomenon, I begin by briefly characterizing valence (Section 5.1.), and introduce widely agreed desiderata for a theory of valence (Section 5.2.). Then, in Section 5.3., I present Prinz's and Carruthers's versions of the non-sensory signal theory of valence (NSS) (Sections 5.3.1. and 5.3.2., respectively), and their explanatory advantages (Section 5.3.3.). Remember that NSS holds that the valenced aspect of a sensory experience is regarded as something that "attaches" to the sensory experience itself, the latter being something distinct from the former. In Prinz's version of NSS, valence amounts to inner-reinforcer signals that command the maintenance or cessation of certain perceptual representations of patterns of bodily changes. Similarly, according to Carruthers, valence amounts to a non-sensory signal that combines with representations of stimuli in different modalities, making the latter attractive or repellent for the agent. As I will show in Section 5.4., NSS faces decisive problems on its own (see above), which makes NSS a poor candidate for a compelling, plausible theory of valence. Thus, the door is open for the view that valence is a sensory item in the furniture of the perceptual predictive machine posited by the PP framework.

In **Chapter 6**, I defend the view that valence properties are indeed a sensory, interoceptive phenomenon. That is, while in Chapter 5 I argue that the case for NSS is rather weak, in Chapter 6 I argue for the view that valence is indeed a sensory phenomenon. Once affective valence is taken to arise via interoceptive inference, the view that valence is a sensory phenomenon can reply to the objections made by defenders of NSS. Thus, affective valence is likely to be a perceptual phenomenon. ITV seems to hold.

In order to show that, I critically discuss some of the main possible ways in which valence properties can be understood in the PP framework (Section 6.1.). Even though each of these views exhibits some difficulties, they all point towards a more global PP view on the nature of valence. All these views point towards the claim that valence amounts to an interoceptive percept, formed via interoceptive perceptual inference, that informs about positive and negative homeostatic changes (Section 6.2.). Theoretical and empirical considerations suggest that the different aspects involved in this claim might be the case (Section 6.3.). In Section 6.4., I show that the view in question can reply to the objections typically faced by views that, as the one defended here, identify valence with mental states that can be felt—e.g., (dis)pleasure. Finally, in Section 6.5, I show that this view on the nature of valence satisfies desiderata for a theory of valence.

However, as I mentioned above, IIE cannot account for emotion *per se*, since there are no regularities pertaining to emotion in the physiological domain. Even though PP can account for valence (affect)—as I argue in this Chapter—this leaves PP without an account of emotion. This is major drawback for PP ambitions.

In **Chapter 7**, I will suggest that the PP approach to interoception can indeed be used to account for emotion *per se*. This requires amending IIE. As I mentioned above, I will argue that, rather than via interoceptive perceptual inference, as IIE holds, emotions arise by minimizing interoceptive PE via *external interoceptive actions*. Here the task consists in minimizing the discrepancy between already formed interoceptive percepts and the hard-wired goal (or expectation) of stable homeostasis. As we saw, *external interoceptive actions* require stored knowledge of 'sensorimotor contingencies'. I defend the view that a certain emotion *E* amounts to a strategy for minimizing interoceptive PE by way of a specific set of representations of 'sensorimotor contingencies'. As I already mentioned, according to this view, emotions are individuated by the kind stored knowledge of 'sensorimotor contingencies' that is brought to bear in the control of valence.

I begin this Chapter by briefly motivating the view that emotion might be forms of action (rather than forms of perception) (Section, 7.1.). Then, after presenting the claim that emotions arise by minimizing interoceptive PE via external interoceptive actions (Section 7.2.), I discus the idea that the kind of stored knowledge of 'sensorimotor contingencies' required for external interoceptive actions can be taken to consist in emotion-specific action-oriented representations, encoded at higher levels of the cortical hierarchy. Insofar as these representations (or 'chunks' of the generative model) encode abstract (Section 7.3.) sets of knowledge about the same category in the world, they can be taken to consist in emotion 'concepts' (Section 7.2.2). The latter have both, mind-to-world and world-to-mind direction of fit (Section 7.4.). In Section 7.5., I highlight the way in which the suggested expansion of IIE can account for those aspects that should be taken into account when developing a satisfactory view on the nature of emotion.

1. Delineating the explanatory target

In this Chapter, I aim to delimitate the kinds of affective phenomena of main interest in this Thesis. Each of them constitute distinct possible explanatory targets. The main distinction drawn here is that between *affect in general* (or 'subjective feeling states') and *emotions per se*. I defend the claim that valence, rather than arousal, stands out as the paramount defining construct in *affect*. Valence is the essence of affect (Section 1.1.). Now, there are several kinds of affective, valenced states. Among them, *emotions per se*. Emotions *per se* are affective states directed at *core relational themes* (Section 1.2.). I also distinguish between *emotion generation* and *emotion regulation*, and between *emotion states* and *emotion episodes*. A theory of emotion is generally taken to have as its explanandum the *generation* of *emotion episodes*. That is the explanatory target of this Thesis. Then, in Section 1.3., I discuss some of the main kinds of views on how emotion episodes are generated, namely, perceptual theories, cognitive theories, and action theories (and also hybrid theories). I briefly discuss those aspects in which these different global approaches turn out to be unsatisfactory, and those aspect in which they result satisfactory. This will give us a sense of those aspects that should be taken into account when developing a satisfactory view on the nature of emotion.

1.1. Characterizing affect: valence as the essence of affect

1.1.1. What is affect? Valence at a certain degree of activation

Affect is the psychological construct in virtue of which certain mental states count as being part of the affective realm. In other words, disparate kinds of mental states, such as moods, emotions, and drives, have in common that they all have *affect* as a necessary component part. In other words, affect is the construct that characterizes *subjective feeling states* of all kinds—'subjective feeling states' amount to mental

states that feel good or bad at a certain intensity level (more on this below)—making the latter *affective* mental states in the first place.

Affect is a central construct in the sciences of the mind. It is usually posited under the label 'emotion' in theories about a wide range of phenomena, such as moral judgment (Greene et al., 2001; Haidt, 2001), decision-making (Bechara & Damasio, 2005), attention (Pessoa, 2008), drives (Craig, 2002), vision (Barrett & Bar, 2009), etc. (see Barrett & Bliss-Moreau, 2009). However, properly speaking, 'emotion' refers to a more circumscribed kind of affective state (more on this below). That is, emotion amounts to just one type of affective state, among many others. Emotion is not itself then the kind of construct that defines the whole affective realm. Affect is such a construct.

Then, what is affect? Affect is the mental state that arises as the physiological, bodily changes that an organism constantly undergoes are continually detected by the central nervous system. As Barrett and Bars remark, "psychology refers to these internal bodily changes as 'affective'" (Barrett & Bars, 2009, p. 1325).

> "In English, the word 'affect' means 'to produce a change'. To be affected by something is to be influenced by it. In psychology, affect refers to a specific kind of influence—something's ability to influence a person's body state. Sometimes, the resulting bodily sensations in the core of the body are experienced as physical symptoms (such as being wound up from drinking too much coffee, fatigued from not enough sleep or energetic from exercise). Much of the time, sensations from the body are experienced as simple feelings of pleasure or displeasure with some degree of activation, either alone or as an emotion (figure 1; see Russell & Barrett 1999; Barrett & Bliss-Moreau in press). At still other times, bodily changes are too subtle to be consciously experienced at all." (Barrett & Bars, 2009, p. 1327-1328)

As it is pointed out in the above quotation, detected bodily changes give rise to mainly two kinds of states. On the one hand, physical symptoms, such as, for example, fatigue, agitation, breathing, muscle tension, and stomach motility. On the other hand, the more diffuse feelings of pleasure and displeasure (i.e., feelings of valence) with some degree of activation or deactivation (i.e., arousal level). However, note that physical

symptoms, even though they are, at the personal level, more circumscribed, easily located feelings, which pull attention to specific regions of the body, they are intrinsically pleasant or unpleasant at certain level of activation or deactivation. This evince that affect is characterized by the latter kinds of states. In other words, even though during the experience of 'physical symptoms' their physical location and physical characteristics are predominantly attended, they always feel good and bad at a certain degree of activation, even though this 'feeling' might not be significantly attended, lying thus at the background of experience. That is, detected bodily changes always give rise to the more diffuse feelings of pleasure and displeasure (i.e., feelings of valence) with some degree of activation or deactivation (i.e., arousal level). Thus, affect can be better characterized by the states of valence and its degree of activation. In fact, it is widely agreed that affect amounts to *valence* and *arousal* (Barret & Russell, 1999; Mauss & Robinson, 2009; Prinz, 2004)[1]. Affect consists then in a mental state that feels good or bad (valence) at a certain intensity level (arousal)[2].

1.1.2. Characterizing valence

Valence can be characterized in the following way. Certain emotions are agreeable, while other emotions are disagreeable. That is, there are positive emotions and negative emotions. For example, joy, pride, love, and amusement typically are positive emotions; while anger, fear, guilt, and contempt typically are negative emotions. Emotions are classified in this way in virtue of the character of its valence. Certain emotions are positive emotions since they have as a component positive valence; and certain emotions are negative emotions since they have as a component negative valence[3]. Not only emotions have valence as a component. Given that valence is the key construct in affect, all affective phenomena are valenced. For example, drives or

[1] Certainly, so as to characterize the minimal components that configure affect, other dimensions have being proposed. However, these proposals all include valence at a certain degree of activation as a defining, non-negotiable component of affect (see Barrett & Bliss-Moreau, 2009).
[2] I characterize 'affect' as mental states that *feel* in a certain way only for introductory purposes. As we will see in coming Chapters, affect (i.e., valence) can take place outside consciousness.
[3] Note that this way of characterizing valence leaves open the possibility that a certain emotion type *E* can have different valence value on different occasions.

motivations also exhibit such a positive and negative character as, for example, hunger and thirst, which are negatively valenced. Also moods are valenced, as depression and anxiety, which typically have negative valence as a component.

More generally, 'valence' refers then to the *positive* and *negative* character exhibited by the mental states that belong to the realm of the affective. Just what such positivity and negativity consists of will depend on which theory of valence turns out to be the case. For example, according to the *hedonic theory*, valence amounts to (dis)pleasure, and according to what might be called the *behavioural theory*, valence amounts to approach/avoidance actions (or dispositions to behave)—in Chapters 5 and 6 I deal with issue of which theory of valence turns out to be the most promising.

1.1.2.1. Valence is a descriptive, psychological construct

Let me clarify a point of potential controversy. Valence is not only part of our folk psychological understanding of the nature of emotion, but it is also a construct that plays a fundamental role in the scientific study of emotion (see, e.g., Barrett, 2006; Russell, 2003; Berridge & Kringelbach, 2015), to the point that, for some theorists, valence is one of the main building blocks of emotion (Barrett, 2006; Russell, 2003). So note that the notion of valence in which I am interested in this Chapter/Thesis is a non-normative notion that plays an explanatory role in psychology. Thus, contrary to what a few researchers have pointed out (e.g., Charchland, 2005; Picard, 1997; Solomon, 2001), when it is said, in the affective sciences, that an emotion is *positive* or *negative* (i.e., that it has positive or negative valence) it is *not* being said that such an emotion is positive or negative in the sense of being *good* or *bad* normatively, in any epistemic, ethical or prudential sense. Valence is neither an ethical nor a prudential construct; it is a psychological construct that plays a role in affective sciences. This psychological notion of valence is the notion with which I am dealing in this Thesis.

1.1.3. Characterizing arousal

Now, *arousal* refers to the degree of *activation* exhibited by affective mental states. Arousal is widely taken to consist in a single dimension, ranging from low activation, as in the case of meditative calmness, to high activation, as in the case of euphoria (e.g., Barrett & Russell, 1999). Properly speaking the kind of activation in question is not the same as the *intensity* that an affective state can have. Think of depression. A state of depression can be very intense, as when the depressed subject lies on the couch the whole day just looking at the tip of her shoes. However, in this kind of case, the subject finds herself in a state of deactivation, so her arousal level is supposed to be low. In this respect, intensity might be seen as the saliency that the characteristic aspects of the affective state in question can exhibit at a certain time. Thus, in a state of intense depression, its typical features, such as lack of motivation, a sense of helplessness, and suicidal thoughts, become particularly salient. However, even though those features become very salient for the subject, there is a constant state of deactivation. Thus, intensity and arousal should be seen as different constructs.

1.1.3.1. Problems regarding the construct of arousal: arousal is not general sympathetic activation

1.1.3.1.1. Heart-rate as a fair way to operationalize arousal

Arousal, understood as degree of activation, has been traditionally taken to depend on the excitation of reticular activating system in the brain stem (Carruthers, 2011; see Colman, 2015). Thus, in this sense, arousal has been traditionally seen as *general physiological activation* (or general sympathetic activation). Insofar as it is seen as general sympathetic activation, besides electrodermal and other cardiovascular responses, arousal is usually operationalized by measures of increased (high activation) and decreased (low activation) heart rate. I think this is a somehow

acceptable way to operationalize arousal. Intuitively, affective states that we take to be high in terms of activation, such as fury and euphoria, consistently exhibit increased heart rate; while affective states that we take to be low in terms of activation, such as sleepiness and calmness, consistently exhibit decreased heart rate. Moreover, affective sates that differ only in terms of their degree of activation (i.e., states with the same valence value), such as the cases of anger and rage, seem to involve corresponding differences in heart rate. Empirical evidence supports this intuition, as people who are comparatively more sensitive to their own heart activity report experiencing more "intense" emotional experiences (Wiens et al., 2000).

1.1.3.1.2. Electrodermal activity as a problematic measure of arousal

Other measures associated with sympathetic activation have been used to operationalize arousal – in the sense of arousal that is relevant here, namely, as the degree of activation that affective mental states exhibit. One of such measures is electrodermal activity (EDA) – also known as 'sympathetic skin response'. EDA tracks sweating, which is triggered by certain kinds of sympathetic activation. Insofar as EDA is considered to be the result of sympathetic nervous system activity alone, EDA has been taken to be a proper indicator of general sympathetic arousal (e.g., Critchley, 2002). However, I think that this is *not* a good way to operationalize arousal, insofar as the latter is taken to be a critical dimension of affective states. This is the case since, intuitively, many cases of affective states that exhibit high activation or intensity do not seem to involve significant increases in sweat glands activity. That is, in many cases one can be in an increasing state of strong pleasantness or unpleasantness, and at the same time undergo rather insignificant sweat glands activity. For example, think of the mystic's elation arising from her sense of oneness with the universe, or the strong amusement that results from a comedian's joke, or the great pride of a father in the graduation of his child. Examples abound. Moreover, affective states that differ in terms of their degree of arousal, such as the feelings of confidence and gladness (Russell, 1999), do *not* seem to involve corresponding, consistent differences in sweating. Certainly, if we take heart rate as the most

acceptable way to operationalize arousal, counterexamples can also be found, but they do not proliferate as in the case of EDA.

1.1.3.1.3. Arousal is not general sympathetic activation

It could be thought that, instead of operationalizing arousal by just one measure, arousal can be better operationalized by a *set* of joint measures. Thus, in conjunction with heart rate, EDA, and other measures usually taken to be indicatives of sympathetic activation, they could all be *jointly* taken to indicate the degree of arousal exhibited by a certain affective state. However, this approach is doomed to fail. Physiological measures commonly used to operationalize arousal, such as the ones above, simply do not show significant co-variation with each other (Bernston & Cacioppo, 2007; Lacey, 1959, 1967), in such a way that they could be jointly taken to reflect the activity of a unified construct, namely, arousal. Crucially, this observation has led to dismiss the construct of *general sympathetic arousal* promoted by arousal theory—and which supposedly underlies "fight-or-flight" responses. Roughly, according to the latter theory, there is a single mechanism that controls all sympathetic/autonomic effectors during affective states high in activation, namely, the arousal mechanism. As I mentioned above, such a mechanism is commonly considered to be realized in the reticular formation of the brainstem. A key prediction of this approach is that physiological measures of sympathetic activity should significantly co-vary within and across individuals. However, as we saw above, that turns out not to be the case (Bernston & Cacioppo, 2007; Cacioppo et al., 2000; Lacey, 1959, 1967). The notion of general sympathetic arousal does not then pick up a real psychophysiological kind. Let me insist, there is no such thing as *general sympathetic arousal*. Taking into account that arousal, understood as the degree of (de)activation exhibited by affective states, is typically identified with general sympathetic activation, the very notion of arousal as a defining dimension of affect begins to look doubtful.

Even more, organism always find themselves in a constant stream of changing affective states (see, e.g., Barrett, 2006a). Thus, if arousal, understood as general sympathetic activation, is a key construct in affect, then it should be constantly occurring. However, general sympathetic activation is not a state that constantly occurs. Think of, for example, a bad posture. After a while, being in a bad posture brings forth a bad feeling at a certain level of "intensity" or activation. In other words, being in a bad posture typically elicit an affective state. In this kind of case, note that there is no general sympathetic activation, even though such a resulting feeling counts as an affective state that exhibit a certain degree of "intensity". Thus, arousal, understood as a key construct of affect, should not be identified with general sympathetic activation.

1.1.4. Valence as the 'mark' of affect

In fact, as some argue, "'arousal' is not a particularly useful overarching construct" (Cacioppo et al., 1991, p. 326). This is the case since, among other reasons, the measures commonly used to operationalize arousal "[provide] ambiguous information about the associated psychological event" (Cacioppo et al., 1991, p. 326), making it explanatorily unappealing compared to other constructs and reconceptualizations (see Cacioppo et al., 1991). Along similar lines, Cochrane (2009) has argued that the 'arousal' dimension of affect overlaps to such an important extent with the dimensions of valence and other dimensions, that he considers that we can simply dispense with such a construct:

> "I identify the arousal/activation dimension as the main culprit for the observed overlapping. Applying a dimension of activation to emotions is like applying a dimension of 'being coloured' to colours. It is far too general. Although the degree of arousal is highly applicable to emotions, it is not doing much conceptually useful work for us. Thus I would be more prepared to jettison this dimension than the dimension of valence." (Cochrane, 2009, pp. 384-385).

In a word, the construct of arousal does not seem to be doing much explanatory work in affective sciences. Besides that, the received view of arousal as general sympathetic activation is deeply problematic. Thus, a proper characterization of affect should not rely on such a construct.

This begins to suggest that valence, rather than arousal, is the key, preponderant construct in the structure of affect. Accordingly, in order to understand the nature of affect, theoretical efforts should be devoted to understand the nature of valence. Arousal might still be a theoretically useful aspect of affect, if it is taken to be something different than general sympathetic activation (I am not taking sides in the debate on whether arousal should be eliminated), but valence stands out as the paramount defining construct in affect, which is clearly not the case for arousal.

This insight is by no means arbitrary. Privileging valence over arousal as the mark of affect is by now a common practice in the affective sciences. For example, according to Ben-ze'ev, possessing an inherent positive/negative component (or inherent evaluative component) is necessary for a mental state to count as *affective* rather than non-affective (Ben-ze'ev, 2010, p.54). Along the same lines, for Charland "the question of valence is […] probably the most important criterion for demarcating emotion from cognition and other related domains" (Charland, 2005, p.84). Deonna and Teroni (2012) regard valence as key construct for understanding emotional phenomena, while simply ignoring arousal—the same can be said of Prinz's (2004) work. According to Fridja (2009), "*affect* refers to pleasure and pain and the processes underlying them" (Fridja, 2009, p.267-268), and considers arousal to be a separate construct than affect. Moreover, valence, but not arousal, consistently appears as the primary component in reports of emotion within and across cultures (e.g., Bottenberg, 1975; Bush, 1973; Galati et al., 2008; Herrmann & Raybeck, 1981; Lutz, 1982). Moreover, the valence dimension of affect is practically always present in reports of affective states (see, e.g., Barrett, 1998). However, the arousal dimension does not systematically appear in reports of affective states. Relatedly, when subjects must draw

distinctions between affective states, they always do it in terms of valence. However, this is not the case for arousal (Feldman, 1995).

Furthermore, privileging valence over arousal is also motivated by the default intuition that positive and negative affective experiences *themselves*, intrinsically, can differ in "intensity", being "intensity" an intrinsic property of positive and negative affective experiences, just as, for example, lower and higher pitches *themselves*, intrinsically, can exhibit quiet or loud volume ('loudness'), being loudness an intrinsic property of pitch. Thus, instead of being a separate construct than valence, arousal could be taken to be a property of valence. For example, arousal could be seen as the "volume" that valence can take. Thus, valence is the key axis of affect, while arousal only a way in which valence can occur.

Now, remember that affect is characterized as the detection of internal bodily changes—"psychology refers to these internal bodily changes as 'affective'" (Barrett & Bars, 2009, p. 1325). Then, it could be argued that arousal, rather than valence, is the key construct in affect, since autonomic responses (i.e., inner bodily changes) are widely assumed to reflect arousal (e.g., Bradley & Lang, 2000; Lang et al., 1993). For example, it has been shown that skin conductance levels increase in line with the experienced degree of arousal, and that this occurs independently of valence value (Bradley & Lang, 2000). However, it is simply not the case that autonomic activity reflects purely arousal. In fact, distinct autonomic measures function independently from each other, and they function even in opposition to each other in certain cases (Bradley & Lang, 2000). A classical explanation of this independence and functional segregation of autonomic activity is to incorporate the valence dimension (Cacioppo et al., 2000; Russell & Barrett, 1999). It is widely agreed that, for example, cardiac activity, heart rate, blood pressure, and skin conductance duration reflect affective valence (see Cacioppo et al., 2000). Also the startle response and also facial EMG indicate valence rather than arousal (Mauss & Robinson, 2009). It is simply not the case then that physiological responses reflect only activation.

Thus, considering all the above, valence, rather than arousal, is the key, preponderant construct in the structure of affect. Valence is the 'mark' of the affective. As we saw, the notion of arousal as general sympathetic activation does not work, and as a defining dimension of affect it looks doubtful. Moreover, as we saw, arousal fails to be a useful construct in affective sciences. Then, if we want to understand affect, valence is the construct that needs to be elucidated.

1.1.4.1. Against valence scepticism

However, even though, as I commented above, valence is widely taken to be the 'mark' of the affective, some argue that the construct of valence also looks doubtful. Solomon (2001) has argued that, considering that 'valence' is used in many differing, orthogonal ways, valence is not a unified construct, and that, consequently, emotion research could just simply dispense with the notion of valence. For example, Solomon notes that if we take negative valence to be identical with suffering, then we have that negative emotions (which are supposed to be unified by their shared valence component) should all involve suffering. However, different emotions involve suffering only in different senses of the term 'suffering'. The suffering of guilt, Solomon maintains, seems so different to the suffering of jealousy. Valence seems not to be a unified phenomenon. However, if in the literature there are many different, orthogonal ways of using the term 'valence'—as Solomon critically remarks—it is simply because we have many competing theories of valence. Just as the fact that there are many different, orthogonal ways of using the term 'attention' (given that there are many competing theories of attention) does not imply that psychology should dispense with this construct, the fact that there are many competing theories of valence does not imply that there might not be a theory of valence that successfully captures the distinction between positive and negative affective states. In fact, as Cochrane (2009) argues, making sense of the evaluative polarity captured by the notion of valence appears to be mandatory precisely because, as I mentioned above, "valence

consistently comes out as the primary factor" in statistical analyses of emotion reports (Cochrane, 2009, p. 387). In other words, contrary to Solomon, given that valence is ubiquitously present in our thinking about affective phenomena, rather than abandon the notion altogether, "it would be more appropriate to specify the concept more exactly, so that possible differences in interpretation do not confound experimental results." (Cochrane, 2009, p. 387). I will defend a view on the nature of valence in Chapter 5 and 6.

1.1.4.2. Let's focus on valence (though there is no need to completely eliminate the construct of arousal)

Privileging valence over arousal as the mark of affect should then be taken to be justified. There is no such thing as *general sympathetic arousal*, and arousal, understood as activation, seems not to be playing an explanatory role in affective sciences. However, as I mentioned above, common sense does tell us that affective states do seem to exhibit some sort of "intensity" together with their positive or negative character. There is no need to dispense completely with the notion of arousal. The only lesson to be drawn from the above discussion is that the view that arousal, understood as the "intensity" that affective states can have, amounts to general sympathetic activation needs to be dropped.

As valence is the defining component of affect, and "intensity" seems to be something that affective states can have, such an "intensity" could be taken to be a property of valence, as I mentioned above. In this sense, arousal could be seen as the 'volume' taken by valenced, affective states. Anyway, there is no need to settle the issue of the nature of arousal in this work, as the goal of this Section is to point to the fact that valence is the mark of affect, rather than arousal. Then, in order to understand the nature of affect, valence is what needs to be understood.

However, as I commented above, intuitively, heart rate seems to be a somehow acceptable way to operationalize such "intensity", which, as I argued, cannot be taken to be identical to general sympathetic activation. Thus, I will take heart rate as an acceptable way to operationalize arousal, whatever the latter might be. I mainly take heart rate as an acceptable way to operationalize arousal, even though arousal is not the central construct in affect, since key studies for a competing view on the nature of valence use heart rate as the key measure of arousal (Chapter 5), and, for the sake of argument, I will concede that heart rate can work in that respect. Perhaps, as I suggested above, arousal is a property of valence, the 'volume' that valence can take, and that this property is determined by heart rate. Anyway, I leave as an open question the issue of the nature of arousal, or the "intensity" that affective states can have[4].

1.2. Taxonomy of the affective space: emotions per se as the explananda

I will then assume the reality of valence, and take it as the construct that best characterizes affect[5]. Thus, *affective* mental states are mental states that characteristically exhibit valence (and arousal), i.e., a positive or negative character (at a certain level of "intensity"). There are several kinds of affective states. That is, mental states that can be taken to be inherently valenced. Together they demarcate the realm of the affective. In what follows, I briefly characterize the main kinds of affective states[6], so as to distinguish them from the kind of affective state that constitutes the target of this Thesis, namely, emotions *per se*. Since in the literature relevant for the affective sciences the label 'emotion' is usually employed to refer to a wide range of disparate affective phenomena, distinct from emotions *per se* (see Barrett & Bliss-Moreau, 2009), this Section serves the purpose of avoiding possible

[4] In this thesis I do not deal with the issue of how many dimensions affect exhibits (or of whether there are such dimensions at all). For the purposes of this Thesis, it suffices to point out that valence is the central, defining dimension of affect. I also do not deal with issue of whether positive and negative valence are independent (Larsen, McGraw, & Cacioppo, 2001) or are inversely related as they lie along a continuum (Russell, 1980). None of these issues will be relevant for the argument developed in this Thesis.

[5] For a defence of the notion of valence, see Prinz (2010).

[6] This is not meant to be an exhaustive list, as the purpose of this Section is to single out 'emotions per se' as the theoretical target of this Thesis.

confusions regarding the kind of mental state that needs to be accounted for by a view on the nature of *emotion*.

Let's get down to business. *Emotions per se* are typically understood as short-lived affective states directed at *core relational themes*. Let me briefly explain. Emotion episodes are short-lived. The emotion episodes of a certain type that an organism undergoes usually last from seconds to minutes. It is extremely rare that an emotion episode of a certain type lasts hours, and it is very hard to even imagine scenarios in which a certain emotion episode can last for more than that. Emotions are directed at core relational themes (Lazarus, 1991). It is widely agreed that emotions are individuated in virtue of what they represent in the external milieu. Emotions are meaningful. Emotions have the function of informing the precise way in which the environment/organism relation is significant for the organism's well-being (Lazarus, 1991; Prinz, 2004). Emotions *represent* such a relation. In the terminology of emotion theory, emotions represent *core relational themes* (CRT) (Lazarus, 1991). CRT individuate emotions. CRT can be seen as the external conditions that each emotion type has the function of discriminating (Prinz, 2004). For example, the external condition every single instance of anger discriminates consists in *a demeaning* offense; while the external condition every single instance of fear discriminates consist in *danger* (Lazarus, 1991). In a word, a CRT is a property of the particular eliciting event that prompted an emotion, in virtue of which the mental state that takes place amounts to a certain specific emotion type and not to another emotion type. Prototypical cases of emotion include, for example, anger, fear, guilt, joy, and pride.

Now, *moods* are one kind of affective state, distinct from emotion, which sometimes figures under the label 'emotion' in the non-philosophical literature. However, emotions and moods are different kinds of affective states. Roughly, moods are typically understood as long-lasting affective states that lack a specific intentional object, but that promote certain associated patterns of thought and action. Prototypical cases of moods are depression, anxiety, and irritability. Moods are not *about* specific object or events. For example, one is never in an irritable mood about the neighbour's

dog. In this sense, moods are 'free-floating' states, unanchored to intentional objects. That is why, as Deonna and Teroni remark, "attributions of moods (e.g., Alison is grumpy) are informative and complete without specification of any object, whereas attributions of emotions (e.g., Alison is angry) may, as we have seen, be informative but remain incomplete as long as the object is not specified" (Deonna & Teroni, 2012, p.4). That is, by knowing that Alison is grumpy one knows that she is not evaluating some specific event in any sort of way; she just finds herself feeling in a certain way, disposed to think and act 'grumpyly'. Moods, contrary to emotions, are long-lasting states that seem to last for a fair amount of time. One can be in an irritable mood for a whole day, or even a week, or months. Finally, even though they do not inform about external events (e.g., core relational themes), moods promote characteristic ways of thinking and acting (as emotions also do). Think of depression. Depression promotes negative thoughts, and biases a number of cognitive functions towards the negative (Ohman et al., 2001; Sizer, 2000). Mood are *not* the explanatory target of this Thesis.

Hedonic states are another kind of affective state, distinct from emotion, which sometimes confusingly figure under the label 'emotion' in the non-philosophical literature. Hedonic states consist of pleasure and its opposite, namely, displeasure. Hedonic states are typically taken to be simple "phenomenological qualities" which present as good (or bad) and attractive (or repellent), *independent of any sort of content*.

> "That pleasure is in itself objectless is sometimes supposed in theorizing in behavioral neuroscience, as well (e.g., Robinson and Berridge, 1993, pp. 261ff.). The same assumption is the basis for the psychologist and emotion specialist James Russell's notion of core affect, which places an in-itself objectless feeling good at the ground level of the construction of more complex positive emotion (Russell 2003)." (Katz, 2016)

On the other hand, hedonic states are typically taken to be parts of emotion (and moods), more simple then than the latter.

> "I will refer to drives and motivations and pain and pleasure as triggers or constituents of emotions, but not as emotions in the proper sense. No doubt all these devices are intended to regulate life, but it is arguable that emotions are more complex than drives and motivations, than pain and pleasure" (Damasio, 1999, p. 341).

Pleasure and displeasure are typically identified with positive and negative valence, respectively (see, e.g., Deonna & Teroni, 2012, pp. 14-15; Prinz, 2004, pp. 167-168; Colombetti, 2005). In this sense, hedonic states are not emotions themselves, but rather constituent parts of emotion that determine the good and bad feelings characteristic of emotion. In this respect, and as Katz remarks in the very first lines of his entry on pleasure,

> "Pleasure, in the inclusive usages important in thought about well-being, experience, and mind, includes the affective positivity of all joy, gladness, liking, and enjoyment – all our feeling good or happy." (Katz, 2016)

Hedonic states are not then emotions themselves, but rather they seem to be objectless constituents of emotion. Hedonic states are *not* the explanatory target of this Thesis.

Another kind of affective state, distinct from emotion, which sometimes confusingly figures under the label 'emotion', are *drives* or *homeostatic motivations*. Examples of homeostatic motivations include sexual desire, hunger, thirst, itch, pain, and temperature. This kind of affective states are undisputable cases of states that track homeostatic imbalances, such as, for example, lack of nutrients in the case of hunger. That is why homeostatic motivations, contrary to emotions, are typically taken not to represent *external* events. Homeostatic motivations are directed at *internal* events related to fundamental physiological needs.

Interestingly, homeostatic motivations, such as hunger, do not simply consist in negatively/positively valenced states that represent internal events. We do not simply call the bad feeling of an empty stomach 'hunger'. Hunger includes an urge to eat. That is why homeostatic motivations are negatively (positively) valenced states which

represent internal events *together with* certain action tendencies to modify the external environment. In the case of hunger, an urge to eat, go to the fridge, put food in your mouth, etc. The fact that homeostatic motivations include such kind of action tendency as a component part will be important for this Thesis, as I will put forward the view that emotions *per se* can be understood under the model of homeostatic motivations. Anger is much closer to hunger than we are used to think. Anyway, drives or homeostatic motivations are *not* the explanatory target of this Thesis.

As we saw, there are several kinds of affective states. I presented the main ones above. The kind of affective states that constitutes the *explananda* of this Thesis are emotions *per se*. That is, short-lived affective states directed at core relational themes, i.e., organism-environment relations, such as danger and loss.

Now, there is also a further distinction within the emotion category that might be useful. This is the distinction between *emotional states* and *emotional episodes* (Schroeder, 2006, p. 257-258). Emotional states correspond to enduring dispositions to undergo certain emotions (e.g., being afraid of spiders, being angry at one's social circumstances). Emotional episodes correspond to occurrent, short-lived emotions (e.g., finding yourself afraid of a particular spider at a certain moment, getting angry at a particular instance of discrimination against you). They are occurrent self-contained states (even though they can certainly be bounded with other mental states forming thus an unitary whole, in the same way that, while watching TV, visual and auditive states become bounded together forming thus an unitary experience). In this Thesis, I am interested in emotional episodes rather than in emotional states. Emotional episodes are what Prinz calls 'state emotions' (Prinz, 2004, p. 179-182), and what Deonna and Teroni call 'episodes' (Deonna & Teroni, 2012, p. 8). Emotional states are close to what Prinz refers to as 'non-occurrent attitudinal emotions' (Prinz, 2004, p. 179-182), and to what Deonna and Teroni call 'emotional dispositions' (Deonna & Teroni, 2012, p. 8-9).

It is also important to have in mind the distinction between *emotion generation* and *emotion regulation*. Roughly, the latter phenomenon consists in changing an emotion once it has already been formed; while the former phenomenon consists in forming an emotion in the first place. The *generation* of emotion episodes is the focus of this thesis.

The distinction between *generating* and *regulating* a mental state can be naturally applied to all kinds of affective states, including affect itself. That is, there is affect regulation, emotion regulation, drive regulation, mood regulation, etc. At least from a theoretical point of view, there seems to be a functional difference between forming a certain mental state and changing such a state once it has already being formed. Think of moods. It looks obvious that one thing is a depressed mood arising in you (mood generation), and another thing is to attend to therapy so as to change that state and feel better (mood regulation). However, note that, interestingly, the distinction in question seems to blur in the case of drives or homeostatic motivations. Think of hunger. Hunger does not simply consist in the detection of a lack of nutrients and the negative feeling that results out of that. As we saw above, hunger also includes the urge to change that state as a constituent part. In fact, we call 'hunger' precisely to the urge to get some food so as to change the state of hunger itself. In other words, the episode of hunger itself seems to arise as we are urged, and try to modify the episode itself. Homeostatic motivations seem to be generated as the attempt to regulate them takes place. Here the distinction between generation and regulation is blurred.

This insight is key for my purposes. In this Thesis, I suggest that emotional episodes amount to specific action strategies for regulating affect, in the same way that it can be held that a component part of what it is for a mental state to be an episode of hunger amounts to specific ways of regulating a certain negative affective state (i.e., the urge to eat so as to change the bad feeling of an empty stomach). So the view I suggest in the concluding Chapter of this Thesis is that *generating* an emotional episode consists in *regulating* affect (i.e., valence) in a distinctive sort of way: via emotion-specific knowledge of 'sensorimotor contingencies'. This sort of action-oriented

representations are engaged so as to minimize high-level interoceptive prediction error ("if I act in this sort of way, interoceptive signals should evolve in such-and-such a way"). This claim should not be confused with the claim that emotion generation amounts to emotion regulation, since the claim is that what is being regulated during an emotion generation episode is *affect*, not an emotion *per se*[7]. In the view suggested in the concluding Chapter of this Thesis, emotion generation consists in affect regulation; while emotion regulation would consists in engaging a control system to prevent the unfolding of the emotion-specific, action-oriented representations used so as to control valence (i.e., a "don't activate those representations" system). Emotions amount to valence plus emotion-specific knowledge of sensorimotor contingencies.

1.3. (Competing) Types of approaches to emotion generation and desiderata

As we saw, the explanatory target of an emotion theory amounts to the generation of emotion episodes, in the more circumscribed sense in which I characterized emotion episodes *per se* above. In what follows, I briefly present the main types of naturalistic approaches to the generation of emotion episodes. I organize these approaches in three general or global categories, namely, *perceptual theories*, *cognitive theories*, and *action theories* (and also hybrid theories). There are certainly other ways in which the different types of approaches to emotion theories can be classified. However, I prefer this way of classification scheme mainly for the following reasons.

In the first place, this classification scheme maintains the spirit of previous classifications (e.g., Deigh, 2010; Prinz, 2004; Scarantino, 2014), in the sense that it captures the main aspects typically considered to be key during emotion generation. Emotions are typically considered to be complex affective states. For example, during jealousy, we typically feel in some kind of way, i.e., we *perceive* our bodies changing

[7] Then, the proposed view differs from currents attempts to dissolve the distinction between emotion generation and emotion regulation (see Gross & Barrett, 2011).

in some manner. We also seem to *think* and *reason* in certain kinds of ways, e.g., coming to believe that someone is threatening a (possible) valued romantic partner. During jealousy we are also moved to *act* in a certain manner, for example, aggressively letting know the third party that one has a special interest in the person in question.

In the second place, this classification scheme achieves to directly relate to the three distinct global functions (or "boxes") of the traditional way of conceiving the more general architecture of the mind. In such a conception, we have three separate kinds of systems, namely, perceptual systems (input systems), cognitive systems (central systems), and action systems (output systems). Thus, the three above mentioned more general aspects of emotion fit the three separate "boxes" of the traditional conception of the architecture of the mind. The three distinct kinds of approaches to the generation of emotion episodes understand then emotions as being primarily dependent on one of such three supposedly distinct, separate "boxes". Interestingly, as it has being emphasized by philosophers working under the predictive processing framework (see Clark, 2016; Hohwy, 2013), these more global functions can be seen as operating under the *same* principle of prediction error minimization (Chapter 2). As we will see (Chapter 2), in the predictive processing framework, cognition seems to dissolve into perception, and action results from an inferential, 'belief-like' process that operates under the same principle than perception, even though action still exhibits a functional difference with perception relative to their direction of fit. Thus, insofar as it dissolves the traditional three "boxes" into the same inferential principle of prediction error minimization, a predictive processing view on the nature of emotion has the potential of integrating the insights of the three distinct kinds of approaches to the generation of emotion episodes. Thus, this classification scheme also serves the purpose of highlighting this potential advantage of a predictive processing view on the nature of emotion.

To sum up this point, one of the reasons for preferring to classify the general approaches to emotion in the global categories of perceptual theories, cognitive

theories, and action theories (and also hybrid theories), is that it highlights the traditional distinct, separate "boxes" that configure the traditional tripartite 'division of labour' of the mind, together with capturing salient aspects of emotion. Insofar as the predictive processing framework dissolves such a division to an important extent, a predictive processing view of emotion is then particularly well positioned to integrate the insights of these three approaches, while avoiding their typical problems. I will suggest such a view in Chapter 7, by expanding on Seth's interoceptive inference view of emotion.

Now, in order to make patent how the more plausible view on emotion that emerges out of the predictive processing framework (Chapter 7) integrates the insights and avoid the typical problems of the above mentioned approaches to emotion, in presenting the latter I show how they succeed and fail in explaining key aspects of emotion. That is, I will briefly discuss those aspects in which competing approaches to emotion turn out to be unsatisfactory, and those aspect in which they are satisfactory. This will give us a sense of those aspects that should be taken into account when developing a satisfactory view on the nature of emotion. As I will argue, the expansion of Seth's view suggested in Chapter 7 turns out to be satisfactory in that sense.

Let me make one last point before introducing the main three general approaches to emotion theories. Typical episodes of emotion generation and their regulation are complex, most of the time involving many aspects. For example, certain movements, facial expressions, thoughts, motivations, concomitant perceptual experiences in different modalities, action tendencies, etc. However, the goal of an emotion theory is to single out the minimal essential components of emotion, and to determine in which way these components interact so as to generate an emotion. The goal of an emotion theory is *not* then to determine which aspects *typically contribute causally* to the generation of an emotion. In other words, the fact that emotions typically involve many aspects does not mean that all of them, even though they interact, are proper parts of the mechanism that constitutes emotion. That is why the different general approaches to emotion theories that I will present below hold that *one* of the aspects typically

involved during emotion (perceptions/feelings, thoughts, or action tendencies) has primacy in respect to the constitution of emotion. In doing so, they relegate the other aspects typically involved during emotion to mere antecedent causes or effects of emotion. This is a wise move. Including too many of the aspects typically involved during emotion as proper constitutive parts of emotion risks obscuring that which is essential to the nature of emotion. An over-encompassing approach risks including aspects that might properly belong not to the generation of emotion, but rather to the *regulation* or *expression* of emotion, or even to the background preconditions that support the emergence of emotion rather than constituting them. In other words, it is desirable for a philosophical theory of emotion not to put too much meat into the oven, to put it this way. Let's get down to business.

1.3.1. Perceptual theories

Common sense tends to regard emotions as feelings. Going through an episode of fear, our intuitions indicate, consists in feeling in some sort of way. Now, leaving common sense intuitions behind, it will be assumed in this Thesis that phenomenal consciousness is populated by nothing over and above percepts; there are no phenomenal qualities beyond sensory, perceptual representations. This assumption simply denies that there can be non-sensory representational vehicles with qualitative character (see Prinz, 2012). If emotions are feelings, and considering that phenomenal consciousness (feelings) is populated by nothing over and above percepts, emotions must be perceptions of some sort. Our common sense intuitions indicate that that which is felt during an emotional episode is our bodies reacting in some way. More precisely, our "guts" reacting. That is, that which is felt during an emotional episode is our physiology changing in some way: heart rate changes, sweating, temperature changes, blood concentrating in different part of the body, our viscera reacting, etc. Thus, emotional experience is the conscious perception of our bodies reacting in some sort of way. This is precisely what the, so-called, *feeling theory of emotion* defends (James, 1884). Considering that phenomenal consciousness (feelings) is populated by nothing

over and above percepts, the feeling theory of emotion counts as a perceptual theory of emotion.

In the naturalistic tradition in the philosophy of emotion, *perceptual theories of emotion* hold that emotions are constituted by the perception of bodily changes. Interestingly, the claim is *not* that emotions cause the bodily changes that are then perceived. To borrow the classical example, it is not that sadness makes us cry, as folk psychology has it. In this respect, the perceptual theory of emotion reverses the order of causation assumed by folk wisdom. The claim is that an emotion is constituted as the bodily changes which have being triggered by some external event are perceived.

Naturalistic perceptual theories of emotion, in the sense here considered, are also called *pure somatic theories*, to borrow Prinz's label (Prinz, 2004). Pure somatic theories, as nicely characterized by Barlassina and Newen (2014), hold that "(i) bodily changes play a causal contribution in emotion generation, and (ii) emotions are entirely constituted by the perception of such bodily changes." (Barlassina & Newen, 2014, p. 640).

As I mentioned above, the feeling theory of emotion (James, 1884) counts as a perceptual theory. More precisely, the feeling theory counts as a pure somatic theory. In the feeling theory of emotion, the conscious perception of certain emotion-specific patterns of bodily changes entirely constitute the emotion. However, the feeling theory understands the perception of emotion-specific bodily changes as the *conscious* perception of such changes. After all, this is what makes it a *feeling* theory.

Among other difficulties, the feeling theory faces a decisive problem. Emotions can occur outside consciousness (see also, Ledoux, 1996; Winkielman & Berridge, 2004; Winkielman et al., 2005; for discussion see Deonna & Teroni, 2012, pp. 16-18). If emotions are perceptions, and percepts can occur outside phenomenal consciousness,

as priming and subliminal-perception studies show (e.g., Naccache & Dehaene, 2001; Winkielman et al., 2005; Wen et al., 2007), then emotions should be able to occur outside consciousness.

We can then extract the following *desideratum*:

> A theory of emotion must explain the fact that emotions can occur outside consciousness.

Damasio's version of the pure somatic theory does not face this problem (Damasio, 1994, 2003). According to Damasio, emotions amount to perceptions of bodily changes. However, contrary to James's view, such perception needs not to be conscious. As in any other sensory modality, bodily perception can also occur outside consciousness. On the other hand, Damasio emphasizes that, also as in any other sensory modality, perceptual representation of the body can be activated without actual bodily changes taking place (Damasio refers to these as 'as-if' bodily changes). Just as visual representations can take place without actual retinal stimulation by light-reflecting objects, bodily representations can take place without the actual stimulation of interoceptors by physiological changes.

Even though this version of the pure somatic theory avoids the problem faced by the feeling theory, it fails to account for a key aspect of emotion. As I commented above (Section 1.2.), emotions are meaningful. Emotions have the function of informing the precise way in which the environment/organism relation is significant for the organism's well-being (Lazarus, 1991; Prinz, 2004). Emotions represent such a relation. Emotions represent *core relational themes* (CRT) (Lazarus, 1991). CRT can be seen as the external conditions that each emotion type has the function of discriminating (Prinz, 2004). For example, the external condition every single instance of anger discriminates consists in *a demeaning* offense; while the external condition

every single instance of fear discriminates consist in *danger* (Lazarus, 1991). In a word, a CRT is a property of the particular eliciting event that prompted an emotion, in virtue of which the mental state that takes place amounts to a certain specific emotion type and not another emotion type. Although they amount to relational properties, CRT are *external* events (Prinz, 2004, 2007). The problem for Damasio's account is that it is hard to see how bodily perceptions *simpliciter* could account for the fact that emotions represent external events. How is it that perception of the body could be about *dangers, losses,* and *demeaning offenses*? The body, without further add-ons, simply does not seem to be meaningful in the required sense.

We can then extract the following *desideratum*:

> A theory of emotion must explain how emotions are meaningful. That is, it must explain how emotions can represent core relational themes.

Prinz's version of the pure somatic theory offers a way out of this problem (Prinz, 2004). In Prinz's view, if certain conditions obtain, bodily perceptions can represent core relational themes (CRT). Let me briefly explain. Relying in naturalistic theories of content, Prinz holds that perceptions of bodily changes get to represent CRT just as any mental item gets to represent something in the world. According to the naturalistic theories of content that Prinz considers (e.g., Dretske, 1981, 1986), a mental state *C* represents *c* in the world in virtue of the fact that *c* causally co-varies with *C* in a reliable way, and has the function (by evolutionary or learning history) of causally co-varying in that way. Thus, if a certain specific pattern of bodily, physiological changes *A* systematically co-varies with a certain CRT *a*, and has the function of doing so, then *A* represents *a*. Prinz holds that, given that CRT relative to, so-called, 'basic emotions' causally co-vary with certain emotion-specific patterns physiological changes (and have the function of doing so), the latter represent the former. Thus, in Prinz's view, emotions get to represent CRT by the registering physiological states that causally co-vary with latter.

For example, in this view, an episode of sadness takes place in the following way. In the external environment something valuable is lost, let's say, a soft and warm hoodie. The closet without this hoodie is exteroceptively perceived. Such exteroceptive state triggers several physiological changes. These physiological changes are registered by an interoceptive perception. This interoceptive perception results directly from the physiological modifications in question. However, such interoceptive state is indirectly caused by the *irrevocable loss* that initially triggered the emotional episode. The interoceptive state that constitutes sadness carries information about *an irrevocable loss* by being sensitive to certain physiological changes. Sadness is the perception of those physiological changes. Therefore, this view implies that emotions are individuated both by the specific pattern of physiological states each of them involves, and by the CRT with which those physiological states causally co-vary (and have the function of doing so).

Then, note that emotions can be taken to have two kinds of content. In this respect, Prinz introduces the distinction between the *real* and *nominal* contents of emotion. Real contents are external things, namely, the CRT that indirectly cause emotions; while nominal contents amount to the physiological changes that cause emotions. Physiological changes are *proximal causes*, while CRT are *distal causes*. Interoceptive percepts have then both, physiological causes and its associated external causes. The interoceptive percepts that constitute emotion inform about real contents by registering nominal contents, that is, via patterns of physiological changes.

Thus, this view implies that for each, so called, 'basic emotion' there must be a specific pattern of physiological changes (Prinz, 2004, pp. 72-74). In other words, every 'basic emotion' has its own physiological signature, which represents its own particular CRT.

The problem with this view is that, as I will discuss in Section 4.1., there are no patterns of bodily changes able to individuate emotions (Barret, 2006b; Quigley & Barrett, 2014). There are no systematic causal regularities linking CRT and patterns of

physiological changes. In other words, there are no emotions configured in the physiological landscape. Thus, emotions do not arise by simply perceiving bodily changes. Certainly, if there were such kind of causal regularities linking individuating patterns of physiological changes and CRT, Prinz's view would be quite compelling. However, as we will see in Section 4.1., evidence to that effect is not compelling.

We can then extract then the following *desideratum*:

> A theory of emotion must explain how emotions can arise without there being emotions configured in the physiological landscape, taking into account the fact that there typically is a bodily aspect to the phenomenology of emotion, and that phenomenal consciousness is populated by nothing over and above percepts.

1.3.2. Cognitive theories

As we saw, emotions are directed at external events, such as for example, the dangerousness of a dark dead-end street, or the offense of a thief stealing your bike. Typically, when an individual determines that external events like those obtain, the emotions of fear and anger take place, respectively. In other words, when someone thinks of a situation in a certain way instead of thinking about that situation in another way, different emotions take place. For example, someone might think of a stock-market crash as a loss, but someone else might evaluate that the stock-market crash is actually one more step towards her goal of world collapse. This kind of case shows that different emotions ensue as one thinks of a certain situation in different ways. Emotional episodes typically involve thoughts and evaluations of this sort. This has led some researchers to conclude that cold cognition is essential for emotional episodes to arise. That is, they defend a *cognitive theory* of emotion.

As I will understand it in this work, a cognitive theory of emotion holds that emotions are identical to certain kinds of thoughts. This is what Prinz has called a *pure cognitive theory* (Prinz, 2004). A paradigmatic pure cognitive theory is the *judgment view* of emotion (e.g., Solomon, 1976, 2003; Nussbaum, 2001). The latter can be characterized as the view that a certain emotion *E* is identical to the judgment that the core relational theme characteristic of *E* obtains. For example, anger is constituted by the judgment that a demeaning offense has taken place, and fear is constituted by the judgment that the faced situation is dangerous. Judgments about core relational themes constitute emotion.

The claim that emotions amount to beliefs or judgments captures well the intentionality of emotion: emotions are *about* core relational themes. Concepts are the quintessentially intentional states. Insofar as beliefs or judgments about dangerousness and offensiveness are constituted by the concepts DANGER and OFFENSE, the emotions of fear and anger, respectively, inherit their intentionality from the judgments or beliefs that constitute them. Thus, cognitivism straightforwardly accounts for the intentionality of emotions. However, the judgment view has difficulties in accounting for the fact that emotions are also motivational states that impel courses of action, as it is very hard to see how judgments or beliefs by themselves could have the required motivational force.

Note that the perceptual view also struggles in accounting for the motivational power of emotions. Perception, insofar as it has mind-to-world direction of fit, informs us about events in the world, rather than telling us what to do in order to modify our position in the world. That is why, for example, Prinz's account needs to add a further component into emotion, besides bodily perception, so as to account for the motivational force of emotion. As we will see in more detail in Chapter 5, for Prinz valence is such a component. According to Prinz (2004, 2010), valence amounts to a non-sensory (and non-conceptual) signal that commands the cessation or continuation

of the bodily changes that take place during emotion. Thus, emotions impel course of action in the world only indirectly, i.e., by impelling to change inner states[8].

We can extract then the following *desideratum*:

> A theory of emotion must account for the fact that emotions have motivational force.

There is another kind of view that it is usually classified as a cognitive theory, since it holds that a certain emotion counts as the emotion it is in virtue of the fact that it essentially involves certain cognitive states. That is, states which are composed of concepts: high-level mental states that have relatively abstract contents. Following Scarantino (2014), the view in question can be called *Belief and Desire (B&D) Cognitivism* (e.g., Green, 1992; Gordon, 1987; Marks 1982). This view holds that emotions are constituted by specific set of beliefs and desires. For example, having anger consists in the belief that a demeaning offense is taking place, and the desire to modify the conditions of the situation.

This sort of view appears to do better in accounting for the motivational aspect of emotion, since desires (combined with relevant beliefs) are typically taken to have motivational power—though see Schroeder (2004); also see Scarantino (2014) for an argument to the effect that this view fails to account for the motivational aspect characteristic of emotion. However, there appears to be no way in which this view could plausibly account for a non-negotiable aspect of emotion, namely, the bodily phenomenology of emotion. Beliefs and desires by themselves simply lack the required phenomenology, considering that, as I commented above, phenomenal

[8] See Scarantino (2014) for a discussion on why this indirectness fails to account for the motivational aspect of emotion.

consciousness is populated by nothing over and above percepts, as I am assuming in this Thesis (for discussion, see, e.g., Prinz, 2012, pp.149-168).

We can then extract then the following *desideratum*:

> A theory of emotion must account for the fact that emotions are intrinsically bodily felt.

1.3.3. Action theories

As one undergoes an emotional episode, one typically has the urge to act in some sort of way. This is a deeply rooted common sense intuition about what does it take to have an emotion. People punch other people in bursts of anger, people run away in fright, hold hands out of love, and smile in pride. Sometimes these urges impel us to do little or nothing at all—which are also things we do. For example, mostly in social situations, in the middle of an episode of fear or anger, people tend to just stay there as if nothing had happened. Also some negative emotions which are low in terms of "activation", such as sadness, tend to motivate the disengagement of overt behaviour, rather than promoting the active modification of relevant aspects of the environment.

Common sense tell us that as one undergoes an emotional episode, one also typically engages in mental actions. For example, during an episode of indignation triggered by a governmental moral transgression, one can begin to imagine oneself taking control of the government residency so as to slap the prime minister in the face (Frijda, 1986).

This common sense observation that emotions are closely tied to action has led some researchers to regard this aspect as the primary aspect of emotion, at the expense of its perceptual (felt) and cognitive aspects. They propose what might be called an *action*

theory of emotion, or an *agential theory*. Action theories hold that emotions are identical to action (inaction) tendencies of a certain sort.

This primacy of the agentive aspect of emotion has been defended throughout the history of emotion research. Dewey (1895) held that emotions inherently involve a readiness to act in context-sensitive ways, rather than feeling in some sort of way. Behaviourists (including philosophical behaviourists) also stressed the close link between emotion and behaviour. For example, Ryle (1949) held that emotion terms are used by people to designate 'liabilities' to do certain things; rather than to designate inner states, such as feelings or thoughts. That is, emotion words refer to the likelihood that someone will act in a particular way, the way characteristic of the relevant emotion type. For Ryle, such 'liabilities' consist then in dispositions to behave in an emotion-specific manner. Along similar lines, Skinner (1953) held that emotions are nothing but probabilities of behaviour given certain circumstances. In this view, jealousy consists, let's say, in the increased probability of attacking physically or verbally the third-party, and insulting the partner and then immediately asking for forgiveness, among other behaviours characteristic of jealousy. Watson (1919) goes even further in its behaviouristic approach by claiming that emotions are overt behavioural responses to reinforcing stimuli. More recently, the agential aspect of emotion has also been emphasized in current neuropsychological research. For example, Panksepp (1998) argued that 'basic emotions' are hard-wired inner states that enable behavioural strategies which are inherited from our ancestors. A key theoretical antecedent of all these views are Darwin's insights on emotion. Darwin (1859) famously held that different emotion types involve distinctive expressive behaviours. As Barrett (2006a) remarks:

> "Darwin (1859/1965) argued that emotion categories are distinguished by expressive behaviors. Researchers who seek to define emotion in terms of species-general aspects have embraced the idea that distinct behaviors occur in the service of distinct emotional states. Because humans share some of their neural circuitry with other animal species (be they primates or rodents), researchers assume that it makes sense to define emotion by what these species all have in common: emotional behavior. And the often-used assumption is that there is one behavior for each putative emotion circuit." (Barrett, 2006a, p. 42)

Then, these early agential theories of emotion, in some way or another, tend to maintain that each emotion type has a characteristic set of instrumental behaviours. This assumption is quite problematic. Empirical evidence shows that there are simply no specific set of behaviours for specific types of emotion (see, e.g., Barrett, 2006a). Different sets of instrumental behaviour are involved in the same emotion type, and the same type of behaviour is involved in different types of emotion. Such assumption is even problematic from a common sense point of view. An episode of fear might involve running away, but it also pretending to be cool, changing your clothes, freezing, lighting a cigarette, closing the window, etc. During anger, one might punch someone, but also lighting a cigarette, close a window, yell someone, whispering something, etc.

We can then extract then the following *desideratum*:

> A theory of emotion must explain the close connection between emotion and action, without identifying emotion types with specific sets of instrumental behaviours.

The most endorsed agential theory of emotion is due to Frijda (1986, 2010). Importantly, this version of the agential theory avoids the above problem. According to Frijda, emotions amount to types of *action tendencies*. The latter can be understood as the readiness to act in a certain manner, in such a way that the actions in question take control precedence over other processes (see Scarantino 2014). Interestingly, the kind of actions that Frijda has in mind are defined at a rather abstract level of granularity. This is what allows Frijda's view to avoid the problem mentioned above. The kind of actions that Frijda has in mind are described at the level in which the goal that these actions aim at are defined by the peculiar type of relationship with the world with which each emotion, for being the emotion that it is, needs to deal. For example, the action tendency of anger consists in "a tendency to regain control or freedom of action—generally to remove obstruction" (Frijda, 1986, p. 88); while the action

tendency of fear consists in making oneself inaccessible to the relevant stimulus so as to avoid it. Action tendencies are goal-states relative to which the agent has the motivation to accomplish. That is, action tendencies do not consist in the readiness to execute specific sort of behaviours. Specific sorts of behaviours respond to situational factors rather than to the type of emotion in question. There are many ways to act so as to regain control or freedom of action to remove an obstruction (the action tendency of anger): punch someone, but also lighting a cigarette, close a window, yell someone, whispering something, and so on. The action tendency of fear is making oneself inaccessible to the relevant stimulus so as to avoid it. This can be achieved by executing actions that overlap with the sort of behaviours that could be involved in the action tendency characteristic of another emotion type. For example, running away, pretending to be cool, crying, changing your clothes, freezing, lighting a cigarette, closing the window, etc. In a word, Frijda's view avoids the above problem by individuating emotions by action tendencies which engage with the world in a peculiar fashion, characterized at a rather abstract level of granularity, rather than by specific behaviours.

The problem with this view must be patent by now. Remember that emotions are meaningful, in the sense that they are *about* core relational themes (CRT). In this sense, emotions *inform* about CRT. Emotions have then a mind-to-world direction of fit. In Frijda's view, emotions are identified with action tendencies, which as such have a world-to-mind direction of fit. They inform then about nothing, but rather they command how the world should be. It is very hard to see how emotions, in Frijda's account, can then get to be about CRT. If fear amounts to the action tendency to make oneself inaccessible to the relevant stimulus so as to avoid it, how is it then that fear gets to be about *danger*? It is difficult to see how the action tendency view could explain the intentionality of emotion.

Scarantino (2014) has offered an agential view of emotion, in line with Frijda's view, that is able to deal with this problem, namely, *the motivational theory of emotion* (MTE). Roughly, MTE is the view that:

> An emotion is a prioritizing action control system, expressed either by (in)action tendencies with control precedence or by action reflexes, with the function of achieving a certain relational goal while correlating with a certain core relational theme. (Scarantino, 2014, p. 178)

MTE refines that which is already in place in Frijda's account, but it also expands the latter so as to show how an agential view of emotion can deal with the fact that emotions are intentional states. MTE refines Frijda's account by proposing a two-level control structure that allows to account for some motivational phenomena involved in emotion, which other accounts fails to explain. However, at this juncture, the more relevant point on which to focus is how MTE expands Frijda's agential theory so as to account for the intentionality of emotion. MTE accounts for the latter phenomenon by embracing a teleosemantic view of content. Roughly, according to the latter, a mental state *M* represent a certain event *E* in virtue of the fact that *M* has the function of being triggered by *E*. Emotions, Scarantino claims, have the function of being triggered by CRT. However, emotions, being action tendencies, they also have the function of achieving the goal that define action tendencies. That is, they also have a motivational function.

> "According to MTE, what explains the intentionality of emotions is that they are (in) action tendencies or action reflexes with the informational-cum-motivational function of *achieving relational goals while correlating with core relational themes*. On this view, fear is about dangers because it is a prioritized avoidance tendency/reflex with the informational-cum-motivational function of achieving the relational goal of one's own safety while correlating with dangers." (Scarantino, 2014, p. 178)

That is, in the case of fear, the past effects of achieving the goal that defines the action tendency of fear, namely, achieving one's own safety, in the presence of danger, explains why fear was selected for.

I think that this view is generally on track. However, let me point to a few potential problems of it, which a PP version of the agential theory might be better positioned to

avoid. In the first place, MTE is silent about why emotions have the motivational force they have. In the second place, remember that there are differentially valenced emotions. That is, some emotions feel good, while other emotions feel bad. This polarity is constitutive of emotion. MTE offers no clue as to what is it that makes certain emotions positive and other emotions negative. Without a view on the nature of valence motivated within the framework of MTE, the latter, as it stands, seems to lack a way of accounting for the polarity in question. That said, as I mentioned above, I think that MTE is, in general, on the right track. The PP view on emotion to be suggested in Chapter 7 draws significantly on MTE, while integrating the insights of the other families of positions in emotion research. Very roughly, I will suggest that emotions amount to action strategies for regulating valence. This, in order to minimize high-level interoceptive PE, i.e., the discrepancy between the expectation of stable homeostasis and the homeostatic state about which the interoceptive percept that constitutes valence informs. I think that this view satisfies desiderata, and avoids the problems exhibited by the types of views presented in this Section.

1.3.4. Hybrid theories

There are few types of theories that instead of emphasizing only one of the aspects commonly associated with emotion, they incorporate more than one aspect into the mechanism that constitutes emotion. Let me briefly present them, as the view on the nature of emotion that I will suggest in Chapter 7 can be taken to be part of this family, although the view I will suggest also emphasizes the agential aspect of emotion, as the action theories just presented. Now, some approaches to emotion tend to include all or almost all of the aspects commonly associated with emotion. However, in this Chapter, I prefer not to include these approaches. I do not include these approaches, because, as I mentioned above, it is desirable for a philosophical theory of emotion not to put too much meat into the oven. That is, the goal of a philosophical theory of emotion is not to list all the aspects that typically contribute in emotion-related phenomena, and show how they are mutually relevant to each other. The goal of an emotion theory is to single out the minimal essential components of emotion generation.

1.3.4.1. Appraisal theories

Appraisal theories are usually regarded as cognitive theories, since in these theories cognition is the driving force of emotion generation (e.g., Arnold, 1960; Lazarus, 1991; Scherer, 1984). However, appraisal theories do not identify emotions with cognition, while including affective states. According to appraisal theories, emotions arise as a certain kind of thought (i.e., appraisals) trigger some sort of affective state, either physiological changes, feelings, etc. Emotions are not identified with appraisals, but rather appraisals cause emotions. Even more, appraisals individuate emotions. That is, the affective states which are caused by appraisals count as a certain emotion in virtue of being caused by them.

What then are appraisals? Appraisals consist in evaluative inferences or judgments about the way in which a certain situation is relevant for the goals and interests of the organism. These evaluative inferences are usually considered to be hierarchically organised in different dimensions. Each dimension reflects different aspect of the organism-environment relation. For example, appraisal theorists generally agree that among such dimensions are the following. Firstly, the organism needs to determine the relevance of the situation, and whether the situation is compatible with her goals. Secondly, the organism needs to infer who is responsible for the relevant situation. Finally, the organism needs to determine the extent to which she has the resources to cope with the situation. Depending on what the organism judges in those respects, a certain type of emotion, instead of another type of emotion, takes place. For example, anger results in case the organism judges that the situation is incompatible with her goals, that someone else is responsible for bringing about the relevant situation, and that she is able to cope with the situation.

Insofar as in this kind of view cognition is in the driver seat in respect to emotion generation, appraisal theories straightforwardly account for the intentionality of emotion. However, note that in appraisal theories, cognition is not part of the

mechanism itself that constitutes emotion. In appraisal theories, cognition is a necessary *cause* of emotion. That is why this kind of view is usually called etiological cognitivism. In this kind of view, given that appraisals lie outside the mechanism of emotion, it fails to show how is it that emotions themselves can be meaningful. That is, if none of the components of the mechanism that constitute emotion account for the intentionality of emotion, then it seems that, in this view, emotions are not *themselves* about CRT. The fact that emotions themselves are meaningful, i.e., that emotions themselves are about core relational themes, is a non-negotiable aspect for a philosophical account of emotion (see, e.g., Prinz, 2004, Deigh, 2010). Thus, this point puts philosophical pressure on standard appraisal views.

1.3.4.2. Two-factor theories

Two-factor theories (Schachter & Singer, 1962) are also usually classified as a cognitive theory, since in these theories cognition is the driving force of emotion generation. However, two-factor theories do not qualify as a pure cognitive theory, because besides including a cognitive component in the mechanism that constitutes emotion, these views also include a felt bodily component (e.g., arousal). According to Schachter and Singer (1962), emotions amount to cognitive interpretations of the physiological changes that constitute arousal. In this view, once the latter is triggered by an external situation, high-level abstract knowledge is brought forth in order to label such an arousal state with an emotion term. Or to put it this way, cognition is brought to categorize the arousal state triggered by the external event via emotion concepts. Thus, emotions exhibit two-factors: bodily perception (arousal) plus cognition.

Contrary to appraisal views, the two-factor approach includes the cognitive aspect as part of the mechanism that constitutes emotion. Thus, in this view, the intentionality of emotion is accounted for by something that is part of emotion itself. Moreover, besides accounting for the intentionality of emotion, this view also accounts for the felt bodily aspect of emotion: that which is labelled or categorized are precisely the

felt bodily changes that, according to Schachter and Singer, constitute arousal. However, this view struggles in accounting for the agential aspect of emotion. Labels do not mandate courses of action, and arousal by itself is action-neutral. Two-factor views are silent as to how emotions have the motivational power they have.

I do not take the above observations to be conclusive arguments against the families of positions presented in this Section (Section 1.3.). The aim of this Section is simply to provide an overview of the main families of emotion theories. This, in order to highlight those aspects in which they are standardly considered to be problematic, and those aspect in which they result satisfactory. Now, the predictive processing framework is beginning to provide a unifying account of *all* the seemingly disparate variety of mental phenomena. As I mentioned above, in the predictive processing framework, cognition seems to dissolve into perception, and action results from an inferential, 'belief-like' process that operates under the same principle than perception—even attention emerges naturally from the inferential perceptual machinery posited by PP (Chapter 2). Thus, insofar as it dissolves the traditional three "boxes" into the same inferential principle of prediction error minimization, a predictive processing view on the nature of emotion has the potential of integrating the insights of the main families of emotion theories. However, the predictive processing framework exhibits some issues when it comes to accounting for emotion (Chapter 4). In the concluding Chapter of this Thesis (Chapter 7), I suggest a way in which the current PP view of emotion can be amended in such a way so as to become a promising view, which compares favourably to the main families of positions in emotion research. It compares favourably since, insofar as PP is a unifying, integrative framework, it has the resources to incorporate the advantages and avoid the typical problems of perceptual, cognitive, and agential views.

To sum up, in this Chapter we saw that valence, rather than arousal, stands out as the paramount defining construct in affect. The explanatory target of this Thesis amounts to one specific kinds of affective state, namely, emotions *per se*. Then I discussed some of the main kinds of general views on how emotion episodes are generated. I classified

these approaches to emotion in the global categories of perceptual theories, cognitive theories, and action theories (and also hybrid theories). I briefly discussed those aspects in which each of these global approaches to emotion turn out to be unsatisfactory, and those aspect in which they result satisfactory. This will give us a sense of those aspects that should be taken into account when developing a satisfactory view on the nature of emotion. As I will argue, the expansion of Seth's view suggested in Chapter 7 turns out to be satisfactory in that sense. A predictive processing view of emotion has the resources to integrate the insights of and avoid the problems of the different approaches to emotion. In the coming Chapter, I present the basics of the workings of the predictive processing machinery. I emphasize how the principles of predictive processing apply to vision, as Seth's view on the nature of emotion is based on a direct analogy between vision and emotion.

2. The basics of the predictive processing framework

The *interoceptive inference view of emotion* (IIE) (Seth, 2013, 2015a; Seth et al, 2012; Seth & Critchley, 2013; Hohwy, 2013) is the first attempt of accounting for emotion in line with the principles of predictive processing (PP). IIE accounts for emotion generation by extending, *in a direct fashion*, the process by which visual (exteroceptive) percepts are formed to the case of interoceptive percept formation. Just as in the PP framework visual perception arises via *visual perceptual inference*, Seth holds that the perception of our homeostatic, physiological condition arises via *interoceptive perceptual inference*. In this view, emotions arise from interoceptive predictions of the causes of current interoceptive afferents. Emotions are then interoceptive percepts of a certain sort. Thus, IIE defends a version of the perceptual theory of emotion (Chapter 1).

Considering that IIE holds that emotions arise in direct analogy to the way in which visual (exteroceptive) percepts are formed, in this Chapter, I present the way in which perception operates in the PP framework, with a special emphasis in visual perception (and exteroceptive perception more generally). Importantly, I limit myself to presenting only the more fundamental aspects of the principles of PP. I present also those aspects that are particularly relevant for the argument I develop in this Thesis, but that do not form part of the more typically emphasized aspects of the architecture of the PP machinery—such as the notion of 'knowledge of sensorimotor contingencies'. I also leave the computational and mathematical details aside, as I deal only with the more 'conceptual' side of the framework. I think that presenting only the more conceptual aspects of the principles of the PP machinery allows to pose the challenges of this new framework in a way that converges with philosophical discussion on the nature of emotion. Finally, in this Thesis, I simply assume that the PP framework is mostly correct. In this respect, I am interested in assessing how much the PP framework can do. Particularly, how much it can do for our understanding of emotion (can the predictive processing framework account for emotion?). In order to do this, the framework needs to be assumed to be in place. Thus, I will exclude from

my exposition the motivations of the framework, and the growing evidence that is beginning to accumulate, suggesting that the PP framework is substantially on track. This work has already been done in an outstanding way in recent works by Clark (2016) and Hohwy (2013). I refer the reader to these works for a review of the evidence and the motivations of the framework.

I begin by presenting the PP account of visual (exteroceptive) percept formation (Section 2.1.). Visual percept formation (and exteroceptive percept formation in general) is the most paradigmatic case of the workings of PP. I emphasize in this Section those aspects of PP that make it especially interesting as an account of perceptual processing. Particularly, (a) the role of *precisions* (Section 2.4.); (b) the importance of cross-tall between modalities across all levels of the perceptual hierarchy, and the constant influence of high-level multimodal and amodal knowledge (Section 2.6.); (c) also the role of stored knowledge about 'sensorimotor contingencies' during active inference (Section 2.5.2.); and (d) the fact that, via learning, the cortical hierarchy recapitulates regularities in the world (Section 2.7.).

2.1. The predictive processing framework

Predictive processing (PP) was mainly conceived and developed as an account (and re-conceptualization) of perceptual processes. Roughly, and leaving mathematical and computational details aside, in the PP framework, the traditional approach to perceptual processing is turned upside down. Traditionally, visual perception is considered to be a bottom-up driven process, in which incoming visual signals or features are accumulatively processed from the bottom-up until a coherent visual percept is finally formed. The idea here is that the input that the visual system receives from the world of middle-sized, light-reflecting objects consists in a rich signal that manages to represent such objects and their features. This signal is passively processed in a step-by-step manner, from earlier regions to higher level regions in the brain. Earlier regions process simple features, and as the flow of information is passed to

higher level regions, the analysis of the input becomes more complex. In the sense that several features that where initially processed in an individual manner become merged together in the complex objects that we visually perceive. Thus, in this account, changes in the configuration of the input stimuli drives modifications in perception, as visual perception is driven by the features of the input stimuli that the visual system captures.

However, bottom-up approaches can hardly deal with the ambiguous inputs that we get from the world. Incoming low-level sensory signals underdetermine their external cause. That is, a certain object (or event) can cause different effects in different senses, and it can also cause different effects in the same sense. Moreover, different objects (or events) can cause the same effect in a certain sense modality. In this sense, there is no linear, one-to-one mapping relation between causes and effects. The effects that the objects and events in the world have in our senses when they impinge our early processing regions are typically *ambiguous* between candidate causes. To take a classical example, convex objects give rise to retinal activity which is ambiguous between different kinds of causes. Such retinal activity can be caused by either a convex object that receives light from above, or by a concave object that receives light from below. Consider also a more colloquial case. Certain retinal and early activity can be highly ambiguous between, and compatible with, all the following worldly causes: a dog, a stuffed dog toy, a few pairs of brown shoes arranged in the form of a dog, a fox, a goat, a design chair, etc. Ambiguity in the incoming sensory signal regarding possible worldly causes is a characteristic issue with which perception needs to deal.

2.1.2. Perceptual inference

Considering that low-level sensory signals underdetermine their external cause, and contrary to the bottom-up approach, Bayesian PP accounts of visual perception hold that visual percept formation takes place via visual *perceptual inference*. That is, the

brain forms visual percepts in a top-down fashion by predicting its incoming lower-level sensory signals from higher-level models of the likely (hidden) causes of those visual signals. These models can be seen as putting forward content-specifying hypotheses about the object or event responsible for triggering incoming visual activity.

The models in question can thus be seen as storing probabilistic knowledge that allows them to link external causes and inputs. In order to do the latter, these models specify hypotheses that could be informally characterized in the following way: "if in fact the (hidden) external cause is an *E*, these lower-level effects or sensory signals are most likely to be expected". These hypotheses consider the likelihood that the considered external causes would produce the encountered lower-level sensory effects, given stored knowledge of the priors (i.e., already maintained 'beliefs' about) of such external causes. The selected hypothesis is the one with higher posterior probability.

Importantly, the selected hypothesis determines the content of the experienced percept. That is, the content of the visual hypothesis that best fits the array of incoming visual signals is the hypothesis that inherits its content to the resulting visual percept. This is the hypothesis that, among competing hypotheses, does best at reducing the difference between predicted signals and actual incoming signals (more on this below). So, let's say that the visual system is considering two hypotheses. The hypothesis that a dog is likely to be causing the incoming visual signals, and the hypothesis that a cat is causing such signals. If the dog-hypothesis is better suited than the cat-hypothesis to fit incoming visual activity, then the percept of a dog is formed, instead of the percept of a cat. The visual experience of a dog then arises.

Now, having stored probabilistic knowledge of the expected activity of the relevant lower layers, the models of the world that the brain harbours predict lower-level sensory activity by *generating* those expected states in the relevant lower layers of the perceptual hierarchy. A contentful percept is formed once a selected hypothesis

achieves to generate the sensory activity that successfully matches, and thus supresses, current lower-level sensory signals (see Clark, 2013; Hohwy, 2013).

> "What we have so far is an internal model that generates a hypothesis—we might call it a fantasy—about the expected sensory input. This is the generative model, which has a number of different parameters that together produce the fantasy (hypothesis) down through the hierarchy. A particular fantasy might do a fine job at matching the incoming sensory input, and thus should determine perception." (Hohwy, 2013, p. 54)

Context and background knowledge are important here. Remember that low-level sensory signals underdetermine their external cause. Many external causes could be giving rise to the same low-level sensory effects. Therefore, in order to successfully predict the presence of, for example, a dog in the environment, the perceptual system has to rely on the relative probability of a dog-encountering in the particular kind of situation in which the organism finds itself. Top-down signals, acting as priors extracted from background knowledge, provide this kind of context-fixing information, which is active previous to encountering the dog in the environment (more on this below). To use a very colloquial (but clarifying) example, let's say that there is a dog in the world. The light reflected by the dog triggers visual activity in early regions of processing. Such activity is ambiguous between its causes in the world. Using the same example as above, let's say that the visual system is considering two hypotheses. The hypothesis that a dog is likely to be causing the incoming visual signals, and the hypothesis that a cat is causing such signals. Considering that you find yourself in the house of a dog-lover person, the hypothesis that there is a dog in front of you exhibits higher anterior probability than the cat-hypothesis. To put it colloquially, in this sort of situation, the visual system tells itself, "if it is a dog, instead of a cat, I expect more likely such and such sensory signals, and not these other ones, because dogs, but not cats, tend to cause such signals in contexts like this". It then generates these expected signals in lower levels of the visual hierarchy. Then it asks itself "Am I getting a good match?" As I commented above, if it is getting a good match, then the percept of a dog is formed. If it is not getting a good match, the hypothesis (model) needs to be updated until a better hypothesis about the causes of the incoming signals is put forward (perhaps there was no dog after all).

2.1.3. The perceptual hierarchy

Importantly, in the PP framework, lower levels of the hierarchy encode regularities that operate at fast time-scales, capturing variant aspects of experience. On the other hand, higher levels encode increasingly more complex regularities, which operate at slow time-scales, capturing relatively more invariant aspects of experience. For example, in the case of vision, low levels of the visual hierarchy encode regularities such as the details of edges and the changing contours of objects as one moves (represented in V1), which have small receptive fields. While high levels of the visual hierarchy encode relatively more invariant information, such as those represented in the temporal lobes, which have wider receptive fields, such as the enduring face and body of someone you know. Now, the higher-levels of the hierarchy are not modality specific (i.e., amodal and multimodal), and encode even more abstract, slow time-scale regularities, not directly related to regularities pertaining to the domain of just one modality. For example, regularities such as, for example, that good people keep a promise.

2.1.4. Prediction error

Crucially, as top-down signals produced by the generative models of the world provide predictions of sensory effects, bottom-up signals provide *prediction error* (PE). The latter notion is of most importance. According to PP, *all what the brain does is to minimize its prediction error*. PE consist in the difference between the sensory signals expected (and generated) from the top-down, and the actual, incoming sensory signals.

PE is critical in the PP framework, not only because, from a more global perspective, PP holds that all what the brain does is to minimize its PE. The latter is also crucial in the PP framework since during online processing, PE is used as feedback for updating and improving models or hypotheses and, in the long term, as a learning signal that

improves the predictions of our stored models of the world. Once updated in this way, a generative model can specify more accurate predictions. Then, in the PP framework, the perceptual system is engaged in the task of 'explaining away' or *suppressing* the sensory signals that fit the model predictions. If there is a *match* between the generated predictions and the incoming sensory inputs, the matched information coming from the bottom-up is suppressed. The remaining non-matched information (i.e., prediction error) is allowed then to 'go up' through the perceptual hierarchy. This, for the sake of online-feedback purposes, and, in the long term, learning. Thus, in the PP framework, PE is what really counts as the input to the system.

> "These predictions, as it were, *query* the world and dampen down predicted sensory input. The result is that only prediction error is propagated up through the system in a bottom-up fashion, and aids in revision of the model parameters (so we need to re-label the bottom-up sensory input, associated with the light grey, upwards arrows in Figure 5, 'prediction error'). The functional role of the bottom-up signal from the senses is then to be *feedback* on the internal models of the world" (Hohwy, 2013, p. 47)

In this respect, the idea is that PE signals play the role of shaping stored priors, so that the percepts which are formed by generating expected signals and minimizing PE achieve to successfully represent the world. In this sense, PE plays the role of preventing that organisms perceive whatever that their brains expect. PE functions as an objective corrective signal triggered by regularities in the world. Thus, expectations are controlled by the world via PE. In other words, sensory evidence constrains the kind of expectations that get to be built by the brain.

The task of minimizing PE by 'explaining away' incoming sensory signals is carried through, in a coordinated fashion, at each level of the perceptual hierarchy. Adjacent levels are connected to each other so that the best hypothesis at one level becomes PE for the level above. Thus, the PP framework posit a cascade of information interchange between levels, ranging from the higher levels, which encode relatively more invariant information, to the lower levels, which encode relatively more variant information.

2.1.5. Inferring precisions

Now, the *variability* of PE in different contexts must be taken into account, so as to veridically revise models and to efficiently sample the world, and thus minimize PE. In order to determine whether a certain perceptual hypothesis fits the incoming sensory data well enough, it is required that the system has prior expectations about the variability of the incoming signals, given context. In other words, PE minimization demands the system to determine the context-dependent reliability of PE across all levels of the perceptual hierarchy: the *precision* of incoming signals also needs to be inferred. Otherwise, the perceptual hypotheses that a certain model puts forward could not be satisfactory in guiding adaptive behaviour—i.e., in getting the world right.

> "What determines such inference about variability is which expectations for variability we have in different contexts. This can be put in the model-fitting terms used earlier. Recall that, in an attempt to fit a statistical model to a set of data, the error between the model and the data should be minimized. But if there are no prior expectations about the variability of the data set (the inverse of which is the same as its precision), then there is no way to reasonably decide how much or how little fitting is enough. If this decision is not optimal, then predictions on the basis of the model will not be good." (Hohwy, 2013, p. 65)

In order to minimize PE satisfactorily, the brain needs to extract from the world regularities relative to precisions, so that it can use them later to improve PE minimization across levels. That is, the brain needs to infer from context in which cases PE is likely to be reliable. In other words, inferring precision amounts to a kind of metacognitive process, in which the brain engages in inference about perceptual inference.

Now, precisions determine the relative influence that top-down expectations have relative to the incoming input. Precisions modulate the balance between top-down and bottom-up influences. Thus, in contexts where the input is deemed to be unreliable, top-down hypotheses will be assigned more weight than usual. In these cases, top-

down hypotheses drive almost completely the processes of percept formation, ignoring thus the input to an important extent.

> "The fundamental purpose for precision processing is to enable the activities of internal models of the world to be driven by reliable learning signals: if there is confidence in a signal then it should be allowed to revise the hypothesis and if there is less confidence then it should tend to carry less weight. The assessment of confidence should therefore impact on the strength of the prediction error message being passed up through the system. Precision expectations are thus thought to be realized in systems that regulate the *gain* on prediction error units in the brain (Feldman and Friston 2010). The more precision that is expected the more the gain on the prediction error in question, and the more it gets to influence hypothesis revision. Conversely, if the reliability of the signal is expected to be poor, then the prediction error unit self-inhibits, suppressing prediction error such that its signalling is weighted less in overall processing." (Hohwy, 2013, p. 66)

For example, think of a case in which the situation does not afford to rely on the incoming stream of sensory data. Let's say that you find yourself driving in a foggy day. This is a typical kind of context in which the visual information one receives from the world is quite ambiguous and noisy. Then, as you drive, the visual system reduces the amount of weight given to the stream of visual input, relative to the amount of weight given to the input in normal conditions. Visual perceptual inferences are then driven almost completely by stored expectations, at the expense of what the world has to say via visual PE: you begin transiently seeing car lights in front of you when they are actually lampposts at some distance, or you transiently see a curve when there the pavement is actually simply a bit wet. Illusions then take place.

A more radical example of a poor sensory signal is the effect known as 'sensory deprivation'. Very roughly, during sensory deprivation people are exposed to a major reduction of stimuli. For example, white noise in the auditive case, or diffuse bright or no light at all in the case of vision. After a while, people begin undergoing hallucinations, such as different kinds of sounds—from beepings to conversations and music—and different types of visual images, ranging from coloured patches to faces and flowers.

These sort of more radical examples of an impoverished signal are cases in which, given that a major variability in the signal is to be expected, the gain on PE signals are dampened down to the extent that it is practically completely ignored. This determines that prior expectations are inferred to have an extraordinary high precision. That is, the visual system (or the auditory system) gives to its top-down predictions significantly more weight than it does in normal conditions. Remember that PE plays the role of impeding that one perceives whatever that one expects, by functioning as a corrective signal triggered by, and dependent on, regularities in the world. Expectations are controlled by sensory evidence. Therefore, in the absence of such a corrective signal, the poor sensory signal involved in the example in question gives rise to the hallucinations characteristic of the sensory deprivation effect.

In other contexts, the input will be deemed to be reliable. In these latter cases, PE will be taken to be informative, and it will then be used to guide the current hypothesis-selection process and, in the long term, learning.

> "Bottom-up prediction error is favoured from units that "feed" well-performing prediction units. That is, an evolving prediction error signal that can continually be explained away well by a particular hypothesis can be assumed to be reliable and should thus be weighted more in the message-passing economy." (Hohwy, 2013, p.61)

Interestingly, note that precision-weighting is remarkably functionally similar to attention. In fact, in the PP framework, precision-weighting is identified with attention (Friston, 2009). Precisions determine which aspects of the sensory stream are going to have more or less gain, and this across all levels of the perceptual hierarchy. In other words, precisions determine where, and to what degree, perception focuses.

> "The complex, precision-weighted play between top-down prediction and bottom-up prediction error messaging marks out a particular functional role. Expectations of precision modulate where and how the perceptual system is focused. In particular, it determines which signal is given preponderance and it determines the extent to which there is a worldly focus, rather than an internal, more general, thoughtful or meandering focus. It turns out that this functional role fits *attention* extremely well." (Hohwy, 2013, p. 70)

The PP framework solves then the underdetermination between (hidden) worldly causes and sensory effects (accessible to the brain) by building expectations about sensory inputs. Given that the regular structure of the world brings about sensory inputs which do exhibit repeatable patterns (i.e., it is not just noise), the latter allow the system to build expectations about subsequent sensory activity. Such expectations are compared to the latter sensory activity, and the resulting precision-weighted difference (PE) can be quantified. If there is a satisfactory match, the model from which the prediction was generated is capturing the relevant aspect of the world. If precision-weighted PE is significant, then the parameters of the models are revised (and learning thus takes place).

The inferential predictive machine sketched above, in which one level attempts to predict the activity of the level below so as to minimize precision-weighted PE, is repeated throughout all the several levels of the perceptual hierarchy. In case that, at a certain level of the perceptual hierarchy, priors do not achieve to predict the incoming signal, the resulting PE drives predictions at the level above depending on how much weight is assigned to such PE. If the latter is assigned a significant weight, then predictions at the level above need to "diligently" deal with such resulting PE, to put it this way. However, if the resulting PE is assigned little weight, then model revision takes place at the same level or at levels below. Thus, during perceptual inference, expectations about precisions (attention) determine the way in which model revision takes place across the hierarchy.

2.1.6. Two ways of minimizing prediction error

Crucially, PE can be minimized via two different strategies, which differ in their directions of fit. PE can be minimized via *perceptual inference* and *active inference*.

2.1.6.1. Perceptual inference

Perceptual inference is the strategy of PE minimization described above. As we saw, in the PP framework, visual percepts are formed by minimizing visual PE in a specific manner: via visual *perceptual inference*. Perceptual hypotheses have mind-to-world direction of fit—models or hypotheses are changed so as to attempt to match incoming signals. During perceptual inference, hypotheses about the worldly causes of sensory input are adjusted so as to fit current incoming data.

2.1.6.2. Active inference

The other manner in which PE can be minimized is action. More precisely, PE can also be minimized via *active inference*. The latter consists in changing the environment so as to obtain sensory data that fits considered predictions or hypotheses. Active inference has world-to-mind direction of fit—model's parameter are kept constant and the system selectively samples the world so as to obtain signals that fit predicted signals.

> "It follows trivially that the upshot of the brain's prediction error minimization activities is to increase the mutual information between the mind and the world—to make the states of the brain as predictive as possible of the sensory input caused by events in the world. This account has largely suppressed a very obvious point, namely that the mutual information can also be increased by making the sensory input from the world more predictive of the states of the brain's model, that is by changing the input to fit the model rather than changing the model to fit the input." (Hohwy, 2013, p. 76)

As I commented above, the predictions that a certain model puts forward become more successful as the more they achieve to match the stream of sensory input. In this manner, the brain achieves to fulfil its only task, namely, minimize PE. Considering that modifying sensory input is quite helpful in achieving such a fit, the brain should be constantly acting so as to minimize PE. That is, the brain should be constantly

changing the sensory array in order to fit its expectations about incoming sensory activity: changing the world to fit the mind. Triggering the sensory stream that is in line with prior expectations can be achieved either by moving oneself around so as to change the relative position of the sensory organs, or it can be achieved by altering the sensory input itself. Action is key for minimizing PE.

Now, in order for purposeful behaviour to ensue, the kind of PE signals that needs to be minimized amounts to signals proprietary of the proprioceptive domain. That is, in order to act, the brain needs to build a model of the musculoskeletal system. Moreover, it must learn the sensory consequences that action has in the proprioceptive system itself. Thus, the brain needs to store counterfactual knowledge of the way in which sensory signals would evolve in case the organism were to act in such and such a way. When this counterfactual knowledge is activated, it triggers a large proprioceptive PE signal. The latter consist in the difference between the actual proprioceptive state of the organism and the counterfactually expected ('desired') proprioceptive signals. In order to minimize this, let's say, 'self-generated' PE signal, the body must move in such a way as to bring about the expected proprioceptive states. In other words, the brain brings action about by simply triggering proprioceptive PE, and by minimizing the latter. No motor commands are part of the story.

> "The mechanism for being a system that acts is thus nothing more than the generation of prediction error and the ability to change the body's configuration such that the antecedent of the counterfactual actually obtains and error is suppressed. Action therefore does not come about through some complex computation of motor commands that control the muscles of the body. In simple terms, what happens is instead that the muscles are told to move as long as there is prediction error. The muscles of the body are thus at the mercy of the prediction error generated by the brain's model of the way the world is expected to be like but isn't. Prediction error is then the simple mechanism that controls action." (Hohwy, 2013, p. 82)

As it is typically described, this manner of bringing about action can be seen as some sort of self-fulfilling prophecy. The proprioceptive system prophesies that it will be in certain sensory states, and by simply entertaining this prophesy the organism will find itself in those sensory states by generating proprioceptive PE. However, just as in any

sensory domain, proprioceptive predictions can go wrong. In this case, the musculoskeletal system will not end up in a position in which it gets the predicted proprioceptive signals. This will trigger another round of proprioceptive PE, which needs to be minimize by engaging the reflex arcs that will eventually put the body in the position that the proprioceptive system prophesises.

Expectations of precisions play a major role in this whole story. As we saw, action is brought about by triggering proprioceptive PE. Then, considering that the brain is exclusively devoted to the business of minimizing PE, the proprioceptive system has two options available, namely, either minimize proprioceptive PE via proprioceptive *perceptual inference*, or minimize it via proprioceptive *active inference*. In other words, instead of intervening the proprioceptive landscape so as to fit the counterfactual proprioceptive expectations about sensory input, the brain could simply change its proprioceptive expectations so as to fit the actual incoming proprioceptive signals. In this latter manner, the brain can perfectly well achieve its only task of minimizing PE. However, note that the latter option implies that action does not ensue, but rather a proprioceptive percept of the current state of the musculoskeletal system gets to be formed. In order for movement to occur, expectations of precisions need to be added to the story. Attention to proprioceptive PE plays a key role during action. Action arises as counterfactual proprioceptive input is inferred to have significantly higher precision than actual proprioceptive input. That is, the actual proprioceptive input is ignored. Thus, perceptual inference does not drive proprioceptive PE minimization, and action can then occur.

Crucially, as Seth (2015a) remarks, active inference can take place not only so as to conform to current expectations, as it occurs during motor behaviour by making proprioceptive data fit proprioceptive expectations (by significantly increasing the *precision* of proprioceptive predictions), so that 'desired' (or expected) movement occurs. Active inference can also take place during percept formation in exteroceptive modalities (i.e., during *perceptual inference*), in order to confirm, disconfirm and disambiguate perceptual hypotheses—this is what Seth calls 'epistemic active

inference' (Seth, 2015a). As Seth (2015a) notes, and as mentioned above, this requires storing representations of the counterfactual relations that obtain between (possible) actions and its prospective sensory consequences ("if I act in this manner, sensory signals should evolve in such-and-such way"). That is, *representations of sensorimotor contingencies*. Active inference requires then this kind of counterfactual knowledge.

The idea is that during perceptual inference in non-proprioceptive modalities, agents sample the external world based on a certain perceptual hypothesis about the cause of incoming sensory activity. The hypothesis in question is the hypothesis that, among competing hypotheses, exhibits higher posterior probability. That is, the best current hypothesis. This hypothesis is used to predict the sensory signals that would take place in case the hypothesis under consideration obtains, and certain actions ensue. Agents then carry out these action to determine whether the considered hypothesis actually obtains. Thus, the brain needs to determine which of the hypotheses that exhibit a significant anterior probability seems to best account for current incoming lower-level sensory activity. This hypothesis, which has now a relatively higher posterior probability, is the one used to guide the active sample of the world. These actions serve the purpose of determining whether the newly triggered sensory activity is compatible with the considered hypothesis. This strategy allows the system to reduce the uncertainty about maintained hypotheses. If the expected sensory activity obtains, then the hypothesis in question is confirmed. However, if the expected sensory activity does not obtain, the considered hypothesis is disconfirmed, and another competing hypothesis is 'put to the test', to keep the scientific metaphor.

> "The situation is then this. Perceptual inference allows the system to minimize prediction error and thus favour one hypothesis. On the basis of this hypothesis the system can predict how the sensory input would change, were the hypothesis correct. That is, it can test the veracity of the hypothesis by testing through agency whether the input really changes in the predicted ways. The way to do this is to stop updating the hypothesis for a while, and instead wait for action to make the input to fit the hypothesis. If this fails to happen, then the system must reconsider and eventually adopt a different or revised hypothesis. For example, if the highest posterior goes to the hypothesis that this is a man's face seen in profile, then the system may predict that by moving visual fixation down towards the chin, a sample will be acquired that fits with this hypothesis. If it does, then

> this further enhances the probability that this is a man's face; if it does not fit then the system may have to go back and revise the hypothesis such that it expects the cause of its input to be, say, a woman's or a child's face." (Hohwy, 2013, p. 79)

As counterfactual knowledge of the sensory consequences of action are required across all levels of the cortical hierarchy, representations of 'sensorimotor contingencies' must also be found across all levels—as any other kind of predictive knowledge posited by PP. Thus, representations of 'sensorimotor contingencies' can be low-level or high-level depending on how variant or invariant are the regularities that they encode. For example, low-level, fast-changing regularities include the very fast eye movements that occur during fixation, known as 'microsaccades'. Microsaccades have a time-scale that varies between 2 and 120 arcminutes. Higher in the hierarchy are the so-called 'drifts', which occur in between microsaccades. Slower actions can include arm movements, or walking a few steps. Even slower time-scale actions can include actions such as taking a bus, or waiting for the night to fall.

This shows that, in the PP framework, perception and action influence each other in a constant cycle. Even more, perception and action operate under the same principle of PE minimization. That is, perceptual inference and active inference are two ways of doing the same thing, namely, minimizing PE. However, this does not imply that perception and action amount to the same thing, that perception confounds with action in such a way that there is no fundamental distinction to be made between the two. Even though perception and action operate under the same principle of PE minimization, they are functionally distinct. While perceptual inference has mind-to-world direction of fit, active inference has world-to-mind direction of fit. This is not just a trivial difference that does not reflect something that exerts a causal difference in the workings of the machinery of the brain. The difference in direction of fit in question is key in this respect. As we saw above, regarding the need of assigning a high gain to counterfactual proprioceptive PE units during active inference for movement, while ignoring actual proprioceptive PE, the machinery of the brain must observe the distinction between perceptual inference and active inference. In other words, the brain must keep functionally separated the task of updating its models of

the world and the task of sampling the world on the basis of an already selected model/hypothesis—otherwise, in the proprioceptive case commented above, movement cannot occur.

This point will be key in this Thesis, as I will argue that in order for an emotion to arise, it is not sufficient to minimize interoceptive PE via interoceptive perceptual inference, as the current PP view of emotion holds. I will suggest that, even though interoceptive perceptual inference is certainly required during emotion, the driving component of emotion amounts to interoceptive *active* inference. As long as there is a relevant functional difference between perceptual and active inference, that is, as long as they do not conflate in the same kind of process, the view I will suggest it is not trivial in this regard: it significantly alters the current PP view on emotion.

2.1.7. Interaction across levels and modalities

Remember that context and background knowledge are important during perceptual inference. During percept formation in a certain modality—e.g., vision—there is a rich interaction across all levels of the visual hierarchy. This interaction operates in both directions, from the bottom-up and top-down.

The interaction in question occurs in such a way that fast-changing regularities are key in the selection of top-down hypotheses. For example, regularities governing shades bias the selection, for example, of the bird-hypothesis over the plane-hypothesis, given certain sensory signals—or also expectations relative to the shape of an object modulate activity in V1. On the other hand, top-down hypotheses are key in recovering low level sensory activity. For example, the encoded low time-scale regularity that birds beat their wings controls the low level dynamics relative to the way in which wings movement changes the contour of the bird. Furthermore, in case there is a learnt association between a certain cue in one modality and another cue in another modality,

cross-modal modulation also occurs (see Quattrocki & Friston, 2014). For example, wine colour likely modulates expectations about its taste (see Clark, 2016, p.55).

More interestingly, the interaction in question also spans the higher levels not only of the visual hierarchy, but also the higher levels of the whole hierarchy, beyond the visual 'channel': the amodal and multimodal higher levels that encode relatively more invariant regularities. To take a typical example, amodal /multimodal, higher level regularities about the strength of a bird can be used to modulate activity of the lower levels of the visual hierarchy, in order to extract regularities about the fast-changing regularities governing the sound of a birdsong (Friston & Kiebel, 2009).

> "Regularities can be ordered hierarchically, from faster to slower. Levels in the hierarchy can be connected such that certain slow regularities, at higher levels, pertain to relevant lower level, faster regularities (for example, slow regularities about aussie rules footy word frequency during the yearly news cycle pertain to faster regularities about the words I end up reading; if I know the slower regularity then I am less surprised by the occurrence of those words). A complete such hierarchy would reveal the causal structure and depth of the world—the way causes interact and nest with each other across spatiotemporal scales." (Hohwy, 2013, pp. 27-28)

By connecting the different levels of the hierarchy in this manner, in the PP framework percepts can be seen as arguably inheriting the depth of the nested causal structure of the world (which the cortical hierarchy recapitulates). In a word, in the PP framework, in order to form a percept in a certain specific modality, the whole system operates across all levels of the cortical hierarchy (for interesting examples, see Clark, 2016, pp. 86-87).

2.1.8. Recapitulating the structure of the world

Importantly, via learning, models manage to recapitulate the structure of the world. This will be key for my argument (Chapter 4). As Hohwy (2013) remarks, as the process of model/hypothesis selection and revision in light of precision-weighted PE

unfolds, and learning thus takes place, visual models manage then to extract regularities of their proper domain, namely, light-reflecting middle-sized objects. More generally, in this manner, generative models manage to extract the causal regularities from the world. In other words, *priors* are learned from experience (i.e., exposure and training), and over time they *recapitulate* the regularities that configure the hierarchically nested structure of the world[9]. In this manner models get to *represent* the world, so that as a certain model continues to improve its capacity to minimize its prediction error, the structure of the world gets to be better represented by the model (this is what is called the model's *accuracy*). This is precisely what allows the system to issue successful predictions of the worldly causes of incoming signals, and thus minimize PE.

Let's remember the example of the song of a strong bird above. In this example, amodal /multimodal, higher level regularities about the strength of a bird can be used to modulate activity of the lower levels of the auditive hierarchy, in order to extract regularities about the fast-changing regularities governing the sound of a birdsong.

Crucially, note that, in this example, for stored knowledge about the size and strength of a bird to strongly modulate, and thus shape the modality-specific representations of sound, there must be a regularity governing specific bird sounds and their size and strength, so that type of size and strength can predict type of sound (i.e., they are not probabilistically independent phenomena). Stronger birds must tend to sing more loudly. If this obtains, the cortical hierarchy can then, via learning, come to extract this regularity in the world. Thus, once recapitulated, this stored regularity can be exploited to predict the modality specific bird sounds in question from contextual knowledge relative to bird strength. In other words, the high-level 'strong-bird hypothesis' shapes percepts in the auditive channel in a certain way, since certain auditive signals, instead of other auditive signals, are expected given this hypothesis. Where do these priors (expectations) come from? From regularities in the world, via exposure and training

[9] Some priors might be hard-wired to some extent.

(i.e., experience). This stored expectation recapitulates then regularities in the world. Considering that this sort of expectations are built from experience, by extracting regularities in the world, their formation requires that there must be a regularity governing bird-strongness and type of sound, so that the former predicts the latter (i.e., they are not probabilistically independent phenomena).

Let me insist. For a certain piece of contextual knowledge *K* to modulate certain low level modality-specific activity *M*, the cortical hierarchy must built a 'prediction-enabling' association between *K* and *M*, so as to be able to predict *M* from *K*. Via learning, this regularity must be extracted from the world and recapitulated by the cortical hierarchy. Then, on the assumption that, via learning, the cortical hierarchy recapitulates the structure of the world, this can be put on more general terms. In order for a specific chunk of amodal/multimodal higher-level knowledge about *H* to systematically shape certain modality-specific lower-level representations about *L*, the following condition must be met: There must be a regularity governing *H* and *L*, so that *H* predicts *L* (i.e., they must not be probabilistically independent phenomena).

Or to put it in slightly different terms. The expectations about which sensory activity to expect given a certain object *O* that the model for *O* encodes are learned from exposure to *O* during training (i.e., experience). Then, if in a certain domain (e.g., detectable vibrations domain) there are no regularities about *O*'s, the model for *O* cannot get to encode this kind of sensory expectations about *O*'s for that domain.

This point will be key for my argument to the effect that the current version of the PP theory of emotion, the interoceptive inference view of emotion, is problematic. As I will argue in Chapter 4, there are no regularities pertaining to emotion in the physiological domain. Thus, it is unlikely that emotions arise by simply minimizing interoceptive PE via interoceptive perceptual inference.

To sum up, according to the PP framework, the brain, in all its functions, is engaged in the single task of minimizing its precision-weighted PE. This can be achieved in two ways. The brain can minimize its precision-weighted PE either via perceptual inference or via active inference. PP holds that percepts are formed via perceptual inference. That is, the brain forms percepts in a top-down fashion via the predictive generation of sensory signals in lower-levels. Such sensory signals are produced by higher-level models of the likely external causes of the lower-level sensory signals in question. Models are updated and improved through precision-weighted PE signals that result from that 'portion' of the incoming signals that did not match with the predicted/generated signals. Perceptual inference has mind-to-world direction of fit. On the other hand, the brain can also engage in active inference in order to minimize its precision-weighted PE. Active inference consists in changing the environment so as to obtain sensory data that fits considered predictions or hypotheses. Active inference requires knowledge of sensorimotor contingencies: representations of the counterfactual relations that obtain between (possible) actions and its prospective sensory consequences. Active inference has world-to-mind direction of fit.

In both kinds of strategies of PE minimization, the latter is minimized by way of a cascade of predictions that span the whole cortical hierarchy. Lower levels of the cortical hierarchy encode regularities that operate at fast time-scales, which capture variant aspects of experience. On the other hand, higher levels encode increasingly more complex regularities that operate at slow time-scales, which capture relatively more invariant aspects of experience. The different levels of the cortical hierarchy constantly interact, so that there are top-down and bottom-up influences that contribute in recovering regularities about the world. Finally, *priors* are learned from experience (i.e., exposure and training), and over time they *recapitulate* the regularities that configure the hierarchically nested structure of the world. That is, via learning models extract regularities in the world. This is allows the brain to issue successful predictions of the worldly causes of its incoming sensory signals.

3. Predictive processing and interoception

In this Chapter, I present, systematize, and refine Seth's (and Hohwy's) proposal that interoception operates under principles of PP. In this Chapter, I limit myself to presenting the interoceptive inference approach to interoception, independently of the question of what is the nature of emotion: How do the principles of PP apply to interoceptive processing *simpliciter*? This will set the basis to better grasp the main commitments of IIE—in the Chapter that follows (Chapter 4), I present and problematize IIE. The aim of this Chapter, then, is to present and discuss the main theoretical groundwork needed to unfold the coming discussion relative to the PP account of emotion and affective valence.

However, I begin this Chapter by discussing the notions of 'interoception', 'interoceptive percepts', and 'homeostasis' (Section 3.1.). Discussing these notions is key for my purposes. In the first place, IIE is articulated in terms of these notions. Then, discussing them is helpful for fully grasping the underlying commitments of IIE and ITV, and also its implications. In the second place, these notions play a major role in the claims defended in this Thesis, so they will be recurring themes in the Sections and Chapters to come.

Then, in Section 3.2., I motivate the view that the principles of PP readily apply to interoceptive processing. After motivating this view, I present, systematize, and refine the interoceptive inference approach to interoceptive processing. In this respect, I highlight three inferential strategies by which interoceptive PE can be minimized: *interoceptive perceptual inference*, *internal interoceptive active inference* (or 'internal interoceptive action'), and *external interoceptive active inference* (or 'external interoceptive action') (allostatic action). Distinguishing between these kinds of strategies is essential for the view proposed in this Thesis. Remember that IIE claims that emotions arise via *interoceptive perceptual inference* simpliciter, and that I expand IIE so as to claim that *external interoceptive actions* is what drives emotion generation.

Importantly, the latter strategy require stored *representations of 'sensorimotor contingencies'*. I discuss these notions in Section 3.3.

3.1. Interoception, interoceptive percepts, and homeostasis

3.1.1. Interoception and interoceptive percepts

Interoception consists in the perception of all aspects of the physiological condition of all tissues of the body (Craig, 2003). Interoception tracks several physiological variables via different sets of receptor types, or interoceptors, located all over the body. Interoceptors, most of which are polymodal – i.e., they are receptive to several physiological variables (Dworkin, 2007) – track *changes* in physiological variables such as, for example, metabolic rate, plasma concentrations of salt, cardiac perfusion, build-up of carbon dioxide in the bloodstream, steroids, temperature, inflammatory cytokine levels, endocrine activity, mechanical stress, thermal activity, extracellular fluid osmotility, lactate, concentrations of glucose, barometric pressure, muscle tension, stimulation of mucosa, and levels of insulin or cortisol, among other variables (Craig, 2015; Dworkin, 2007). That is, the kind of variables relative to which the spinothalamic pathway is sensitive (Craig, 2003). In other words, interoception consists in the perception of the autonomic, hormonal, visceral and immunological homeostatic changes (and their physiological effects) that together constitute the physiological state of the organism (see also Barrett, 2015).

Let me digress. By characterizing 'interoception' in this way, I am taking sides with, what might be called, a *moderately restrictive* characterization of interoception—or, if you are more optimistic, a *moderately inclusive* characterization of interoception. Let's call it simply the *moderate* view of what counts as 'interoception'. Ceunen and his colleagues (Ceunen et al., 2016) have brought attention to the fact that, in the relevant literature, there are two main kinds of views on what counts as interoception:

restrictive and *inclusive* views. *Restrictive* views (e.g., Dworking, 2007; Sherrington, 1948) consider that only the perception of activity originating in the viscera counts as interoception. Such activity ultimately results from the triggering of visceroceptors. On the other hand, according to Ceunen et al., for *inclusive* views, 'interoception' refers to the phenomenological experience of a body state, which ultimately results from activity in the central nervous system (CNS), independent of the kinds of receptors that trigger such an activity. If the resulting experience is felt as a bodily experience, then it counts as an interoceptive experience. Ceunen et al. consider that this kind of experience is multimodal, as it arises from "the integrated cross-modal CNS perception of the body state" (Ceunen et al., 2006, p.743). They consider that interoception is multimodal, since key regions in which interoceptive features are re-represented, such as the mid-insula and the anterior insula, integrate the latter with exteroceptive sensory states (vision, audition, etc.) (Craig, 2008). Interoception is the phenomenological experience of the body, but only insofar as the latter is integrated with information from all modalities.

I think that both views go too far. The restrictive view is too restrictive; while the inclusive view is too inclusive. The restrictive view fails to consider that, as it will be discussed below, the function of interoception is to track the physiological changes that define the homeostatic system of an organism. The function of interoception is to track homeostatic afferent information (Craig, 2015). All tissues of the body, and not just the viscera, are innervated by interoceptors that convey homeostatic afferent activity to the lamina I spinothalamic tract, known as the homeostatic pathway (Craig, 2015). Therefore, since all tissues of the body are innervated by interoceptors, and thus trigger homeostatic afferent activity, the restrictive view is not appropriate in light of anatomical and functional criteria.

The inclusive view holds that 'interoception' refers to the phenomenological experience of the body, which results from the integrated cross-modal perception of the body, independent of the kinds of receptors and specific sensory-channels involved. This view fails to consider that, for example, vision also exhibits the kind of

integration that Ceunen et al. consider. In fact, cross-modal sensory integration is a characteristic phenomenon of cortical organization at its higher levels, regarding all modalities or sensory channels. As Craig remarks:

> "Progressive re-representations that combine feature extraction and cross-modality integration are present in the serial processing streams observed in the visual, auditory, and parietal somatosensory cortical regions and are consistent with the evolutionary development of new processing regions in primate cortex" (Craig, 2008, pp. 279-280).

Certainly, at higher levels, interoceptive representations, just as sensory/perceptual representations in any modality, are integrated with representations coming from other modalities. However, we do no say from this that, for example, vision is multimodal or cross-modal in the sense that Ceunen et al. emphasize. That is, we are not *inclusive* about vision, nor audition, etc. There are no *a priori* reasons to claim that interoception is *sui generis* in this respect. On the other hand, the inclusive view relies too strongly in phenomenology as a criterion for demarcating what counts as interoception. Generally speaking, such a criterion by itself tends to produce faulty classifications (see, e.g., Macpherson, 2011). In the case at hand, it is not hard to find cases in which the relevant experience is exteroceptive, but it is felt as a bodily experience. For example, the experience of touch inside the mouth is felt as a bodily experience. However, touch is uncontroversially classified as an exteroceptive modality. The inclusive view is too inclusive.

It is also sometimes claimed that certain representations in the somatosensory system count as interoceptive representations, besides the uncontroversial interoceptive representations that result from the insular pathway (e.g., Khalsa et al., 2009; Barlassina & Newen, 2014). I think that including the somatosensory pathway from skin afferents to the somatosensory cortex as part of the interoceptive system is also too inclusive. Among other anatomical differences, this somatosensory pathway—the dorsal column-medial lemniscal pathway—is constituted by large-diameter sensory fibers from mechanoreceptors in the skin, contrary to the homeostatic pathway (the spinothalamic pathway), which is constituted by small-diameter fibers (Craig, 2015)

that innervate all tissues of the body. There is an important anatomical difference then between the somatosensory pathway and the insular pathway. There is also a functional difference. Activity in the somatosensory pathway results eventually in the experiences of discriminative touch and limb position. Both these sensory experiences are uncontroversially classified as exteroceptive and proprioceptive, respectively. Thus, I will align with what I called above a *moderately restrictive* characterization of interoception.

Adopting either the restrictive or the inclusive view has consequences for the nature and scope of the claim put forward by the *interoceptive inference view of emotion* (Seth, 2013; Seth et al, 2012; Seth & Critchley, 2013; Hohwy, 2013), and consequently, for expansions of the latter, such as the one suggested in Chapter 7. The main claim put forward by the interoceptive inference view is that emotions arise from interoceptive perceptual hypotheses about the causes of current interoceptive afferents. In this view, *in direct analogy* to the way in which visual percepts are formed (Seth, 2015a, Seth & Friston, 2016), for an emotion to arise, emotion models/hypotheses need to supress from the top-down incoming interoceptive inputs.

On the one hand, adopting the restrictive view amounts to the adoption of a notion of interoception already abandoned by the scientific community (for discussion, see Craig, 2015). Therefore, if the interoceptive inference view (and expansions of the latter) is read as embracing the restrictive conception of interoception, then its claim would be a non-starter. There is much more to interoception than just visceral perception (does any interoceptive theory of emotion wants to exclude from bodily perception, for example, temperature increases in the face?).

On the other hand, adopting the inclusive view of what counts as interoception leads to the trivialization of the interoceptive inference view of emotion as it has been put forward. This is the case, since, if the inclusive view is adopted, the claim put forward by the view in question would amount to the claim that emotions arise as prediction

errors from different modalities (including exteroceptive modalities) are minimized via emotion multimodal perceptual hypotheses. As long as the inclusive view trivializes the notion of interception by conflating it with integrated multimodal perception, but with an special focus on the body, the *interoceptive* inference view of emotion risks stop being an *interoceptive* view of emotion, properly speaking: it risks becoming a multimodal view of emotion. That is, it would end up holding something distinct from what it aims to claim, namely, that *in direct analogy* to the way in which visual percepts are formed (Seth, 2015a), for an emotion to arise, emotion models/hypotheses need to supress from the top-down incoming *interoceptive* inputs.

Now, the above should not be taken to imply that a multimodal view is implausible. It might turn out to be the case that some sort of multimodal view better captures the nature of emotion. It should certainly be explored. However, in this Thesis I am interested in discussing the interoceptive inference view insofar as it amounts to a PP version of the perceptual, interoceptive view of emotion. According to the latter, even though multimodal and abstract knowledge typically causally contributes in emotion generation, emotions only arise as interoceptive perception takes place, *in direct analogy* to the way in which vision operates (more on this in Section 4.4.4.). End of digression.

3.1.2. The interoceptive system as a perceptual system

The interoceptive system is certainly a sensory/perceptual system: it has evolved transducers and processes for capturing information about the inner physiological milieu, and its fined-grained representations can be centrally integrated for the purpose of action (see Matthen, 2015; Picciuto & Carruthers, 2014; Ritchie & Carruthers,

2011). I take Ritchie & Carruthers (2011) to have successfully shown that this is the case: the interoceptive system is a sensory/perceptual system[10].

As any other sensory system, the interoceptive system is a dedicated input system. As such, it uses disparate kinds of mental representation, and it has its own sensory code (Prinz, 2002). That is, the interoceptive system is sensitive to particular kinds of inputs, distinct from the inputs to which, for example, the visual system is sensitive. And this inputs are processed in their own proprietary: the 'interoceptive code'.

The end products of sensory processing are sensory/perceptual representations of the stimuli that cause such processing, i.e., percepts. Insofar as interoception is a sensory system, it then gives rise to percepts—and of a distinctive sort, namely, percepts of the physiological inner milieu commented above: *interoceptive percepts*.

3.1.3. Interoceptive percepts and phenomenal consciousness

In line with the standard view, I take percepts to be representations constituted by *bundles of bounded sensory features*. For example, in the case of a visual percept of a dog, many features, which are processed in different brain structures at different time scales, are coherently bounded together so as to form a unitary perceptual

[10] The interoceptive system likely comprises a collection of interoceptive sensory modalities. Ritchie & Carruthers (2011) do not go through the trouble of arguing for a specific view regarding which might be the correct list of such distinct interoceptive modalities. In fact, no philosophical work has been done on this issue yet. What Ritchie & Carruthers (2011) do is to consider the interoceptive system as a whole and show, by focusing on different aspects of it, that it satisfies criteria for counting as a sensory system rather than as another kind of mental mechanism (e.g., a purely motivational or cognitive mechanism). They do not consider each candidate single interoceptive modality (or some of them) and show that each of them (or some of them) is indeed a distinct individual modality, by showing how each of them (or some of them) satisfies all the just mentioned criteria. They neither discuss the issue of what makes a (candidate) single interoceptive modality the individual modality it is. Then their argument must be taken to only show that the interoceptive system as a whole is a sensory/perceptual system. I am agnostic as to what is the correct list of interoceptive modalities and sub-modalities. This ultimately depends on what is the better way to individuate the senses (see Macpherson, 2011; Mathen, 2015). Since nothing in my argument will depend on these matters, I will leave such discussion for another occasion.

representation of a dog: an oval-like shape, borders and edges that compose four legs, colours that attach to all these parts, a moving tail, contours that define a face, etc.

Let me briefly digress. I align with the default view in cognitive science that percepts and mental images (i.e., the products of imagery) are, structurally, basically the same kind of mental entity. They certainly arise under different conditions (during online and offline processing, respectively), and they might exhibit, in many occasions, differing degrees of granularity. However, to put it this way, both percepts and mental images are the same kind of mental end product: representations constituted by bundles of bounded sensory features. Even though I will be concerned here mainly (if not exclusively) with percepts, for convenience I will use 'percept' to refer to both percepts and mental images. End of digression.

It will be assumed in this Thesis that, under normal conditions of neural operation, phenomenal consciousness is populated by nothing over and above percepts, i.e., experience is restricted to percepts. Percepts *can* occur outside phenomenal consciousness, as priming and subliminal-perception studies show (e.g., Naccache & Dehaene, 2001; Winkielman et al., 2005; Wen et al., 2007). However, when consciousness arises, it is exhausted by percepts: consciousness does not outstrip percepts (see Prinz, 2012)[11]. That is, it will be assumed that there are no phenomenal qualities beyond sensory representations—independently of the issue of whether the latter can also have abstract contents. Note that the claim in question simply denies that there can be non-sensory representational vehicles with qualitative character (for discussion, see Prinz, 2012, pp. 149-168).

[11] The escape clause at the beginning of the paragraph is meant to cover cases of agnosia (particularly, visual agnosia). In such cases, there is consciousness without optimally formed percepts (Prinz, 2012). That is, agnosia patients seem to experience "broken percepts": bundles of features which are not optimally bounded together. However, note that the case that their consciousness is populated by "broken percepts", since their perceptual mechanism is malfunctioning after a stroke or other kind of lesion, bolsters the view that sensory features bounded together (i.e., percepts) is the material out of which the brain gets consciousness. There seems to be no cases of consciousness without percepts (or "broken percepts").

I am also assuming that there is unconscious perception, i.e., that coherently bounded sensory representations or percepts can occur outside phenomenal consciousness (so they cannot be reported) during online processing. As studies of subliminal perception show (e.g., Naccache & Dehaene, 2001; Winkielman et al., 2005; Wen et al., 2007), presented stimuli can be semantically processed, *and thus represented*, without being phenomenally conscious for the agent. This is evidenced by the fact that, after presentation, those stimuli can have a consistent impact on the subsequent behaviour of the agent. In these cases, a percept is formed—even though it does not need to be a rich, full in details percept or sensory representation—but it does not reach consciousness. The same holds for percepts in all modalities, as the brain conserves the same principles of organization across its sensory systems, and there is no *a priori* reason to exclude interoception (for a discussion of this assumption, see Block, 2011).

3.1.4. Interoceptive percepts are coherently unified representations that track cascading whole-body physiological changes that constantly evolve through time

Now, what sort of physiological activity is tracked by the interoceptive features that constitute interoceptive percepts?

As I mentioned above, interoception tracks several physiological variables. Importantly, once a local physiological change relative to a certain variable takes place, cascading changes are triggered in the activity of many variables and effectors all over the body landscape. These local physiological changes and effectors coordinate and modulate each other in parallel and across levels in a mutually constraining, network-like fashion (Craig, 2015, p.20; Dworkin, 2007). For example, muscle afferent activity in a certain tissue modulates cardiovascular activity; and changes in lactic acid and deprotonated phosphate in certain tissues modulates muscle sympathetic nerve activity and vasodilatory effect of metabolic product (Dworkin,

2007). Also the activity in certain thermoreceptors (a type of interoceptor) linearly modulates respiratory parameters, which in turn constrain heartbeat, circulatory activity, tissue metabolic rate, thermogenic, brown adipose tissue, compartmental vascular perfusion, panting, and sweating. In turn, each of these changes triggers a cascade of mutually constraining physiological changes on their own in a constantly evolving cycle (Craig, 2015; Diesel et al., 1990). Thus, by constraining and modulating cascading physiological changes in other variables across the body landscape, each local physiological change in a certain variable configures a novel shape in the whole-body physiological landscape.

It has been claimed that interoceptive percepts are better seen as a multimodal percepts that integrates sensory representations coming from all modalities, including the exteroceptive modalities (Ceunen et al., 2016). However, as I argued above (Section 3.1.1.), such an inclusive view of interoception is too inclusive, as it conflates the notion of interoception with the notion of integrated multimodal perception. After all, physiological perception (interoception) occurs whether or not it is integrated with what the organism is currently seeing or hearing.

3.1.5. The interoceptive hierarchy

Distinct types of interoceptors—which are mostly polymodal— located in different tissues all over the body landscape track these constantly evolving cascades of mutually constraining local physiological changes. This information is received by the brain[12]. As it is the case with sensory/perceptual systems in general, the interoceptive system is likely to be hierarchically organized (see Chapter 2). After initial processing in functionally specific interoceptive thalamic sensory regions – e.g., VMpo, VMb, MDvc –a somatotopically organized interceptive primary sensory region represents local physiological changes as they occur in different tissues of the body (Craig, 2010).

[12] Not all interoceptors project to the cortex. Some of them work only locally (Craig, 2015).

To take a speculative example, at these levels of the interoceptive hierarchy, representations track the rapidly evolving activity of different fibers located in different parts of the heart, and in different parts of other organs and tissues. That is, at these levels of processing, the rapid activity of restrictive receptive fields is represented.

Good candidate structures for realizing these levels are the posterior and middle insula, which constitute a primary interoceptive cortex (Craig, 2015), roughly analogous to V1 in the case of vision. These regions receive visceral inputs from regions such as the nucleus of the solitary tract, and from the lamina-1 spinal tract, which form part of a functionally specific interoceptive pathway (Craig, 2002).

As is the case with respect to standard sensory systems, such as vision, the interoceptive system must also exhibit levels of processing where comparatively more invariant features are encoded, such as for example, features pertaining to the rate of the whole heart, and also features pertaining to the more global physiological condition of other single organs and tissues. There must also be levels of processing that represent even more invariant features, such as for example, the physiological condition of the whole body. Candidate structures involved in realizing these latter levels of processing include, for example, regions of the insula, regions of ACC, and regions of the amygdala, among others (Craig, 2015; Prinz, 2012, pp.65-66).

As we saw in Chapter 2, where I presented the PP framework, forming world revealing percepts requires finding a stable solution across all levels of the perceptual hierarchy. The same should hold for *inner*-world revealing percepts. Interoceptive percepts then merge features that track changes in several kinds of cascading, mutually-constraining physiological variables all over the body, giving rise to a coherent and unified representation of the physiological condition of the whole body at its different time-scales. In other words, interoceptive percepts provide a global, hierarchically organized, and constantly evolving "view of the body landscape" (Damasio, 1994).

The insular cortex is key for interoceptive percept formation. The insula is a complex structure, composed of several subregions. All of those subregions serve the function of representing "visceral" activity (Deen et al., 2011). Interestingly, the insula is a structure that is richly interconnected with regions involved in reward processing, exteroception, motor control, and cognition (Medford & Critchley, 2010; Deen et al., 2011). Thus, even though the insula is an interoceptive structure, it is highly sensitive to information coming from other senses, cognition, reward systems, proprioception, etc. Particularly relevant here is the anterior insula (AIC), which is highly sensitive to exteroceptive information and realizes a whole body representation of the physiological condition of the body (Craig, 2015). In AIC, interoceptive percepts are integrated with such exteroceptive representations. As we will see below, taking into account all sorts of information is a critical part of the process of interoceptive percept formation via *perceptual inference*. This makes then the insula, particularly AIC, a key region for instantiating predictive interoceptive inference during percept formation (Seth, 2013). On the other hand, AIC is critical for the conscious representation of the inner milieu (Craig, 2015). Thus, considering that consciousness is restricted to percepts, this supports the claim that the anterior insula is critical for interoceptive percept formation.

However, even though the anterior insula is critical for interoceptive percept formation, it should not be seen as necessary and sufficient for interoceptive percepts. Damasio (Damasio et al. 2013) reports that a patient with a major bilateral lesion in the insular cortices exhibits normal bodily experience. Thus, interoceptive percepts are likely to be realized in a distributed manner, across all subcortical and cortical regions of the interoceptive hierarchy.

3.1.6. Interoceptive percepts as 'multimodal interoceptive percepts'

Insofar as interoceptive percepts incorporate coherently unified features that represent changes *in several kinds* of cascading, mutually-constraining physiological variables all over the body, they can be taken to be *multimodal* interoceptive representations: they track physiological changes for which there are distinct types of sets of interoceptors. For ease of exposition, from now on I will refer to these representations simply as 'interoceptive percepts' instead of as 'multimodal interoceptive percepts' (see also footnote 10).

3.1.7. Representing proximal physiological changes via interoceptive percepts: the content of interoceptive percepts

Insofar as the interoceptive system is a sensory/perceptual system, it *represents*. As I mentioned above, I take Ritchie & Carruthers (2011) to have successfully shown that the interoceptive system is a sensory/perceptual system, and that its final products do count as truly representational. That is, the interoceptive system issues proper percepts[13].

[13] I am assuming that such final products are proper sensory *representations*. I am aware that in some corners of the cognitive science community it is held that sensory processing, and cognition more generally, does not harbour representations (Varela et al., 1991; Chemero, 2009). Enactivists emphasize that it is unlikely that the mind's job is to recover a mind-independent world in the way that a mirror captures the things that get to be in front of it. Minds evolved so as to act within its own ecology, to put it that way. What an organism is able to do specifies what she perceives, and vice-versa. The mind and its own ecology mutually specify each other. As long as sensory representations are taken to be mirrors of an agent-neutral world, sensory representations should certainly be looked with suspicion. I think these insights are on the right track. However, they do not speak against representations. There is no need to take representations as mirrors of an agent-neutral world. In fact, representations are arguably *action-oriented* (Clark, 1997; Millikan, 1996). That is, they jointly encode aspects of the world and specify relevant actions. In line with the insights of enactivism, the aspects of the world that they encode are better seen as capturing the task-relevant, ecologically salient aspects of the niche that the organism contributes to specify through action.

Now, it is natural to hold that what the end products of interoception represent are the physiological changes that impinge our interoceptors. On the one hand, the mental states that are the final product of interoception causally co-vary with such physiological changes. Moreover, they have the function of representing those physiological changes in order to maintain homeostasis (Craig, 2015)—they are *used* by the organism *for* that purpose. Then, by the lights of any family of naturalistic theory of content – either a causal or a consumer theory – what the end product of interoceptive processing represents are the physiological changes involved in homeostasis maintenance (more on this below).

Besides the argument above, the claim that interoceptive representations have physiological changes as contents is supported by the fact that interoceptive states such as pain and orgasm do exhibit such kind of content, as Tye (1995) has successfully shown. For example, part of the content of pain is, roughly, tissue damage; while part of the content of orgasm is, roughly, muscle contractions in certain pelvic regions. On the other hand, all sensory systems typically have content: vision, audition, touch, taste, proprioception, etc. It would be simply arbitrary then to exclude the interoceptive system (see also Schroeder, 2001) (for discussion, see Block, 2006) (more on this below, in Section 6.3.2.).

Now, external sensory modalities such as audition, vision and touch, besides representing proximal stimuli, they (mainly) represent distal stimuli. For example, vision represents both the patterns of light that impinge our transducers, and medium-sized objects in the external environment (e.g., a dog), respectively. However, in the case of interoception there seems to be no proximal/distal distinction (Ritchie & Carruthers, 2011, p. 357). Then, the interoceptive system seems to only represent the physiological changes that impinge our interoceptors; it does not represent more distal

objects beyond the skin[14]. Interoceptive percepts then represent the inner world: patterns of constantly evolving physiological changes.

3.1.8. Interoceptive percepts represent changes

Note that interoceptive afferents respond to *changes* in physiological variables, not to the properties of physiological variables as such. That is, interoceptive afferents respond to *increases and decreases* in the activity of physiological variables; they do not signal an indication of the precise quantitative value taken by a certain variable at a certain time. Interoceptive percepts then are better seen as representing physiological *changes*. Importantly, whole-body increases and decreases in the activity of physiological variables occur as the organism (or some of its parts) attempts to keep its overall physiological workings within a certain range of viability. This process is known as *homeostasis*, and it will prove to be a key notion in the coming Sections and Chapters.

3.1.9. Homeostasis and its network-like nature

In the sense that will be relevant in this work, homeostasis basically consist in maintaining an optimal overall physiological regulatory level, or balance, able to keep the entire organism's body within its limits of viability. The interoceptive system, which tracks whole-body physiological changes, evolved for maintaining homeostasis (Craig, 2015). Interoceptive percepts represent then the physiological changes that occur during homeostasis regulation. In other words, interoceptive percepts represent

[14] Prinz (2004) holds that interoceptive representations also track distal properties, namely, core relational themes. However, Prinz's argument for that claim it is based on a dubious hypothesis, as I will argue in Chapter 4. Then, and to borrow Prinz's terminology, interoceptive percepts do not have *real contents* beyond the skin; they only have *nominal contents*, which correspond to physiological changes.

the physiological changes that configure a homeostatic system. Thus, interoceptive percepts inform us how we are faring in maintaining homeostasis.

Now, contrary to what is usually assumed without argument in the philosophical literature (e.g., Corns, 2014), homeostasis does not operate in a "thermostat-like" fashion: single, segregated variables (e.g., osmotic pressure) each being separately regulated to a certain static built-in value, which supposedly each variable individually has (Craig, 2015, p.20; Dworkin, 2007):

> The brain does not contain a collection of thermostats, each controlling a separate condition, and there is no single, overarching command center that controls all functions. For instance, thermoregulation is now recognized to involve a redundant set of anatomically distinct neural mechanisms at several levels of the neuraxis that are the products of successive evolutionary improvements. These mechanisms control multiple effectors, such as tissue metabolic rate, thermogenic brown adipose tissue, compartmental vascular perfusion, panting, sweating, shivering, and behaviour. […]Furthermore, thermoregulation interacts with a variety of homeostatic conditions, such as energy metabolism, salt and water regulatory hormones (renin, aldosterone, atrial natriuretic peptide), sweat and saliva production, cardiac and respiratory functions, renal filtration, and most important, behaviour (cold-seeking or heat-seeking). For example, a single intravenous injection of hypertonic saline in a rabbit, rat, or human simulating dehydration and salt imbalance will raise the blood pressure and cardiac output (after transient decreases) but reduce the metabolic rate, respiration, and core temperature; all of these effects can be viewed as responses that conserve water. Yet it will also raise the core temperature threshold for sweating, a heat defence mechanism, and at the same time increase cold-seeking/heat-escape behaviour in either a thermoneutral or a warm environment; notably, these effects would seem contradictory in a "thermostat" model. (Craig, 2015, p.20-21).

As I mentioned above, once a local physiological change relative to a certain variable takes place, cascading changes are triggered in the activity of many variables and effectors all over the body landscape, which mutually constrain each other in a network-like fashion. Multiple whole-body variables are then orchestrated by central, autonomic and endocrine processes, so that changes in one variable coordinate compensatory changes in other variables; each of which is adjusted in turn by another set of physiological changes in a constantly evolving cycle. Thus, by constraining and modulating cascading physiological changes in other variables across the body landscape, each (homeostatically relevant) local physiological change in a certain variable configures a novel shape in the whole-body physiological landscape.

Homeostasis requires then coordinating many effectors and variables in parallel all over the body, so that an overall, whole-body adaptive physiological balance can be achieved.

This constantly evolving whole-body physiological landscape that results from the aim of maintaining homeostasis gets nicely illustrated by borrowing Damasio's waterbed analogy: "when someone walks on it [a waterbed] in varied directions: some areas are depressed, while others rise; ripples form; the entire bed is modified as a whole […]" (Damasio, 1994, p. 135). Interoceptive percept represent then this ever-evolving wide landscape of local changes as they take part on a whole-body network of further compensatory changes.

Since the regulation of a certain local physiological change mobilizes a wide range of mutually-constraining whole-body physiological changes in other variables in a constantly evolving cycle, the achievement of homeostatic balance is then not best described by the thermostat metaphor, where the regulatory target amounts to a single variable that is individually regulated to a rigid regulatory level. The network-like nature of homeostasis maintenance also implies that homeostatic regulatory levels are *flexible* to a certain degree. There still is a regulatory level toward which homeostatic processes aim. Nonetheless, the fact that homeostasis involves multiple dynamic processes that negotiate changes across many variables implies that oscillations around this level are rather broad and set to change (Bernston & Caccioppo, 2007)—that is why some researchers prefer to use the term 'heterostasis'. Then, contrary to the thermostat view of homeostasis, there is no single set-point, but rather a collection of dynamic functions that "interact to maintain an optimal use of energy in the body across all conditions at all times" (Barrett, 2015, p. 422). Moreover, the network-like nature of homeostasis—i.e., local changes determining changes in many other variables in the whole-body landscape—also implies that any triggered physiological change in some variable is homeostatically relevant, insofar as it will trigger further changes in other variables, modifying thus the shape of the whole-body physiological landscape. It is no surprise then that variables such as blood pressure seem to have no

set-point: it varies significantly during the day of a normal adult depending on the demands implied by changes in other variables (Sterling, 2004).

3.1.9.1. Two ways of rectifying homeostatic imbalances: internal and external actions

Importantly, in order to overcome a homeostatic imbalance, an organism has two kinds of actions available. They might be called *internal actions* and *external actions*. The former consist in automatically executing "physiological policies" by making use of resources that are already available within the organism. For example, as when an organism deals with an increase in effective osmotic pressure of plasma (which is one of the ways whereby the feeling of thirst can be eventually triggered), the body responds by secreting vasopressin, stimulating the renin-angiotensin-aldosterone system, reducing renal solute and water excretion, among other whole-body "physiological policies". These are internal actions. In the case of internal actions, when physiological changes move in the direction of going beyond viability limits, homeostatic balance is rectified then by triggering internally accessible physiological resources.

On the other hand, an organism can deal with an increase in effective osmotic pressure of plasma by modifying her environment, for example, looking for a beverage. These are external actions—sometimes also referred to as 'allostatic actions' (Gu & Fitzgerald, 2014; Sterling, 2004; Seth, 2015a). Homeostasis drives behaviour. External actions require motivation. Regions of the interoceptive pathway encode urges to change physiological changes already in early stages of interoceptive processing. The main regions involved are the brainstem, the hypothalamus and ACC, which receive projections from interoceptors' activity via the parabrachial nucleus, and are heavily interconnected with interoceptive regions likely involved in realizing interoceptive percepts. This makes plenty of evolutionary sense, for, let me insist, in many cases the only way in which certain physiological imbalances can be corrected is by acting in the world. Then, the interoceptive system includes motivations that impel the

behavioural responses required to maintain homeostasis (it is no coincidence then that mental states such as hunger, thirst, cold, etc. are known as 'motivations').

3.1.9.2. Homeostatic regulatory standards

Maintaining homeostasis can be taken to be a hard-wired goal of biological systems (e.g., Friston, 2010; Friston & Stephan, 2007). Goals set standards according to which something can be evaluated. Thus, body landscapes can be taken to be *good* or *bad*, *positive* or *negative*, according to whether they tend to *approach* or *deviate* from the aimed-at (flexible) regulatory level of homeostasis maintenance, respectively.

By this I do *not* mean to claim that body landscapes can be taken to be positive or negative in case they tend to *promote* or *threaten* homeostasis: given that homeostasis maintenance works in a network-like fashion, departures from a homeostatic norm in certain variables triggers compensatory physiological changes in other variables. The latter changes involve sacrificing regulatory balances for the sake of avoiding even major imbalances in target variables. In this kind of case, such compensatory changes, in a sense, can be taken to *promote* homeostasis maintenance. However, insofar as they deviate from its aimed-at homoeostatic range of levels, they do not count as *positive* body landscapes in the sense I intend.

To sum up, the interoceptive system gives rise to interoceptive percepts. The latter amount to bundles of coherently unified interoceptive features that proximally represent the constantly evolving cascade of whole-body physiological changes involved in homeostasis maintenance. Physiological changes can be positive or negative relative to homeostatic standards. Homeostatic imbalances can be rectified by internal and external actions.

3.2. Motivating PP as applied to interoception

As I mentioned in Chapter 2, the PP framework has been mostly applied to the explanation of exteroceptive percept formation, in particular to vision (e.g., Rao & Ballard, 1999). In this and other exteroceptive (and motor) domains, PP has proven to be explanatorily successful and surprising, showing thus that the hypothesis of the predictive mind is on the right track. Indeed, the fact that PP offers compelling accounts in those domains at several levels of processing—including the retina (Srinivasan et al.,1982), thalamus (Jehee & Ballard, 2009), and sensory cortex (Rao & Ballard, 1999)—suggests that the computational principles of PP constitute a general solution of the nervous system.

This begins to suggest that the process by which interoceptive percepts are formed is quite likely to be also governed by the principles of predictive processing (PP). Indeed, as Seth (2015a) and Hohwy (2013) remark (see also Friston et al., 2013), the PP framework applies much more naturally, and fundamentally, to the interoceptive processes that take place during homeostasis maintenance. The PP framework is then, to say the least, a more than promising approach to the nature of the workings of interoceptive percept formation. Let me briefly explain some of these claims.

As we saw in Chapter 2, PP holds that the brain forms percepts by minimizing the difference between the actual sensory signals it receives from the world and the signals it predictively generates in a top-down fashion on the basis of its models of the most likely worldly causes of such incoming signals. This difference is known as 'prediction error' (PE). As we saw in Chapter 2, the brain is engaged in the single task of minimizing PE, so that both perception and action result from the brain's best attempts at minimizing PE on the basis of its stored predictive models.

Importantly, PP can be taken to be the brain's evolved perceptual and behavioural strategy for inferentially dealing with the more fundamental task of complying with the *free energy principle* (Friston 2005, 2009, 2010). Roughly, according to the latter, organisms must ensure that they maintain themselves within expected bounds of 'surprisal'. In other words, in the PP framework, perception and action "emerge as a consequence of a more fundamental imperative towards avoiding 'surprising' events" (Seth, 2015a, p.5). Roughly, a certain organism is in a 'surprising' state in case it is outside the subset of possible states that are most probable for such organism to be in—considering the way the organism in question is constituted—simple because there is low probability of finding such organism in that state, given its constitution.

Now, a defining characteristic of biological systems is that they maintain homeostasis, i.e., they regulate their inner physiological milieu so as to keep it within viability bounds (see Friston, 2010, p.1). 'Surprising' states, in the sense defined above, reflect conditions incompatible with homeostasis. That is, avoiding 'surprising' events roughly amounts to avoiding putting the organism away from homeostatic balance. This is the case arguably because 'surprisal' is a measure that is quantified relative to the constitution of an organism, and the latter, in turn, determines the kinds of transactions with the environment in which it is most probable to find the organism in question. Thus, "the long-term (distal) imperative—of maintaining states within physiological bounds—translates into a short-term (proximal) avoidance of surprise" (Friston, 2010, p.2). In a word, organisms require minimizing 'surprisal' so as to stay within its peculiar limits of physiological viability: "low free-energy systems will look like they are responding adaptively to changes in the external or internal milieu, to maintain a homeostatic exchange with the environment" (Friston & Stephan, 2007, p. 428). However, organisms cannot estimate 'surprisal' directly, so they need to rely on the PP strategy of PE minimization in order to achieve finding themselves in low 'surprising' states (i.e., states compatible with homeostasis). Then, considering that the function of interoception is to track physiological changes so as to maintain homeostasis, minimizing interoceptive PE is a more fundamental task than minimizing exteroceptive PE, given the imperative of keeping 'surprisal' quantities low. In fact,

that is why minimizing exteroceptive PE (via perceptual inference and action) "emerges as a consequence of a more fundamental imperative towards homeostasis" (Seth, 2015a, p.3), or if you want, it "emerges as a consequence of a more fundamental imperative towards the avoidance of "surprising" events" (Seth, 2015a, p.5). Thus, PP straightforwardly, and fundamentally, applies to interoception, given the deeper background of the free energy principle.

Compared to exteroceptive (and motor) processes, PP research on interoception has been scarce and still remains more speculative – probably because it is technically more challenging to gather evidence in the interoceptive domain, since interoceptive processes involve deep structures in the brain, and homeostatic processes are highly complex. However, early research on homeostatic physiological regulation recognized the need of mechanisms that hint some of the elements that PP incorporates into its architecture. For example, already Cannon recognized the need of mechanisms able to *anticipate* visceral activity by way of learned responses, so as to rectify physiological imbalances (Cannon, 1928). Also control architectures that include a predictive forward model that estimates visceral activity based on learned responses have been long recognized as key for homeostatic regulation (e.g., Dworkin, 1993).

More interestingly, as Seth (2013) remarks, various strands of current evidence suggest that interoception operates under the principles of PP. Firstly, contrary to the assumption that interoception works in a feed-forward, bottom-up fashion, interoceptive regions in the brain, besides constantly exchanging information across levels, exhibit significant top-down projections to physiological control regions in the brainstem and spinal cord (Critchley & Harrison, 2013). Secondly, as I mentioned above, AIC is a key region of interoceptive representation that instantiates (together with other regions) a level of sensory/perceptual processing at which interoceptive representations can become conscious (Craig, 2015). Neuroimaging studies show that AIC is involved in expectation processing and responds to prediction errors in cases of homeostatically relevant phenomena, such as pain perception (Ploghaus et al., 1999; Seymour et al., 2004), affective touch (Lovero et al., 2009), itchiness (Holle et al.,

2012), and in cases of affective responses to external situations (Xiang et al., 2013). Finally, there is evidence of interoceptive PE signals in interoceptive regions, such as the dorsal middle and posterior insula, in the case of inspiratory load during exercise in professional athletes (Paulus et al., 2012), and during threat processing in the case of elite war fighters (Paulus et al., 2010).

Furthermore, several studies on pain, which amounts to a homeostatic affective state grounded in interoception (Craig, 2002, 2003, 2015), suggest that it does not simply result from bottom-up feed-forward processing. For example, there is evidence that top-down modulation attenuates the strength of incoming nociceptive signals, suggesting that there is in the case of pain a phenomenon similar to signal suppression in the way postulated by PP (Calejesan et al., 2000). It has been also shown that contextual factors, expectations, and attentional phenomena can shape the processing of pain (Jepma & Wager, 2013; Wiech et al., 2008; Ploghaus et al., 1999).

To sum up, the process by which interoceptive percepts are formed is quite likely to be governed by the principles of PP, for it follows rather directly from the fact that PP is a case of the free-energy principle, and some strands of evidence suggest that it might be so.

3.3. The workings of interoceptive (predictive) inference

Before presenting Seth's and Hohwy's versions of the interoceptive inference view of emotion in the next Chapter, in the coming Sections I systematize and refine the ways in which the principles of PP operate in the case of interoception in general. That is, in the coming Sections, I systematize the workings of interoceptive inference, independently of whether or not they can be used to account for emotion *per se*.

The workings of interoceptive (predictive) inference

The view that interoceptive and homeostatic processes are better understood along PP lines has been mainly proposed by A. Seth (Seth, 2013, 2015a; Seth et al, 2012; Seth & Critchley, 2013)—see also Hohwy (2011, 2013). According to Seth, the interoceptive processes that occur during homeostasis maintenance take place via predictive *interoceptive inference*. More precisely, according to Seth, *in direct analogy to the way exteroceptive percepts are formed* (Seth, 2015a; Seth & Friston, 2016), the subjective feeling states characteristic of interoceptive processing arise by way of interoceptive perceptual inference. That is, subjective feeling states arise by minimizing interoceptive prediction error (PE) via interoceptive inferences of the likely causes of incoming interoceptive signals. Let's call this view the *interoceptive inference approach* (IIA).

Importantly, insofar as drives (or homeostatic motivations), such as hunger and thirst, are quintessentially states that result from interoceptive processing, I will consider those kinds of states in my exposition of the interoceptive inference approach to interoception as representative of the kind of subjective feeling states that result from interoceptive processing. The issue of whether emotions *per se* and valence as such can be understood as arising via interoceptive inferences will be discussed in the next Chapters.

Allow me now to clarify a minor terminological issue. Even though Seth phrases IIA in terms of subjective *feeling* states, IIA should be understood as an account of interoceptive *percept* formation—percepts, remember, can occur outside consciousness (i.e., they can be *unfelt*). IIA should be understood in this way since IIA basically amounts to the direct extension to interoception of the process by which percepts are formed in the exteroceptive domain (e.g., vision). In other words, IIA basically amounts to the working of visual inference during visual percept formation, but applied to interoception. However, given that percepts generally take place in consciousness (i.e., they are typically *felt*), it is not misleading if IIA is taken as an account of subjective feeling states. Nonetheless, it should be kept in mind that IIA is

an account of inner feelings only insofar as it is an account of interoceptive percepts—which *can* occur outside consciousness, as I commented above.

How does interoceptive inference work in the IIA framework? According to IIA, subjective feeling states (or interoceptive feelings) arise via interoceptive inference in the following way. Firstly, actual physiological changes take place. This can occur under several conditions. For example, actual physiological changes can be triggered by (non-centrally triggered) autonomic control signals (e.g., the reduction of renal solute and water excretion), and more indirectly by demands of the musculoskeletal system (e.g., having to run away) and by external conditions (e.g., snow falling). These occurring physiological changes produce actual incoming interoceptive signals that need to be 'explained away'. Such signals need to be 'explained away' by a hypothesis about their cause in the physiological domain. For example, let's say that it is snowing. Certain physiological changes take place, for example, peripheral vasoconstriction. They cause certain incoming interoceptive signals. The latter need to be 'explained away'. A good hypothesis (one with high prior probability and likelihood, so with high posterior probability) is that such signals are being caused by low body temperature (cold). Incoming signals are successfully explained away. The experience of cold arises, i.e., the felt interoceptive percept of cold is formed.

Let me briefly digress. Remember that the above actual incoming interoceptive signals need to be 'explained away' (so as to minimize interoceptive PE) because this is the manner in which the brain can achieve its distal goal, which is reducing this other quantity, distinct from PE, namely, 'surprisal'. Also remember that 'surprisal' can be taken to be at optimal (low) levels in case the organism stays in ranges of physiological balance. Then, 'explaining away' actual incoming interoceptive signals is the *proximal*, immediate goal that *interoceptive inferences* need to achieve, just as the proximal goal that visual inferences need to achieve is 'explaining away' incoming visual signals. While minimizing 'surprisal', or keeping an overall physiological balance in the long run, is the *distal*, long-run goal that the brain, in *all its functions*, including interoception, needs to achieve. Then, even in the case of visual inference the distal goal can be taken to be minimizing 'surprisal'. Also note that the *direction*

of fit relative to the task of 'explaining away' actual incoming interoceptive signals is mind-to-world, i.e., finding an hypothesis that successfully models the way the (inner) world is. On the other hand, note that the direction of fit relative to the distal goal of keeping an overall physiological balance is world-to-mind, i.e., changing the (inner) world so as to fit an expectation of homeostatic balance (i.e., keeping surprisal low) (more on this below). End of digression.

Now, according to Seth, actual incoming interoceptive signals are met, and compared, with signals generated from the top-down. These latter signals amount to predictions about the likely interoceptive signals that should be actually taking place, given a content-specifying model of the likely causes of such signals. These models then can be seen as putting forward hypotheses about the causes of incoming signals (e.g., "if it is hunger, I expect these interoceptive signals to be taking place. Let's generate those signals. Do they match incoming signals?"), and they are hierarchically organized in several layers of processing, where the layer above attempts to predict the activity of the layer below. In a word, models of the likely causes of incoming interoceptive signals specify predictions by generating from the top-down the expected activity in lower layers of processing, and this strategy is repeated across all levels of interoceptive hierarchy.

Importantly, the comparison of actual incoming interoceptive signals and the interoceptive signals generated from the top-down gives rise to *interoceptive PE*. The latter amounts then to the difference between expected and actual signals. The task of the brain during interoceptive inference is to minimize this difference by suppressing or 'explaining away' from the top-down its incoming signals, or by acting so as to change its inputs, as we will see below. This resulting PE is used to refine current predictions, so as to issue better predictions able to successfully minimize occurring interoceptive PE. PE signals are also used as a learning signal to improve, in the long-run, the models responsible of generating predictions. According to Seth, *in direct analogy* (Seth, 2015a; Seth & Friston, 2016) to the case of vision (and exteroceptive processing more generally), where a visual percept is formed when visual PE is successfully minimized, subjective feeling states arise when interoceptive PE is

successfully minimized. The content of the resulting experience or percept is specified by the content of the predictions generated from the top-down. At the higher-levels, the models that contribute in specifying interoceptive hypotheses encode not only interoceptive information, but also encode multimodal and amodal information. This sort of information contributes in linking interoceptive states with exteroceptive states. In other words, the brain uses all it has so as to minimize interoceptive PE and thus form an interoceptive percept: amodal high-level contextual knowledge, and knowledge from other modalities that laterally feed interoceptive hypotheses. These pieces of knowledge inform interoceptive inference across all levels of the interoceptive perceptual-hierarchy.

However, note that during interoceptive PE minimization, amodal and multimodal knowledge serve *interoceptive* ends: finding the better interoceptive hypothesis. In other words, forming an interoceptive percept requires meeting incoming *interoceptive* signals alone, and this is achieved by generating from the top-down *interoceptive* signals—i.e., signals proprietary of the visual channel won't do the job of suppressing *interoceptive* incoming data. Let me insist, only once the bottom-up activity proprietary of the interoceptive channel is 'explained away', an interoceptive percept is formed. Then, even though the models that contribute in specifying interoceptive hypotheses encode, besides interoceptive information, also multimodal and amodal information, such non-interoceptive information is used for the sake of finding the interoceptive signals that better match the incoming signals proprietary of the interoceptive channel. So during interoceptive PE minimization, amodal and multimodal knowledge serves *interoceptive ends*: influencing the interoceptive data that eventually will meet activity in the lower layers of the interoceptive channel, so as to minimize interoceptive PE.

As we will see below, the anterior insula (AIC) is key for interoceptive percept formation, not only because is thought to be critical region for comparing top-down and bottom-up interoceptive signals, but it is also sensitive to exteroceptive and proprioceptive information during interoceptive inference (Seth, 2013, 2015a; Seth et al, 2012; Seth & Critchley, 2013; Seth & Friston, 2016). This is analogous to the case

of vision—remember that IIA is built on the *direct analogy* to visual percept formation via visual PE minimization (Seth, 2015a; Seth & Friston, 2016). In the case of vision, forming a percept requires meeting incoming *visual* signals alone (to put it this way, one does not need to hear so as to see). Remember that sensory systems are dedicated input systems. Certainly, in the PP framework, during visual percept formation, amodal knowledge and knowledge from non-visual modalities contribute across all levels of the visual hierarchy. However, they contribute by constraining the kind of visual activity that is generated from the top-down so as to successfully 'explain away' the incoming signals proprietary of the visual channel. If these signals proprietary of the visual channel are not 'explained away', no world revealing visual percept gets to be formed. This is an aspect of the PP framework that is not usually stressed, but I think is critical to have it in mind when it comes to philosophically determine the more theoretical consequences of the framework. This aspect of the PP framework will be key later (Chapter 4) when evaluating the interoceptive inference view of emotion (IIE).

Expectations about *precisions* play a key role during this whole process. Precisions is the quantity that reflects the inferred reliability of the signal. Remember that depending on how reliable the signal is inferred to be, precision estimations control the balance between the weight that top-down and bottom-up influences have at different levels of the perceptual hierarchy. If the signal is deemed to be unreliable, top-down predictions are given increased weight relative to bottom-up influences. However, if the signal is deemed to be highly reliable, bottom-up influences are given more gain relative to top-down predictions. Importantly, precisions are encoded by post-synaptic neurotransmission, and remember that, in the PP framework, this process amounts to attention. Expectations about precisions then determine the balance between bottom-up and top-down influences during percept formation (and action, as we will see below). This latter point will prove to be important in Chapter 6, when it comes to dealing with some objections to the hypothesis that valence is grounded in the interoceptive system.

3.3.1. On the causes of interoceptive signals

IIA is then the view that, in direct analogy to the way in which exteroceptive percepts are formed, subjective feeling states arise by minimizing interoceptive prediction error (PE) via interoceptive inferences of the likely causes of incoming interoceptive signals. Thus, we must assume that instead of determining what it is the object or event in the external world that is likely causing incoming signals, as in the case of exteroceptive inference, during interoceptive inference the causes that need to be determined amount to the causal regularities that obtain in the inner world. That is, the physiological changes that take place during homeostasis maintenance. To put it this way, given that interoceptive signals (after transduction) are directly triggered by physiological changes in the 'inner environment' (such signals are precisely what the interoceptive system evolved to track), the causes that need to be inferred by interoceptive models and hypotheses must be physiological in nature. After all, patterns of changes in the physiological landscape are the kind of thing that causes interoceptive input. I take this to be obvious. Remember that sensory systems are dedicated input systems. Insofar as the interoceptive system is a sensory system, it is sensitive to particular kinds of inputs, distinct from the inputs to which, for example, the visual system is sensitive. The physiological inner milieu is precisely the domain to which the interoceptive system is responsive. This is particularly the case, if we embrace what I called above the 'moderate view' about what counts as interoception (Section 3.1.1.)

Certainly, as long as exteroceptively represented external objects and events can cause physiological changes via learned associations between a certain external cue and a certain specific physiological pattern (e.g., seeing a nutritious slice of pizza can cause the physiological changes characteristic of hunger), the *indirect or distal causes* (nutritiousness) of certain current incoming interoceptive signals could be taken to be *external* objects and events.

However, note that in order to infer such indirect or distal cause (nutritiousness) from incoming interoceptive signals alone a condition must be met. Such condition is that there must be causal regularities governing the presence of a certain specific external

object (or event) and the occurrence of certain specific physiological changes able to directly trigger interoceptive signals (after transduction). Without such regularities, interoceptive activity simply cannot be informative about external causes. In other words, subjective feeling states have proximal causes, but they can have, under certain condition, indirect or distal causes. Physiological changes are proximal causes, while external events are distal causes. Bodily, interoceptive percepts have then both, physiological causes and its associated external causes. The interoceptive percepts that constitute subjective feeling states inform about distal causes by registering proximal causes, that is, via patterns of physiological changes. Importantly, a certain distal cause has an associated proximal cause that informs about the latter only insofar as there is a causal regularity linking the two (for an analogous approach applied instead to the content of emotion, see Prinz, 2004). To put it this way, "blindfolded", an organism can determine that something external is nutritious only by accessing what interoceptive percepts informs about. In a word, the causes responsible for interoceptive input, and which need to be inferred by interoceptive inferences, can certainly be external.

This granted, it must be recognized that this sort of causal regularities that link specific external objects and specific physiological changes are rare. That is why we are rarely informed about the identity of external objects and events from just feeling our bodies (from felt interoceptive percepts)[15]. Certainly, during the formation of interoceptive percepts, determining (via exteroceptive inference) the identity of the external, indirect cause of incoming interoceptive signals can contribute in assigning prior probabilities to interoceptive hypotheses about the direct physiological cause of such signals. For example, let's say that certain interoceptive signals are caused by certain pattern of physiological activity. *Visually* determining that a likely indirect cause of such signals is the slice of pizza in front of you will increase the prior probability of the hunger-

[15] Certainly, it is ridiculous that an interoceptive percept could inform about the external causes that typically trigger exteroceptive receptors, such as a dog approaching or a loved one dying. "Blindfolded" we cannot tell that there is a pizza on the table by only accessing interoceptive percepts, even though food systematically triggers very specific interoceptive perceptions (those that constitute hunger). The claim is that *nutritiousness* or *dangerousness* (which are external things, even though relational properties (Prinz, 2004, 2007)) are the external things that an already formed interoceptive percept informs about.

hypothesis compared to the thirst-hypothesis. But note that in this sort of case determining the external cause of incoming interoceptive signals occurs via *exteroceptive* inference, so that the proximal, physiological cause of incoming interoceptive signals can be better inferred. Anyway, it might worth taking into account that, as I just mentioned, incoming interoceptive signals can also be indirectly triggered by external objects or events. When causal regularities linking specific external objects and specific physiological changes obtain, the cause to be determined by interoceptive inference can be external (but this is rare).

That said, however, there is an important point to have in mind at this juncture, so as to avoid becoming confused. We are dealing here with the manner by which subjective feeling states are formed. That is, we are *not* dealing with the task of forming an *exteroceptive* percept by making use of the information provided by incoming interoceptive signals or by already formed interoceptive percepts. For example, forming the visual percept of a salad by using interoceptive information relative to hunger so as to determine that incoming visual signals are better explained by the salad-hypothesis than by the grass-hypothesis (since in that way the salad-hypothesis increases its prior probability relative to the grass-hypothesis). This latter kind of task is the one with which, for example, Pezzulo (2014) deals.

On the other hand, we are neither dealing with the non-perceptual task of determining the identity of the cause responsible of the *already formed percepts* that currently populate one's mind. This task consists in *explaining* the origin of one's already formed percepts, rather than forming such percepts in the first place. In this task, one infers the origin of such percepts so as to make sense of them, without shaping their configuration. They become some sort of fixed *explanandum*, to put it that way. The task here is merely explaining or making sense of current bodily, interoceptive experience. Consequently, *an interoceptive percept has already being formed*. Let me exemplify. You are experiencing thirst. That is, you already formed an interoceptive percept. You (or, better, your brain) now try to infer whether that thirst experience (i.e., interoceptive percept) was caused by that last night extra drink or by that salty lunch. Both hypotheses amount to external phenomena. Note that any hypothesis that

gets to be selected will leave the percept in question (thirst) just as already is. Thus, this amounts to a case of explanation (or making sense of) of an already formed percept, rather than to a case of percept formation (the interoceptive percept was already formed). This is analogous to the following case. Imagine you walk into your office one morning and see that things on your desk are not in the order you left them the night before. You consider two explanatory hypotheses: a burglar came in during the night, and the hypothesis that the cleaning guy came in very early this morning. Note that any of the considered hypotheses, if selected, will *not* change the already formed percept of your desk being in such and such a way (your desk looks the same, you are just trying to know what happened that it looks that way). In this sort of cases, inferring causes plays the role of making sense of already formed percepts, rather than the role of forming percepts in the first place[16].

In a word, according to IIA, subjective feeling states arise by minimizing interoceptive PE via interoceptive inferences of the likely causes of incoming interoceptive signals. Those causes belong to the physiological domain, but incoming interoceptive signals can also be indirectly triggered by external objects or events, as long as there are causal regularities that link such external events to certain physiological changes, so that formed interoceptive percept can result informative about external, indirect causes (but this is rare). So, let me insist, what are the causes that must be inferred by interoceptive models? The physiological regularities that obtain in the inner milieu, and, only indirectly, external events systematically related to certain physiological happenings (but this is rare).

3.3.2. Strategies for minimizing interoceptive PE

The aim of this subsection is to systematize the ways in which interoceptive PE can be minimized. This will require systematizing and refining the ways in which the strategies for minimizing interoceptive PE have been described in the literature (Hohwy, 2011, 2013; Seth, 2013; Seth et al, 2012; Seth & Critchley, 2013). This will

[16] Certainly, in this sort of case, once a certain *explanatory* hypothesis has been selected so as to account for an already formed percept, it can influence *further* percept formation. If this occurs, we are not then dealing with the task of merely explaining already formed percepts.

prove to be useful for gaining some clarity on the aspects of the interoceptive process of interoceptive PE minimization that are the most relevant for the view on the nature of valence and emotion to be defended in the coming Chapters. I will show there how different strategies of interoceptive PE minimization serve different functions during emotional episodes. Very roughly, I will argue that the valence component of emotion results from interoceptive PE minimization via perceptual inference; while the main driving component of emotions *per se* consists in interoceptive PE minimization via (external) active inference.

There are three kinds of strategies by which interoceptive PE can be minimized. These kinds of strategies are analogous to the ones used during exteroceptive inference (discussed in Chapter 2), namely, *perceptual inference* and *active inference*. However, in the interoceptive domain, there are two kinds of strategies of active inference available, not just one kind of strategy as in the exteroceptive domain. As it is to be expected, these two kinds of strategies of active inference parallel the distinction made above regarding the ways in which homeostatic imbalances can be rectified. As I discussed above, homeostatic imbalances can be rectified by *internal actions* and *external actions*. The former consist in automatically executing "physiological policies" by making use of resources that are already available within the organism (e.g., centrally generated autonomic control signals). External actions consist in modifying things in the external environment so as to deal with a physiological imbalance. Then, interoceptive PE can be minimized via, what might be called, *interoceptive perceptual inference*, *interoceptive internal actions*, and *interoceptive external actions*. Let me unpack these strategies.

3.3.2.1. Interoceptive perceptual inference

Interoceptive perceptual inference is the strategy I presented above. It is the kind of strategy of interoceptive PE minimization by which interoceptive experience arises. More precisely, by which subjective feeling states arise. Remember that phenomenal consciousness is populated by nothing over and above formed percepts. Thus, interoceptive experience must arise via the process responsible for percept formation:

interoceptive perceptual inference. The task of interoceptive perceptual inferences consists then in forming an interoceptive percept that informs about events in the inner milieu. Interoceptive perceptual inference consists in updating hypotheses so that the generated predictions fit the incoming signal. In this sort of strategy the *direction of fit* is mind-to-world, i.e., finding a hypothesis that successfully models the way the (inner) world is (or, more precisely, a hypothesis that models the ecologically relevant aspects of the way the inner world is).

In this kind of strategy, minimizing interoceptive PE is achieved in the following way. Let's say that certain physiological changes are triggered, such as, for example, an increase in plasma osmolarity levels, a decrease in blood volume, and a decrease in blood pressure, among other physiological changes—these physiological changes amount to the physiological regularities that consistently occur during thirst. They cause interoceptive signals at the lowest levels of the interoceptive hierarchy. Considering that it is already 2pm, and that sounds of cooking utensils come from the neighbour's kitchen, the interoceptive system is already expecting hunger to take place, so it puts forward the hypothesis that it is likely that hunger is causing the incoming signals—remember that at the highest levels interoceptive models encode multimodal and amodal information that allow them to make these links between exteroceptive phenomena (sounds and time of the day) and expected specific interoceptive signals. The interoceptive signals expected for the hunger hypothesis are then generated from the top-down, such as, for example, signals relative to decreasing glucose levels, increasing fatty acids levels, and decreasing body temperature, among other physiological changes. Let's say that even though some interoceptive activity was successfully predicted, interoceptive PE is significant and it exhibits an inferred high precision-weighting. Then, certain incoming interoceptive signals still demand to be 'explained away', to put it this way.

Now, in this kind of strategy, given that PE has not been successfully minimized by the hunger-hypothesis, the interoceptive system must put forward another content-specifying hypothesis that might better match the incoming signal[17]. In other words,

[17] This does not need to occur serially, as different hypotheses can be rehearsed in parallel.

interoceptive models are updated so as to put forward a content-specifying hypothesis that fits the inner world, instead of another competing hypothesis that involves a different content. Let's say that this time the interoceptive system goes for the thirst-hypothesis. It generates then from the top-down the interoceptive activity expected for thirst: signals relative to an increase in plasma osmolarity levels, a decrease in blood volume, and a decrease in blood pressure, among others. This hypothesis achieves to match the incoming signal, so the latter is successfully 'explained away'. Consequently, according to IIA, a subjective feeling state arises. Remember that in the PP framework, the content of the experience or percept that results from successfully 'explaining away' incoming data is specified by the content of the predictions generated from the top-down. Thus, the subjective feeling state that arises in this scenario corresponds to the feeling of thirst, or the percept corresponding to thirst.

3.3.2.2. Interoceptive active inference

Just as in the case of vision, in the case of interoception, PE can be also minimized by changing the input so as to fit the hypothesis (i.e., changing the inner world so as to fit the model), instead of changing the hypothesis to fit the input (i.e., changing the model to fit the inner world), as in the above strategy. That is, interoceptive PE can be also minimized via *active inference*. In this sort of strategy, the direction of fit is world-to-mind, i.e., changing the (inner) world so as to fit the expected activity.

As I mentioned above, in the case of interoception there are two kinds of strategies of active inference, namely, what might be called *interoceptive internal actions* and *interoceptive external actions*.

3.3.2.2.1. Interoceptive internal actions

Interoceptive internal actions operate in the following way. Let me continue with the example of thirst above. As we saw, once the initial incoming interoceptive signals triggered by the physiological regularities systematically involved during thirst are successfully 'explained away' by the thirst-hypothesis, and consequently interoceptive

PE is minimized, the experience of thirst takes place. That is, the felt percept that constitutes thirst is formed.

Remember that such incoming interoceptive signals need to be 'explained away' (so as to minimize interoceptive PE) since this is the manner by which the brain can achieve its distal goal, which is reducing this other quantity, distinct from PE, namely, 'surprisal'. Also remember that 'surprisal' can be taken to be at optimal (low) levels in case the organism stays in ranges of physiological balance. Then, 'explaining away' actual incoming interoceptive signals is the *proximal*, immediate goal that *interoceptive inferences* need to achieve (in all its kinds of strategies); while minimizing 'surprisal', or keeping an overall physiological balance in the long run, is the *distal*, long-run goal that the brain, in *all its functions*, including interoception, needs to achieve.

Now, states of thirst (formed via interoceptive perceptual inference) are incompatible with the distal goal of maintaining homeostasis. Organisms have the hard-wired expectation of stable homeostasis. Interoceptive states of thirst differ from the 'desired' interoceptive states that an homeostatic balance requires. So once an interoceptive percepts that constitutes the feeling of thirst is formed (via interoceptive perceptual inference), *high-level* interoceptive PE is triggered. High-level interoceptive PE consists in the difference now between the expected 'goal state' of homeostatic balance and the current interoceptive percept that constitutes the experience or state of thirst. That is, the interoceptive percept that informs that a certain homeostatic imbalance is taking place. The main task of the interoceptive system is *not* now forming a percept, but rather bringing physiological variables to their expected state by minimizing such high-level interoceptive PE.

Interoceptive internal action consists then in changing physiological inputs so as to fit an interoceptive 'goal state'. In order to achieve this, what I called above 'automatic physiological policies' (roughly, autonomic reflexes) are engaged. The latter make use of resources that are already available within the organism. For example, secreting vasopressin, stimulating the renin-angiotensin-aldosterone system, reducing renal

solute and water excretion, in the case that effective osmotic pressure of plasma increases (thirst).

This is directly analogous to the way active inference works in the case of proprioception so as to bring about physical action or motor behaviour, as stated in Chapter 2. In the proprioceptive case, the difference between a 'desired' proprioceptive state and the current proprioceptive percept is minimized by engaging motor reflex arcs, which bring about actual behaviour. In the interoceptive case, autonomic reflexes are executed so as to minimize interoceptive PE and thus maintain homeostasis. In other words, during interoceptive internal action, top-down interoceptive predictions reflect homeostatic standards relative to which current interoceptive activity is compared. This triggers high-level interoceptive PE. In the PP framework this means that active inference needs to be engaged: actions must be bring forth so as to fulfil predictions. This means that interoceptive models must predict the physiological activity that will help to achieve homeostatic balance. Physiological policies (roughly, autonomic reflexes) transcribe then such top-down interoceptive predictions that reflect homeostatic standards into physiological changes that put the organism closer to homeostatic balance.

Remember that for motor behaviour to arise by way of active proprioceptive inference, proprioceptive PE signals need to be inferred to have low precision weighting. This amounts to the attenuation of attention (sensory attenuation), and is required in order to impede that model revision takes place (which would tramp movement). Analogously, interoceptive internal actions require decreased attention to interoceptive PE, so as to impede model revision that could tramp 'physiological policies' (see, Barret, 2015). In other words, precisions relative to the ascending high-level interoceptive PE must be attenuated to allow top-down, descending homeostatic expectations to 'run the show', so that autonomic reflex could take place.

> "The role of precision and attenuation of interoceptive prediction errors may therefore be fundamental in the organization and selection of autonomic reflexes. In other words, homoeostatic regulation may require the temporary suspension of interoceptive precision—attending away from the current interoceptive state of the body, such that top-down predictions can elicit peripheral sympathetic and parasympathetic reflexes." (Quattrocki & Friston, 2014, p. 419).

Now, interoceptive internal actions, by engaging automatic 'physiological policies', trigger now another cascade of physiological changes aimed at rectifying homeostasis (or to put it this way, interoceptive internal actions trigger 'inner physiological behaviours', to keep the analogy with motor behaviour). These newly triggered physiological changes result in new incoming interoceptive signals that need to be 'explained away' via interoceptive *perceptual* inference. This in order to form a percept that informs the organism how the inner milieu is now faring, after those internal actions (or 'inner physiological behaviours') took place. This cycle of perceptual interoceptive inference and (internal) active interoceptive inference occurs continually as the organism struggles to satisfy the imperative towards maintaining homeostasis (or keeping 'surprisal' levels low).

However, internal interoceptive actions rarely can rectify homeostatic imbalances by themselves (Craig, 2015). They trigger compensatory changes that help the organism to only momentarily deal with the imbalances in question. Think of the case of thirst. We simply lack the physiological resources to re-hydrate ourselves by producing water or some other liquid. Behaviour needs to be engaged in order to find some water and put it into our mouths, to put it this way. Here is when *external interoceptive actions* come into play.

3.3.2.2.2. External interoceptive actions

The difference between expected interoceptive states compatible with homeostasis and the interoceptive percepts that inform the organism about homeostatic deviations can be also minimized via *external interoceptive actions* (allostatic actions)[18]. As I mentioned above, we lack the capacity to rectify physiological imbalances by producing the needed physiological resources by ourselves. So behaviour needs to be motivated. *External interoceptive actions* consist precisely in modifying the external environment, and your situation in it, in order to rectify homeostasis. For example,

[18] These are what Seth (2015a) calls 'allostatic actions'. 'Allostasis' typically refers to the processes by which homeostatic balance is regained through behaviour.

finding a coat in case of a drop in temperature, or going to the fridge for a snack in the case of hunger (these latter motor actions require minimizing proprioceptive PE, as discussed in Chapter 2).

As Seth (2015a) remarks, external interoceptive actions require engaging the higher-levels of the interoceptive hierarchy. At these levels, as I commented in Chapter 2, predictive models are multimodal (and amodal), encoding associations between interoceptive, exteroceptive and proprioceptive information. These stored associations allow predictive models to predict temporal sequences of linked exteroceptive and interoceptive signals (also proprioceptive), which can then be generated form the top-down by modulating the activity of the lower, modality-specific levels of the hierarchy. Stored links between exteroceptive and interoceptive signals allow the system to know which interoceptive states can be obtained via which exteroceptive (and proprioceptive) states. This permits that an action can be found that achieves to cause the expected interoceptive state that homeostasis demands. To put it this way, at higher levels, predictive models know that if effective osmotic pressure of plasma is high (as the thirst-percept achieved via interoceptive perceptual inference informs), and you find water and put it in your mouth (exteroceptive and proprioceptive knowledge), this will likely trigger the interoceptive signals that will eventually match the 'desired' interoceptive goal-state (let's say, normal levels of osmotic pressure) via external interoceptive action. Given the mentioned constant cycle of perceptual and active inference typical of PP, after such an interoceptive action is executed, and thus new interoceptive signals triggered, interoceptive perceptual inference needs to be engaged so as to inform the organism about it current physiological condition. Does it fit the hard-wired goal of keeping 'surprisal' levels low (i.e., maintaining homeostatic balance)?

3.3.3. Counterfactual knowledge of the sensory consequences of action: representations of 'sensorimotor contingencies'

Interestingly, for this kind of inferences to occur, interoceptive predictive models at higher levels require to encode, as Seth (2015a) notes, counterfactual knowledge of

the sensory consequences of action. More precisely, knowledge of the counterfactual relations between, on the one hand, particular exteroceptive states, motors states, and changes to the environment, and, on the other hand, the interoceptive activity that they would ensue. This is what Seth (2014, 2015a) calls 'knowledge of sensorimotor contingencies'.

> "Counterfactually- equipped predictive models encode not only the likely causes of current sensory input, but also the likely causes of fictive sensory inputs conditioned on possible but not executed actions. That is, they encode how sensory inputs (and their expected precisions) would change on the basis of a repertoire of possible actions (expressed as proprioceptive predictions), even if those actions are not performed." (Seth, 2015a, p. 17)

3.3.3.2. 'Active inference for percept formation' vs 'active inference for action'

As we saw in Chapter 2, in the case of vision this kind of counterfactual knowledge is used during what might be called *active inference for percept formation*. This kind of active inference can be distinguished from what might be called *active inference for action*. The latter consist in minimizing proprioceptive and interoceptive PE via classical reflex arcs and autonomic reflexes, respectively. This requires, as we saw, transiently reducing expected precision of ascending proprioceptive and interoceptive PE. Here the task is getting the body/the inner milieu to move in conformity with proprioceptive/interoceptive 'goal-states' or expectations. These tasks presuppose that the hypotheses in question have world-to-mind direction of fit.

In the visual case, *active inference for percept formation* is the kind of active inference that takes place during *perceptual inference* in order to gather new visual samples, so that the best perceptual hypothesis, among competing ones, can be selected. Here the task consist in forming a percept in the first place, rather than bringing about adaptive action. Contrary to the above kind of active inference, this task presupposes that the relevant perceptual hypotheses have mind-to-world direction of fit. Certainly, in the PP framework, motor behaviour always results in new signals that need to be suppressed by perceptual hypotheses, so that there is a constant cycle of perceptual and active inference. However, there is still a distinction to be made between moving so as to put yourself in another place *simpliciter*, once you already know what is

occurring in the world (let's say, you want to get something you saw), and moving for the sake of better determining what is occurring in the world.

The distinction in question is analogous to the distinction between epistemic and instrumental active inference (Seth, 2015a, 2015b).

> Epistemic (active) inference involves selecting actions that we expect to increase the fit between predictive models and hidden causes of sensory signals. This form of inference may characterize, for example, saccadic eye movements or exploratory body movements to inform self-models. (Seth & Friston, 2016, p. 5).

Instrumental active inference consists in the task of controlling sensory activity in line with expectations. This kind of active inference is the one involved in motor behaviour and internal interoceptive actions.

3.3.3.2.1. Types of 'active inference for percept formation': confirmatory, disconfirmatory, and disambiguating actions

Now, Seth (2014, 2015a) identifies three types of action that depend on the kind of stored counterfactual knowledge I mentioned above (i.e., representations of 'sensorimotor contingencies'). In the visual domain, these types of action can be selected so as to gather visual samples that allow to the visual system to put forward the best perceptual hypothesis about the causes of sensory input. The three types of action that Seth identifies amount to actions that gather samples so as to either *confirm*, *disconfirm*, or *disambiguate* hypotheses. For example, given that the visual system has stored counterfactual knowledge, when it is considering that the dog-hypothesis makes fine predictions regarding the flow of incoming signals, it knows that *if* it is a dog what is causing such signals, *and* it performs certain saccades or moves in certain way, the flow of incoming data *should* change in such and such expected way, given its expected precisions of PE. If the flow of incoming sensory information changes in such expected ways, the dog-hypothesis gets *confirmed*. The signals expected for that hypothesis are then generated form the top-down in order to suppress incoming data. Let's say that the latter is achieved: the percept of a dog is formed. But if the flow of incoming sensory information does *not* changes in expected ways, the dog-hypothesis

gets *disconfirmed*, so predictive models need to be updated. On the other hand, the visual system might be considering that both the cat-hypothesis and the dog-hypothesis are equally good candidate hypotheses, given the current incoming data and priors. In this case, given that the visual system has stored counterfactual knowledge, knows that *if* it is cat rather than a dog, *and* performs certain actions, visual signals *should* evolve in such and such a way, and not in these other ways, which are more expected for a dog. In this manner action can be used to *disambiguate* between both such competing perceptual hypotheses.

However, in the case of active inference for *interoceptive* percept formation (or action during interoceptive perceptual inference), these types of epistemic actions seem unlikely. They would require triggering physiological changes as a way of gathering new samples of interoceptive activity. This would generally involve driving our own physiology beyond viability, which is incompatible with the distal goal of keeping surprisal levels low (or, roughly, maintaining homeostasis):

> "[…] it may not be adaptive (in the long run) for organisms to continually attempt to disconfirm current interoceptive predictions, assuming these are compatible with homeostatic integrity. To put it colloquially, we do not want to drive our essential variables continually close to viability limits, just to check whether they are always capable of returning." (Seth, 2015a, p.20)

During interoceptive percept formation, such (dis)confirmatory and disambiguatory sampling 'experiments' are ill-advised. Thus, in order to form an interoceptive percept, the interoceptive system needs to rely just on model updating on the basis of precision-weighted actual interoceptive PE. That is, on the basis of standard interoceptive perceptual inference. This same worry applies not only to active inference for interoceptive percept formation (or action during interoceptive perceptual inference), but it also applies to internal and external interoceptive actions.

However, as we saw above, the fact that (dis)confirmatory and disambiguatory actions do not readily apply to interoception, does not imply that the interoceptive system does not store counterfactual knowledge that links exteroceptive (and motor) and interoceptive states. Counterfactual knowledge of this kind seems to be key for unfolding adaptive *external interoceptive actions* so as to reach expected interoceptive

states compatible with homeostasis. To put it this way, if you want to rectify a subjective feeling state incompatible with homeostasis (e.g., pain, hunger, thirst), you better know under what possible external conditions those states are likely to vanish.

The counterfactual knowledge in question is key, because putting the organism in new exteroceptive and proprioceptive states able to trigger 'desired', expected interoceptive activity (i.e., external interoceptive action) is really the only way to thoroughly minimize the difference between interoceptive 'goal-states' and current interoceptive percepts. Let me insist, internal interoceptive actions can be taken to be momentary compensatory policies, not viable in the long-run. That is why external interoceptive actions are likely to be always motivated, even though in parallel autonomic reflexes (internal interoceptive actions) might be also triggered. That is, external interoceptive actions are likely to be the default strategy for achieving 'desired' interoceptive states. This seems to be the case considering that it is more efficient, in term of energy usage, to exploit the environment rather than depleting body's energy. Moreover, this is also suggested by the fact that, as the case of thermoregulation shows, external interoceptive actions are phylogenetically older than internal actions (Craig, 2015, p.21). Thus, external actions are likely to be automatically motivated, and its selection over internal action does not depend on contextual factors.

Now, when an external (or internal) interoceptive action is engaged so as to achieve an interoceptive 'goal-state', interoceptive *perceptual* inference takes place in order to form an interoceptive percept that tracks the resulting physiological activity. To continue with the example of thirst above, let's say that, for example, the interoceptive system puts forward now the 'satiety-hypothesis'. The interoceptive system generates then the expected signals for that hypothesis in order to 'explain away' incoming data, and it updates hypotheses if needed. This might lead then to another round of *active* interoceptive inference for the purpose of correcting the interoceptive states that are now taking place. Homeostasis then requires the interoceptive system to engage in this constant cycle of perceptual and active inference.

Then, the above begins to suggest that what we feel/perceive consists in the consequences of what we do. More precisely, we feel/perceive the interoceptive consequences of our interoceptive (external and internal) actions.

3.3.4. The brain basis of interoceptive inference

Seth (Seth, 2013, 2015a; Seth et al, 2012; Seth & Critchley, 2013) suggests that interoceptive inferences are mainly realized in a network of regions that likely constitute a (loose) perceptual hierarchy. Among them are regions of the brainstem – such as the nucleus of the solitary tract, periaqueductal gray, locus coeruleus – limbic regions – such as substantia innominata, nucleus accumbens, and the amygdala, and cortical regions. Among the latter, particularly relevant structures are the posterior and middle insula, the anterior insula, the anterior cingulate, and the orbitofrontal cortex (most of them involved in the, so called, 'salience network'[19]).

As I mentioned above, the posterior and middle insula realize a primary interoceptive cortex, analogous to V1 and A1 in the case of vision and audition, respectively (see also Craig, 2015, p.186). The anterior insula (AIC) is particularly relevant for the interoceptive inference view. This region is hypothesized to be the main structure where comparisons between expected interoceptive signals and incoming interoceptive signals take place, constituting a comparator or error-module (Seth et al., 2011). Taking into account that the AIC is considered to harbour phenomenally conscious coherent interoceptive representations of the global condition of the body (Critchley et al., 2004; Singer et al., 2009), this region is a fine candidate for being one of the main regions that realize interoceptive percepts. Moreover, there is evidence showing that the AIC is a region that "integrates" interoceptive and exteroceptive signals. This claim, however, should be taken carefully (that is why I used inverted commas). I think that AIC "integrates" interoceptive and exteroceptive signals only in the sense that it is a site where *associations* (in a broad sense) between interoceptive and exteroceptive information take place, i.e., learned interoceptive activity responses for certain exteroceptive activity. I take this to be the case because the evidence

[19] See Menon V. (2015).

generally used to support the claim that AIC is an "integration" site in that sense consist in evidence simply showing that the AIC gets also activated by exteroceptive stimuli (see Craig (2015) for a complete review). This only supports the claim that AIC harbour associations between certain external stimuli and interoceptive responses—analogously, for example, to the links between exteroceptive stimuli and bodily responses that the ventromedial prefrontal cortex allows in Damasio's (1994) account). This sort of associations likely forms the counterfactual knowledge that is required for external interoceptive actions to take place.

Finally, remember that what I called above internal interoceptive actions involve generating interoceptive signals that modulate the activity of lower interoceptive control regions. This, in order to reach interoceptive 'goal-states'. These kind of actions are realized in regions of the anterior cingulate (ACC). This region together with the orbitofrontal cortex are proposed to be involved in the generation of top-down interoceptive predictive signals. This seems to be the case, for these regions exhibit rich projections to lower-level interoceptive and autonomic control regions, including limbic structures, the brainstem, midbrain, and the hypothalamus. The latter is known to be key for homeostatic regulation.

Similarly, in their EPIC model of predictive interoceptive inference, Barrett & Simmons (2015) identify the mid-cingulate cortex, ACC, posterior vmPFC, posterior OFC, and AI, as the main visceromotor cortical regions involved in generating predictions during (internal) active interoceptive inference and interoceptive perceptual inference. Lower levels of the interoceptive hierarchy include the amygdala, the ventral striatum, the hypothalamus, the periaqueductal grey, and spinal cord nuclei. These latter regions are especially relevant during (internal) active interoceptive inference, as they control internal systems involved in homeostasis. Barrett & Simmons (2015) hold that these regions receive predictions during active inference from the visceromotor cortices, so that they can execute the "interoceptive policies" expected to be useful for the contextual demands. Now, the primary interoceptive cortex, constituted by the mid- and posterior insula, are hypothesized to serve the role of computing the difference between expected interoceptive signals and

actual interoceptive signals (PE), and of propagating the resulting PE signal to higher interoceptive levels of the perceptual hierarchy. The EPIC model emphasizes that AI should not be taken to be the only region responsible for realizing interoceptive percepts, but it is just one of the regions of the interoceptive network mentioned above that contributes in realizing interoceptive percepts.

To sum up, interoceptive percepts represent the constantly evolving positive and negative physiological changes that occur during homeostasis. They are positive or negative relative to standards set by the hard-wired distal goal of maintaining homeostasis (roughly, keeping 'surprisal' low). Interoceptive percepts are likely to be formed by computations that operate under the principles of predictive processing. Subjective feeling states arise as an interoceptive percept is formed. This occurs via interoceptive perceptual inference. In order to minimize the difference between the built-in expectation of homeostasis and the physiological state about which a formed interoceptive percept informs. Such high-level interoceptive PE can be minimized by way of internal and external interoceptive actions[20]. The latter involve instrumental actions that require counterfactual knowledge of the interoceptive consequences of action. All these discussed notions set the main theoretical groundwork needed to unfold the coming discussion relative to the PP account of emotion (and valence). I discuss the interoceptive inference view of emotion in the coming Chapter.

[20] There is something missing in the account of percept formation so far. Features need to be bounded together. Roughly, according to one proposal (Hohwy, 2013), in the PP framework, binding occurs as a result of the sub-personal top-down expectation that the considered features are likely to belong to the same object.

4. The interoceptive inference view of emotion: a critique

As we saw in the previous two Chapters, in the PP framework, all what the brain does, in all its functions, is to minimize its precision-weighted PE. According to PP, visual percepts are formed by minimizing visual PE in a specific manner: via visual *perceptual inference* (Chapter 2). That is, the brain forms visual percepts in a top-down fashion by predicting its incoming lower-level sensory signals from higher-level models or hypotheses of the likely (hidden) causes of those visual signals. A visual percept is formed once a certain content-specifying hypothesis achieves to successfully match, and thus supress, current lower-level visual signals.

In the *interoceptive inference view of emotion* (IIE) (Seth, 2013, 2015; Hohwy, 2013), the principles of PP have been extended to account for emotion. IIE holds that, in direct analogy to the way in which visual percepts are formed (Seth, 2015; Seth & Friston, 2016), emotions arise from interoceptive predictions of the causes of current interoceptive afferents. In other words, emotions amount to interceptive perceptions formed via higher-level, content-specifying emotion-hypotheses. Emotions result then via *interoceptive perceptual inference*, just as visual percepts result via visual perceptual inference.

In this Chapter, I will argue that IIE is problematic. I will show that IIE is committed to the assumption that there must be different regularities pertaining to different emotion types in the physiological domain. However, this is unlikely to be the case. Therefore, it is unlikely that emotion models get to encode interoceptive expectations in the way required by IIE.

The PP framework seems then to lack the resources to account for emotion. Emotions are at the core of our mental life. If PP cannot account for that aspect of mentality, it simply fails as an overarching, unifying framework in cognitive science. This is a

major drawback for PP ambitions. However, in Chapter 7, I will suggest that, *if* PP is on track, interoceptive PE minimization can indeed account for emotion. However, this demands amending IIE in a key respect. I will briefly suggest that emotions do *not* arise via interoceptive *perceptual inference* (as IIE claims), but rather they arise by minimizing interoceptive PE in another fashion. That is, emotions arise via *external interoceptive active inference*. This proposal avoids the problematic assumption of IIE. Therefore, if the proposed view holds, PP's ambitions are safe: interoceptive PE minimization can account for emotion.

In what follows, I present IIE (Section 4.1). In Section 4.2., I show that IIE is indeed committed to the assumption that there must be regularities pertaining to emotion in the physiological domain. In Section 4.3., I argue that such an assumption is likely not to be the case. Then, in Section 4.4., I reply to possible objections regarding the actual commitments of IIE. Among them, the objection that IIE is not committed to the claim that there must be different regularities pertaining to different emotion types in the physiological domain, since IIE should be read as claiming that all emotion hypotheses expect the *same* physiological states. I also deal with the duck/rabbit objection, and with the possible objection that IIE is not committed to the claim in question, because IIE should be read as holding that emotions consist in an 'amalgam' of several multimodal states. I conclude this Chapter with some final remarks (Section 4.5).

4.1. The interoceptive inference view of emotion

PP is already doing explanatory work in a wide variety of psychological domains. However, PP was conceived and developed as an account (and re-conceptualization) of *perceptual* processes. That is why its principles have been mainly applied in the explanation of mental phenomena that, in some way or another, can be readily understood as perceptual in nature – e.g., visual perception, binocular rivalry, illusions and delusions, etc. (for a review, see, e.g., Clark, 2013; Friston, 2005, 2009).

According to the Jamesian view of emotion (James, 1884), emotions can be understood as *perceptions* of bodily, interoceptive changes. Considering that the Jamesian view that emotions can be understood as a perceptual process has recently seen a resurgence of interest in emotion research (e.g., Prinz, 2004), an obvious next step for PP's explanatory ambitions is to apply its principles in accounting for emotion.

To date, there is no fully developed PP account of emotion on offer. However, A. Seth (Seth, 2013; Seth et al, 2012; Seth & Critchley, 2013) and J. Hohwy (2013, pp. 242-244; see also Hohwy, 2011) have recently offered a first sketch of how such extension might go. Taking into account the fact that PP mainly works as an account of perception, these first sketches have suggested a PP version of the perceptual, interoceptive view of emotion. According to this kind of view, emotions are *perceptions* of distinct physiological changes. These first PP accounts of emotion then see emotion as arising from *interoceptive perceptual inferences*. This view might be called the 'interoceptive inference view of emotion' (IIE).

According to IIE, emotions arise by minimizing interoceptive PE. However, in IIE, emotions arise by minimizing interoceptive PE in a specific sort of way. According to IIE, *in direct analogy* to the way in which visual percepts are formed (Seth, 2015; Seth & Friston, 2016), emotions arise from interoceptive predictions of the causes of current interoceptive afferents. For an emotion to arise, emotion models/hypotheses need to suppress from the top-down incoming interoceptive signals.

In order to achieve this, analogously to the case of vision (Chapter 2), emotion models generate expected data in lower layers of interoceptive processing. Informally, emotion models can be seen then as putting forward hypotheses about which interoceptive activity is more likely given a certain emotion instead of another emotion – e.g., "if it is fear, instead of anger, I expect more likely such and such interoceptive signals, and not these other ones, because fear tends to cause such signals in contexts like this one. Let's generate those signals. Do they match incoming signals?"

Mismatches between predicted interoceptive signals and actual interoceptive signals result in interoceptive PE, which causes to replace the considered perceptual interoceptive hypothesis. An emotion arises once a certain emotion hypothesis fits incoming interoceptive signals, and thus it minimizes interoceptive PE. In other words, an emotion is generated as an interoceptive percept is formed, i.e., an interoceptive percept that is formed driven by an emotion-hypothesis. Just as visual percepts result via visual *perceptual inference*, emotions result via *interoceptive perceptual inference*. In this view, the content of a certain high-level emotion model/hypotheses—to put it in this way, 'the anger-hypothesis' or 'the fear-hypothesis'—determines the content of the interoceptive percept that is formed, and consequently, it determines the content of the bodily experience that ensues, which according to IIE, constitutes the experience of emotion—one experiences anger or fear, to keep the example from above.

Emotion hypotheses then shape interoceptive percepts from the top-down. This solves, analogously as in the case of vision mentioned above, the underdetermination between emotion types and physiological input. In this sense, emotion differentiation can be explained in terms of the content of the high-level hypotheses (e.g., anger-hypothesis vs fear-hypothesis) that are brought to bear on the modulation of interoceptive perceptions (see Hohwy, 2013).

This aspect of IIE makes it a particularly interesting account of emotion. Insofar as it incorporates high-level knowledge into interoceptive perception, and claims that the content of interoceptive experience is determined by the content of higher-level emotion models, IIE puts together key insights of both, Jamesian and two-factor, Schachterian views of emotion[21].

[21] Roughly, Jamesian views of emotion holds that emotions amount to bodily perceptions, while Schachterian views hold that emotions amount to cognitive interpretations of current bodily experience ('arousal' for Schachter). Thus, in Schachterian views, emotions require one more 'factor' than Jamesian views. Schachterian views exhibit two-factors: bodily perception plus 'cognition'.

However, note that the claim that high-level emotion knowledge shapes interoceptive perception does not make IIE a strictly two-factor, Schachterian view. This is the case because, in the latter kind of view, 'cognition' has the function of merely explaining or making sense of current bodily, interoceptive experience. Consequently, *an interoceptive percept*, which in Schachterian views, and contrary to Jamesian views, is ambiguous concerning a specific emotion type, *has already been formed*. Thus, in strictly two-factor, Schachterian views, contextual knowledge only merely explains (or makes sense of) an already formed percept, without shaping it or playing a role in the formation of such percept, as a PP perceptual view must claim.

Now, importantly, analogously to visual percept formation, finding a fitting interoceptive hypothesis requires that the whole system constantly contributes to that task across all levels of the interoceptive hierarchy. Thus, contrary to other views that see emotion as arising from bodily perception (e.g., Prinz, 2004), IIE explicitly recognizes the constant influence of other modalities and high-level amodal and multimodal knowledge in feeding interoceptive hypotheses. Such non-interoceptive activity modulates, across all levels of the perceptual hierarchy, the interoceptive predictions involved in the task of suppressing incoming interoceptive signals. In this sense, the generation of an emotion can be seen as typically involving an 'amalgam' (Clark, 2016, p.234) of different kinds of information. The latter must be seen as contributing in finding interoceptive data able to explain away incoming *interoceptive* signals, since emotions "arise as *interoceptive* prediction error is actually explained away" (Hohwy, 2013, p. 243) (more on this below).

Interoceptive models need to be integrated with high-level amodal and multimodal information for the purpose of linking interoceptive responses with the external and contextual cues (e.g., a snake approaching) that could be causally relevant in bringing about incoming interoceptive signals. We can assume that such integration helps fix a context that can be used as priors for issuing a proper interoceptive hypothesis. Determining that some event in the external environment is causally relevant in bringing about low-level interoceptive activity might be useful in choosing between

competing interoceptive hypotheses. For example, if an individual has two interoceptive hypotheses with the same posterior probability about the emotion that might be causing certain interoceptive signals. Let's say the fear-hypothesis and the excitement-hypothesis. In this case, she could decide between the two by exteroceptively and cognitively determining the nature of the context in which she finds herself: Does the context make more likely the fear-hypothesis or the excitement-hypothesis? Let's say that, in the case in question, the individual sees that a snake is approaching. Thus, the interoceptive hypothesis for fear acquires higher posterior probability: the interoceptive signals expected for the fear-hypothesis are generated form the top-down, interoceptive PE is appropriately minimized, and fear is perceived and experienced.

However, note that during this process of interoceptive PE minimization, amodal and multimodal knowledge serve *interoceptive* ends: finding the better interoceptive hypothesis. Such kind of knowledge contributes in influencing the interoceptive data that eventually will meet activity in the lower layers of the interoceptive channel, so as to minimize interoceptive PE. In other words, forming an interoceptive percept requires meeting incoming *interoceptive* signals alone, and this is achieved by generating from the top-down *interoceptive* signals—i.e., signals proprietary of the visual channel won't do the job of suppressing *interoceptive* incoming data. Remember that IIE's claim is that, in direct analogy to visual percept formation, an emotion arises when *interoceptive* PE is minimized via interoceptive perceptual inference. This requires 'explaining away' incoming data in the proprietary code of the interoceptive channel alone. Generally speaking, in the PP framework different modalities and high-level amodal knowledge provide background and contextual knowledge for generating a successful perceptual hypothesis that will result in a coherently well-formed percept in one specific modality. So, just as visual hypotheses might be disambiguated, for example, by *auditive* information—e.g., a bark might help decide in favour of the *visual* dog-hypothesis instead of the *visual* wolf-hypothesis—interoceptive hypotheses might be disambiguated by information provided by other modalities and background knowledge. In the interoceptive inference view of emotion, cognitive and multimodal knowledge contributes in disambiguating hypothesis for *interoceptive* percepts.

Note that for a certain external cue (a snake approaching) to modulate interoceptive activity, the system must built an association between such an exteroceptive, contextual cue and certain interoceptive activity (to take the example above, a snake approaching and the interoceptive activity relative to fear instead of the interoceptive activity relative to another emotion type), so as to be able to predict such interoceptive activity from this external cue.

Considering the point remarked by condition *C* above (Section 2.7.), in order to learn such a 'prediction-enabling' association, the system must extract from the world the regularity linking the event *snakes approaching* (or *danger*, more generally) and the triggering of certain patterns of physiological changes. In other words, the event *snakes approaching* (or *danger*, more generally) must predict those patterns of physiological changes (i.e., they must not be probabilistically independent phenomena). This also applies to relatively slower time-scale or abstract regularities.

Now, according to Seth, top-down interoceptive signals are mainly issued from the anterior cingular cortex (also the orbitofrontal cortex is involved). Matches and mismatches (prediction error) between autonomic afferents and interoceptive inferences are calculated in the anterior insular cortex, a visceral sensory region associated with interoceptive awareness (Craig, 2002). In other words, the brain network responsible for generating emotions through interoceptive inference amounts to what is known as the *salience network*, which also includes other paralimbic structures such as the amygdala and the inferior frontal gyrus (Seth et al., 2012). That is, in the interoceptive inference view, emotions are anchored in structures usually associated with autonomic representation and visceral sensory processes.

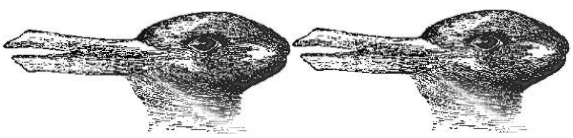
Hohwy (2013, pp. 242-244) has proposed an extension of PP to emotions along the same lines as the one proposed by Seth. Both propose an interoceptive inference theory of emotion. Following James (1884) and Prinz (2004), Hohwy holds that emotional

experience arises from inferences to the causes of the type of interoceptive signals that are prompted in a given context. Here emotions are "reduced to basic interoceptive states" (Hohwy, 2013, p.243) and our perception of them: "emotion arises as a kind of perceptual inference on our own internal states." (*Ibid*.) In other words, emotions result from interoceptive inferences:

> "The key is to view interoceptive signals as internally generated and highly ambiguous prediction error, which is explained away by interoceptive expectations in a hierarchical manner that can draw on all sorts of statistical regularities in the overall prediction error landscape." (Hohwy, 2013, p.243)

According to Hohwy, interoceptive signals are ambiguous between hypotheses about the world. As in Seth's proposal, background knowledge helps here in choosing between competing hypotheses. In this respect, Hohwy offers as an example an analogy with visual perception. In the case of the double duck-rabbit (Fig. 1), received or activated knowledge that the duck is eating the rabbit disambiguates visual signals in favour of that hypothesis. In a PP context, this means that knowledge of that extra piece of information results in the top-down generation of the visual signals corresponding to that winning hypothesis (and thus incoming visual data is better explained away), instead of the top-down generation of the visual signals corresponding to a competing hypothesis (let's say, one rabbit following another rabbit). As a result of that, one forms and experiences the visual percept of the duck eating the rabbit, and not the visual percept of one rabbit following another rabbit.

Fig. 1



In the case of emotion there is an analogous story. Let's assume that in a certain emotionally significant context certain interoceptive signals are prompted. The

emotion system considers then two hypotheses, namely, the fear-hypothesis and the joy-hypothesis. Receiving or activating an extra piece of contextual knowledge—for example, that your boss is approaching, and that she tends to be threatening—makes the fear-hypothesis more likely. Consequently, the interoceptive data for *that* hypothesis is generated from the top-down. Let's say that interoceptive PE is successfully minimized by such data, so fear then "[…] arises as interoceptive prediction error is actually explained away" (Hohwy, 2013, p. 243)[22].

To sum up, according to IIE, emotion arise when a content specifying emotion-hypothesis minimizes interoceptive PE across the whole interoceptive hierarchy. This requires that amodal and multimodal models constantly contributes to that task across all levels of the cortical hierarchy. However, as Hohwy remarks, in IIE, emotions are "reduced to basic interoceptive states" (Hohwy, 2013, p.243) and our perception of them: "emotion arises as a kind of *perceptual inference* on our own internal states." (*Ibid*. italics are mine). Emotions result then via interoceptive perceptual inference, just as visual percept result via visual perceptual inference. Emotions are interoceptive percepts which are formed guided by a content-specifying emotion-hypotheses.

4.2. IIE is problematic

As it stands, IIE exhibits a problematic key assumption. Remember that in the PP framework, the generative models from which hypotheses are put forward extract the regularities that configure the hierarchical structure of the world. In other words, priors are learned from experience (via model/hypothesis selection and revision in light of precision-weighted PE), and over time they manage to recapitulate the causal regularities in the world at its different time-scales (Hohwy, 2013). This is what allows

[22] See Pezzulo (2014) for a proposal on how the interoceptive inference approach is relevant for standard cognitive and exteroceptive inferences.

models to issue successful predictions about the worldly (hidden) causes of its incoming signals.

Now, according to IIE, emotions result from *interoceptive* predictions—more precisely, via interoceptive perceptual inference. Where do interoceptive priors come from? From the causal regularities that obtain in the inner physiological world, as patterns of changes in the physiological landscape are the kind of thing that causes interoceptive input (this is analogous to the platitude that visual priors come from regularities involving light-reflecting objects). Thus, IIE is committed to the assumption that there must be causal regularities pertaining to emotion in the physiological domain.

However, as I will argue below (Section 4.1.), evidence points to the claim that there are no regularities pertaining to emotion in the physiological domain (emotion types and patterns of physiological changes are statistically independent phenomena). There are no distinct bodily, physiological regularities relative anger, fear, joy, sadness, etc. In other words, evidence strongly suggests that there is no significant causal regularity connecting emotion types and patterns of physiological changes, so that a certain emotion type could predict physiological patterns. This is the case simply because, contrary to perceptual, interoceptive theories of emotion (Prinz, 2004; James, 1884), the physiological landscape does not configure emotions. There are no emotions in the inner milieu. Considering that there are no emotions in the physiological landscape, it is unlikely that the brain stores expectations about what *interoceptive* signals to expect given a certain emotion type. In other words, taking into account that there are no emotions configured in the physiological landscape, the learning of priors relative to which interoceptive signals to expect given a certain emotion-hypothesis seems unlikely. There are no such regularities to extract so as to build the relevant interoceptive expectations. Emotion models (which I am safely taking to be high-level) must encode then, primarily, expectations about *other* sort of information. Therefore, it is unlikely that emotion models get to encode interoceptive expectations in the way required by IIE. That is, in the way that interoceptive perceptual inference demands—

the experience of emotions does not arise by minimizing interoceptive PE by generating *interoceptive* signals from emotion-models. IIE should be amended.

This argument is analogous to the following, more familiar argument. Considering that there are no regularities pertaining to cloud-types in the auditive domain (or in the 'detectable vibrations domain')—i.e., there are no auditive regularities pertaining to cirrus, nor to cumulus, stratus, etc.—it is unlikely that cloud-models encode auditive expectations. Without clouds in the auditive domain, it is hard to see how cloud models could get to build expectations about which auditive signals to expect given a certain cloud-hypothesis. Or to put it this way, given that there are no regularities pertaining to clouds in the auditive domain, the experience of clouds does not arise, primarily, by minimizing auditive PE by generating auditive signals from cloud-models. If this argument works in this cloud case, it should also work in the emotion case.

Let me insist. Remember that, as I commented in Section 2.7., a learnt 'prediction-enabling' association between a certain piece of high-level knowledge about *H* and certain low-level modality specific sensory activity about *L* is required for that piece of high-level knowledge to strongly modulate, and thus shape, percepts in such low-level modality specific layers of processing. This sort of associations recapitulates regularities in the world. Considering that this sort of associations are built from experience, by extracting regularities in the world, their formation requires that there must be a causal regularity governing *H* and *L*, so that the former predicts the latter. Now, evidence suggests that there are no regularities governing emotion states and physiological patterns, so that the former could predict the latter (see below, Section 4.1.). Thus, it is unlikely that emotion-models (which I am safely taking to be high-level) store expectations about which interoceptive signals to expect given a certain emotion type. It is unlikely then that emotion-models shape interoceptive percepts, so that the latter could constitute the experience of a certain emotion.

Let me put the argument in question in slightly different terms. The sensory predictions about a certain object (or event) *O* that the model for *O* encodes are learned from exposure to *O* during training (i.e., experience). Then, if in a certain domain there are no *O*'s, a model cannot get to encode sensory predictions about *O*'s for that domain. There are no emotion types configured in the physiological/interoceptive domain. Thus, it is unlikely that emotion-models (which I am safely taking to be high-level) encode interoceptive predictions regarding emotion types—i.e., it is unlikely that emotion-models are simply in the business of predicting interoceptive signals via interoceptive perceptual inference. Emotion-models, from which it is assumed that emotions arise, must then encode, primarily, other sort of information. Below I discuss the evidence that points to the claim that there are no regularities pertaining to emotion in the physiological domain. There are no emotions to be found in the physiological landscape.

4.3. The evidence

Against some versions of the 'natural kinds' view of emotion, L.F. Barrett and her colleagues have compellingly made the case for the claim that the physiological landscape does *not* exhibit "distinctive sets of correlated properties" (Barret, 2006b, p.33) that could configure anger, fear, joy, etc. That is, there are no physiological response patterns that instantiate regularities pertaining to emotion (see, e.g., Barrett, 2006b; Quigley & Barrett, 2014). In this respect, this is mainly the case since statistical analyses of meta-analytical studies on emotion evince that there is no robust specificity in autonomic activity measures across emotion studies. This is not the place to unfold this one-hundred-years-controversy in any detail, so I refer the reader to Barret's work on the matter. However, it is worth mentioning that, within philosophy, her arguments to the effect that there are no regularities pertaining to emotion *per se* in the physiological domain have been widely taken to support this conclusion (e.g., Carruthers, 2011; Ritchie & Carruthers, 2015; though see Colombetti, 2014, pp.35-36). Even influential sympathizers of the view that emotions are biological natural kinds, which have autonomic 'signatures', such as Scarantino (Scarantino, 2009;

Scarantino & Griffiths, 2011), have recognized that Barrett (2006b) indeed achieves to show that the conclusion in question is the case. Scarantino (2009) recognizes that Barrett has shown that there are no physiological response patterns that instantiate regularities relative to the kind of mental states that *we take emotions to be* (anger, fear, joy, etc.)—i.e., the kind of mental states that, as we saw in Chapter 1, typically constitute the *explandum* of an emotion theory[23].

It could be objected that meta-analytical studies allow us to draw conclusions only for *individual* measures of autonomic responses (see Cacioppo, 2000). Meta-analytical studies then have little impact on current versions of interoceptive theories of emotion (e.g., Prinz, 2004). This is the case, since a key claim defended by these views is that, in the physiological landscape, emotions are configured by several, *patterned* physiological responses. Is there positive evidence for this claim?

To date, the most detailed defence of the interoceptive theory of emotion is due to Prinz (2004). Prinz only presents one piece of evidence as evidence for the claim in question, namely, the study of Levenson, Ekman, and Friesen (1990). This study is then particularly relevant for the claim in question. Let me briefly discuss it.

This study aimed at finding whether there is physiological differentiation between, so-called 'basic emotions'. In the study, subjects were instructed to produce facial configurations previously associated with specific emotional states, and to report which emotion they experienced (if any). At the same time, physiological variables such as heart rate, finger temperature, somatic activity, and skin conductance were jointly monitored. They found that the reported emotions correlated with autonomic

[23] Scarantino goes on to propose, however, that emotion research should change the explanatory target of emotion theories. Emotion research should not have as explananda the mental states that we take emotions to be. Instead, emotion research should find another explanandum, though similar to the mental states that we take emotions to be. However, I think this is not a satisfactory move, as it attempts to offer poor substitutes of the mental states that we want to understand in the first place (see Dennett, 2009). It simply changes the subject. Anyway, this controversy cannot be resolved here.

differences both between negative and positive emotions, and among the negative emotions of anger and fear. Differences in autonomic activity between negative and positive emotions are not much interesting, because they are straightforwardly explained by the fact that valence properties are distinctly realized in the autonomic system (Russell, 2003; Barrett, 2006a). Thus, they only show that there is a distinctive pattern of physiological changes for valence properties. However, differences in autonomic activity among negative emotions, which share valence properties, elude such straightforward explanation. Accordingly, the reported autonomic differentiation between anger and fear is the truly interesting finding on which to focus.

The results in question are not as suggestive of the truth of the claim that there are emotions configured in the physiological landscape as it might appear. There is an alternative explanation, not considered in the study. The supposedly observed emotion-specific autonomic patterns for fear and anger can also be explained by another hypothesis. Anger and fear differ not only in that they are different emotions, they also differ in a key respect: they differ in, what might be called, 'coping potential properties', namely, preparatory activity for threat or challenge behaviour. Nothing in the study in question discards the possibility that the observed results are driven by threat and challenge rather than by fear and anger (see Barrett, 2006b).

> "The vascular patterns that differentiate anger and fear also distinguish between threat and challenge appraisals (e.g., Blascovich & Mendes, 2000; Mendes, Reis, Seery, & Blascovich, 2003; Tomaka et al., 1993, 1997). In attempting to generate fear and anger, researchers may have inadvertently manipulated threat and challenge appraisals, leading to the observed vascular effects. Similarly, skin conductance reactivity is associated with increased attention allocation (Blakeslee, 1979; Frith & Allen, 1983), and may have little to do with emotional responding per se." (Barrett, 2006b, p.41)

Now, challenge and threat *cut across* "basic emotions". Therefore, the observed autonomic activity in question does not distinguish between "basic emotions" *per se*.

On the other hand, it is worth considering that autonomic activity measures differ from occasion to occasion depending on the method used to trigger emotions in the lab (see also, Larsen et al., 2008). Variation in autonomic patterns across studies which differ in their triggering methods is the norm. For example, the patterns of autonomic activity found in the study of Christie & Friedman (2004), which used films as triggering method, vary significantly to the patterns found in the study of Nyklicek et al. (1997), which used music as a triggering method. The study of Stemmler et al. (2007) is particularly interesting in this respect. Stemmler et al. work looked for specificity in autonomic patterns for anger vs. fear. Also, they used two different triggering methods (in two different studies). No robust consistency across both studies was found (for a critical review of this kind of studies, see Quigley & Barrett, 2014).

Now, as Prinz (2004) (and others) emphasizes, physiological resources are allocated according to the kinds of behaviours that the faced situation affords. That is, physiological responses prepare for action. This is a widely acknowledged phenomenon in emotion research. In light of this phenomenon, the above mentioned inconsistency in results across studies makes plenty of sense, considering that it is not implausible to assume then that certain triggering methods tend to relate to certain kinds of situation more than other triggering methods. This points to a more telling argument against the view that the physiological landscape configures emotions.

Physiological resources are allocated according to the nature of the behaviour that the faced situation demands. Then, if different emotions usually differ in their associated behaviours, but they also usually involve the same kind of behaviour (because not only the same type of situation, but also different types of situations might demand the same kind of behaviour), and the same type of emotion usually involves quite different behaviours, it is expected that the different emotion types are not distinctively configured by patterns of physiological responses. Further studies are needed to resolve this issue, but current evidence, taken as a whole, does not support the view that there are regularities pertaining to emotion in the physiological domain.

Consequently, it should not be taken for granted when theorizing about emotional processes.

There are no emotion types configured in the physiological/interoceptive domain. There are no regularities governing emotion states and physiological patterns, so that the former could predict the latter (i.e., they are probabilistically independent). Thus, it is unlikely that emotion-models store expectations about which interoceptive signals to expect given a certain emotion type. It is unlikely then that emotion-models shape interoceptive percepts, so that the latter could constitute emotion. Emotion-models, from which it is assumed that emotions arise, must then encode other sort of information.

4.4. Reply to possible objections

4.4.1. On the causes of interceptive signals

As I commented above, IIE is the view that, in direct analogy to the way in which exteroceptive percepts are formed, emotions arise by minimizing interoceptive PE via perceptual inferences of the likely *causes* of incoming interoceptive signals. Thus, we must assume that instead of determining what it is the object or event in the *external* world that is likely causing incoming signals, as in the case of *exteroceptive perceptual inference*, during *interoceptive* inference the causes that need to be determined amount to the causal regularities that obtain in the *inner*, physiological landscape, as patterns of changes in the physiological landscape are the kind of thing that causes interoceptive input (I take this to be a platitude).

Now, as long as exteroceptively represented external objects and events can consistently cause physiological changes via learned associations between a certain

external cue and a certain physiological pattern (e.g., seeing food can cause the physiological changes characteristic of hunger), the *indirect or distal causes* of certain current incoming interoceptive signals could be taken to be *external*. Then, it could be argued that the task of interoceptive perceptual inferences is to infer which emotionally relevant *external event* is causing incoming interoceptive signals. As such an external cause is successfully inferred, its expected interoceptive signals are generated from the top-down, meeting, and thus 'explaining away', incoming interoceptive signals. This gives rise to the formation of an interoceptive percept: emotion experience thus arises.

In this respect, think of the distinction made by Prinz (2004) between the *real* and *nominal* contents of emotion. As IIE, Prinz identifies emotions with bodily perception. Real contents are external things ('core relational themes': danger, loss, etc.) that cause emotions; while nominal contents amount to physiological patterns that cause emotions. Physiological changes are *proximal causes*, while external events (core relational themes) are *distal causes*. Bodily, interoceptive percepts have then both, physiological causes and its associated external causes. The interoceptive percepts that constitute emotion inform about real contents by registering nominal contents, that is, via patterns of physiological changes. Importantly, a certain real content has an associated nominal content that informs about such real content only insofar as there is a causal regularity linking the two (see Chapter 1). According to Prinz, the mind/brain can determine then that a certain external situation is dangerous by forming an interoceptive percept that informs the organism about danger. In other words, 'blindfolded', an organism can determine that an external situation is dangerous only by accessing what interoceptive percepts inform about. In a word, the (distal) causes responsible for interoceptive input, and which need to be inferred by interoceptive inferences, can be external.

A defender of IIE could argue then that it does not matter that there are no emotions in the inner physiological landscape, since interoceptive perceptual inferences are in the business of inferring the *external causes* of interoceptive signals, and there are external situations that can be said to be characteristic of a certain emotion instead of

another emotion. To put it this way, the signature of an emotion type lies in the external environment. For example, an approaching *dangerous* dog is characteristic of fear, while the *loss* of a loved one is characteristic of sadness[24]. Certainly, it is ridiculous to claim that an interoceptive percept could inform about the external causes that typically trigger exteroceptive receptors, such as a dog approaching or a loved one dying. 'Blindfolded' we cannot tell that there is a pizza on the table by only accessing interoceptive percepts, even though food systematically triggers very specific interoceptive perceptions (those that constitute hunger). The claim is that *nutritiousness* or *dangerousness* (which are external things, even though relational properties (Prinz, 2004, 2007)) are the external things about which an already formed interoceptive percept informs.

Then, in the reading of IIE under consideration, the task of interoceptive perceptual inferences is to infer which *external event* characteristic of emotion (e.g., core relational themes: *danger, loss,* etc.) is causing incoming interoceptive signals.

There is an important point to have in mind at this juncture, so as to avoid becoming confused. We are dealing here interoceptive perceptual inference, which according to IIE is the manner by which emotions arise. In direct analogy to visual perceptual inference, the task of interoceptive perceptual inferences consists in forming a percept. In this case, an interoceptive percept. That is, we are *not* dealing here with the task of forming an *exteroceptive* percept by making use of the information provided by incoming interoceptive signals or by already formed interoceptive percepts. For example, forming the visual percept of a salad by using interoceptive information relative to hunger so as to determine that incoming visual signals are better explained

[24] However, as I mentioned above, as Prinz (2004) has argued, an interoceptive theory of emotion needs to say that what makes an external situation *dangerous* (or any other property that characterizes an emotion type, such as *loss* in the case of sadness) is that the representation of that external situation is accompanied by the perceived physiological changes that constitute fear. This perceived pattern of changes is what makes that that situation is perceived as *dangerous* in the first place. Thus, in an interoceptive theory of emotion, *being dangerous* is not something that is determined exteroceptively. For expository purposes, I will ignore this problem for the defender of the objection in question.

by the salad-hypothesis than by the grass-hypothesis (since in this way the salad-hypothesis increases its prior probability relative to the grass-hypothesis). This latter kind of task is the one with which, for example, Pezzulo (2014) deals (see Pezzulo, 2014).

On the other hand, we are neither dealing with the *non-perceptual* task of determining the identity of the cause responsible of the *already formed percepts* that currently populate one's mind. This task consists in *explaining* (or making sense of) the origin of one's already formed percepts, rather than *forming* such percepts in the first place. In this task, one infers the origin of such percepts so as to make sense of them, without shaping their configuration. They become some sort of fixed *explanandum*, to put it this way. The task here is merely explaining or making sense of current bodily, interoceptive perceptual experience. Consequently, *an interoceptive percept has already been formed*. Let me exemplify. You (or better, your brain) are undergoing the experience of cold. That is, you (or better, your brain) already formed an interoceptive percept. That is, an interoceptive perceptual hypothesis was already successful in minimizing incoming interoceptive PE. You (or better, your brain) now try to infer whether that cold experience (i.e., interoceptive percept) was caused by your drinking a lot of cold water a few minutes ago, or by that breeze of air conditioning in the library. Both hypotheses amount to external phenomena. Note that whether the cold-water-hypothesis or the air-conditioning-hypothesis turns out to be selected as the better hypothesis, both leave the configuration of the interoceptive percept in question (cold) just as already is. Thus, this consists in a case in which a hypothesis is selected in order to *explain* (or *make sense* of) an already formed interoceptive percept; rather than to a case in which the task is attempting to determine what is going on in the inner milieu in the first place, since a satisfactory perceptual hypothesis was already found (an interoceptive percept was already formed). This is analogous to the following case. Imagine that you wake up very early in the morning and your partner is not right next to you on the bed. This is very rare. It demands explanation. You then consider two possible explanatory hypotheses. The hypothesis that your partner had an early meeting and did not tell you, and the hypothesis that your partner went to the supermarket to buy milk for the breakfast. Note that any of the considered hypotheses,

if selected, will *not* change the already formed percept of your empty bed. The bed looks the same, you are just trying to make sense of this rare situation. This is the kind of case that I commented above (p. 139) regarding what I referred to as strict two-factor, Schachterian views.

Then, in the reading of IIE under consideration, the task of interoceptive perceptual inferences is to infer which *external event* characteristic of emotion (e.g., danger) is causing incoming interoceptive signals. As such an external cause is successfully inferred, let's say *dangerousness*, its expected interoceptive signals are generated from the top-down by the fear-model, meeting, and thus 'explaining away', incoming interoceptive signals. This results in the formation of an interoceptive percept that informs about *dangerousness*: the experience of fear thus arises.

Note that for a hypothesis about something external to be able to 'explain away' incoming *interoceptive* signals by generating the latter (and thus form an interoceptive percept), it must encode expectations about which interoceptive activity to expect given that hypothesis instead of another hypothesis. The danger-hypothesis must encode interoceptive expectations which are distinguishable relative to the loss-hypothesis. More generally, for a hypothesis about something external (danger) to modulate interoceptive activity, the system must built an association between such an exteroceptive cue and certain interoceptive activity (instead of other interoceptive activity), so as to be able to predict such interoceptive activity from this external cue.

Where do such interoceptive priors come from? As I mentioned in Sections 4.2. and 2.7., in order to learn such a 'prediction-enabling' association, the system must extract from the world the regularity linking such an external phenomenon (danger) and the triggering of certain patterns of physiological changes able to directly trigger interoceptive signals (after transduction). In other words, *danger* must predict those patterns of physiological changes—remember the similar point made by Prinz that a

certain real content has an associated nominal content that informs about the latter only insofar as there is causal regularity linking the two.

Considering the above, this reading of IIE cannot work, and for the exact same reasons discussed in Section 4.2. There are no regularities linking the external events that characterize emotions (loss, danger, etc.), and certain pattern of physiological changes (instead of other patter of physiological changes), so that the latter could be predicted from such external cues. That is, the argument and the evidence from Section 4.2. and 4.3. also apply in this case. The evidence from Section 4.3. directly apply to this case, since the external events that characterize emotions (loss, danger, etc.), precisely insofar as they characterize emotions, are the kind of stimuli used for triggering emotions in the lab in the studies discussed Section 4.1.

4.4.2. The double duck-rabbit objection

There is another similar objection, closely related to the objection above. Think of the double duck-rabbit case (fig.1) (see Hohwy, 2013, p. 243). In this case, one single kind of very ambiguous stimulus can be perceived as a duck eating a rabbit or as a rabbit following another rabbit (among other combinations), depending on which hypothesis becomes more likely given context. We could think of physiological changes as highly ambiguous in this sense, so close to just noise that this is why it has been so hard to find emotions configured in the physiological landscape (section 4.3.). Then, it could be argued that a certain emotion arises instead of another emotion as contextual knowledge makes one emotion-hypothesis more likely than another emotion-hypothesis. Then, there is no need there to be non-noisy physiological changes that configure emotion types, because emotion differentiation can be accounted for in terms of contextual knowledge alone. Contextual, external cues determine which emotion is taking place, and not the physiological cause that gets to be perceived.

Now, note that in the double duck-rabbit case, when the contextual piece of knowledge that the duck is eating the rabbit is activated, it disambiguates visual signals in favour of that hypothesis. In a PP context, this means that knowledge of that extra piece of information results in the top-down generation of the visual signals corresponding to that winning hypothesis, instead of the top-down generation of the visual signals corresponding to a competing hypothesis (let's say, one rabbit following another rabbit). At the same time, PE signals relative to the selected hypothesis is given high-precision, while the 'portions' of the incoming signals compatible with the competing hypothesis are given less weight. As a result of that, one forms and experiences, for example, the visual percept of the duck eating the rabbit (hypothesis *A*), and not the visual percept of one rabbit following another rabbit (hypothesis *B*).

Thus, in this sort of case, even though the visual stimulus is highly ambiguous, certain visual signals are more expected for hypothesis *A* than for hypothesis *B*. These hypotheses then involve different precision regimes for different 'portions' of the total incoming signal, and 'explain away' then different 'portions' of the incoming data. More generally, any top-down sensory hypothesis must specify distinguishable sensory data relative to another competing top-down sensory hypothesis. This implies that these two distinct hypotheses 'explain away' different 'portions' of the total precision-weighted incoming signal. Hypotheses *A* and *B* then expect distinct visual data to generate from the top-down, and different precision regimes relative to the total incoming data (both hypotheses expect different precisions regimes for different portion of the total incoming signal). Thus, both hypotheses should minimize different aspects of the incoming visual signal. This is what occurs (for the sake of argument) in the duck-rabbit example, or in any case of 'ambiguous figures', in which two distinct percepts can be formed by extracting different regularities in the same input.

Where do those visual priors came from? From regularities pertaining to ducks and from regularities pertaining to rabbits in the domain of light-reflecting objects (i.e., the visual domain). Since there are indeed regularities pertaining to those animals in the visual domain, the visual hierarchy can, via learning, recapitulate these regularities

over time. The visual hierarchy can then built expectations regarding which visual data is more likely given a duck or a rabbit. When the incoming input is highly ambiguous, these learnt visual expectations are then used so as to 'explain away' incoming visual signals, once contextual knowledge has contributed in deciding which hypothesis is more likely.

The same applies in the emotion case above. For example, two interoceptive hypotheses about the emotion that might be causing certain highly ambiguous interoceptive signals are considered, let's say, the fear-hypothesis and the excitement-hypothesis. Exteroceptively perceived contextual cues can be used so as to decide between the two hypotheses: does the context make more likely fear or excitement? Let's say that a snake is approaching. Thus, the interoceptive hypothesis for fear becomes the better hypothesis. In a PP context, this means that knowledge of such a contextual cue results in the top-down generation of the interoceptive signals corresponding to that winning hypothesis (the fear-hypothesis), instead of the top-down generation of the interoceptive signals corresponding to the competing hypothesis (the anger-hypothesis). More generally, and as I mentioned above, any top-down sensory hypothesis must specify distinguishable sensory data relative to another competing top-down sensory hypothesis. This implies that they 'explain away' different 'portions' of the total precision-weighted incoming signal. Now, according to IIE, emotion-hypotheses are interoceptive in nature (otherwise they could not be compared to actual interoceptive data, which is what it is claimed emotion-hypotheses need to 'explain away' so as to minimize interoceptive PE). Therefore, a certain emotion-hypothesis must specify distinguishable interoceptive data relative to another competing emotion-hypothesis. Then, in the case being considered, once the fear-hypothesis becomes (indeed) the better hypothesis, the interoceptive signals expected for the fear-hypothesis are generated from the top-down, meeting and thus successfully 'explaining away' their relevant precision-weighted 'portions' of the incoming signal. In this manner, the experience of fear arises, instead of the experience of anger.

Then, just as in the visual case above, in this sort of case, even though the interoceptive stimulus is highly ambiguous, certain interoceptive signals are more expected for the fear-hypothesis than for anger-hypothesis. Where do these interoceptive prior expectations come from? From regularities pertaining to emotion in the inner milieu. This reading of IIE also requires then that there must be regularities pertaining to emotion in the physiological domain. The argument from Section 4.2. applies to this case.

4.4.3. All emotion hypotheses expect the same interoceptive activity

There is another closely related objection. IIE could be read in the following way, closer to the way in which Hohwy (2013) presents it. According to this reading, the claim that IIE puts forward is that basically *all* emotion hypotheses expect the *same* interoceptive states. In this respect, perceptual interoceptive hypotheses about which emotion might be taking place encode expectations about which affective properties should occur given such a considered emotion. The affective properties in question can be taken to consist in an arousal state together with its hedonic value, which are assumed to be states in the interoceptive system. All emotions involve arousal, and the tasks of 'interoceptive inferences' is to infer which emotion is causing it. So there is an arousal state, and the emotion system needs to infer what is causing it: does the situation make more likely that fear, instead of anger, is occurring? Emotions differentiate then only in terms of our distinct 'cognitive' responses to the same expected interoceptive states (Hohwy, 2013, p.243). In this respect, the direct analogy with vision does *not* go, as different visual hypotheses encode *different* sensory expectations; while in this reading of IIE, different emotion-hypotheses all expect arousal properties. In other words, different emotion-hypotheses (the joy-hypothesis, the fear-hypothesis, etc.) they all predict the interoceptive system to encode the sensory states characteristic of arousal (together with its hedonic value). Thus, there is no need there to be regularities pertaining to different emotions in the physiological landscape. In this respect, it is only required that there must be regularities relative to arousal properties in the physiological domain, and it is not controversial to assume that

arousal together with its hedonic value are states in the interoceptive system (see, e.g., Barrett, 2006b; Quigley & Barrett, 2014). Once these interoceptive states are predicted by a certain perceptual emotion-hypothesis, an emotion arises.

Now, note that a state of arousal is typically considered to be a bodily experience, i.e., an interoceptive experience. Thus, a state of arousal implies that an interoceptive percept (the one that constitutes the felt experience of arousal) has already been formed, as phenomenal consciousness is populated by nothing over and above percepts. It might seem natural then to consider this reading of IIE along the lines of a strict two-factor, Schachterian view. In this sort of view, emotions amount to cognitive interpretations of current bodily experience ('arousal' for Schachter). Thus, Schachterian views exhibit two separate factors: bodily perception plus 'cognition'. In this kind of view, 'cognition' has the function of merely explaining or making sense of current bodily, interoceptive experience. So, in the first place, a state of arousal emerges, and then 'cognition' makes sense of the latter in light of current external contextual aspects. In strictly Schachterian views, contextual knowledge merely contributes in making sense of an already formed interoceptive percept, without playing any modulatory role in in the formation of the latter (the first, separate factor), as a PP view must claim. Here high-level knowledge does not shape percepts.

However, the reading of IIE in question should not be understood in a strictly two-factor, Schachterian manner. This, since a PP view of emotion, which takes emotions to arise as *interoceptive* PE is minimized as bodily activity is perceived, must claim that higher level emotion-knowledge modulates lower levels of the interoceptive hierarchy, thus shaping interoceptive perception from the top-down. This, by generating descending interoceptive predictions that meet incoming interoceptive signals. In a word, the reading of IIE in question should not be understood along the lines of the case of the burglar from Section 3.3.1. (p. 119).

In that sense, to count as a proper PP view in which top-down predictions shape percepts, the reading of IIE in question should be understood in roughly the following way. A certain event activates interoceptors. This triggers interoceptive activity at lower levels. These incoming interoceptive signals need to be 'explained away'. At the same time, in parallel, exteroceptively perceived contextual aspects make the fear-hypothesis the most promising hypothesis. The interoceptive signals expected for fear are then generated from the top-down. These latter signals amount to the interoceptive signals expected for a state of arousal (together with its hedonic value), since these are the kind of interoceptive state that all emotion hypotheses expect. As the expected descending interoceptive predictions successfully 'explain away' such incoming interoceptive signals, an interoceptive percept is formed. This percept constitutes the experience of arousal (together with its hedonic value). However, such percept (arousal) was shaped by top-down predictions generated by the 'fear-model' (the 'fear-model' expects arousal to take place). Thus, given that such percept is formed driven by the fear-hypothesis, in this case such experienced arousal (together with its hedonic value) amounts to the experience of fear. The emotion of fear thus arises. Then, according to this reading of IIE, at the same time that the experience of arousal (together with its hedonic value) is constituted by such descending predictions, the experience of fear arises.

This reading of IIE looks plausible, as it does not face the problem highlighted in Section 4, and so it is also immune to the counter-arguments discussed in this Section. This reading of IIE also significantly departs from a perceptual, interoceptive view of emotion along the lines of the James-Lange view, since this reading of IIE does not assume that there must be emotions configured in the physiological landscape. This reading of IIE is only committed to the claim that the physiological landscape configures arousal, together with its hedonic value (i.e., valence). In this sense, this reading of IIE is closer to the view defended in this Thesis, according to which valence properties (and its degree of "intensity") are configured in the inner milieu, and they are shared by all emotion types (Chapters 5 and 6).

However, it might not be the most plausible way to read IIE. Evidence shows that firstly the affective properties of stimuli are determined between 120 and 180ms after stimulus onset (i.e., their positive or negative valence and degree of arousal), and only then more fine-grained aspects of the content of such stimuli are determined between 200 and 350ms after stimulus onset (e.g., sad/loss, fear/danger) (see, Barrett et al., 2007; Barrett & Bar, 2009; Eimer & Holmes, 2007; Palermo & Rhodes, 2007; Vuilleumier & Pourtois, 2007). That is, during perceptual tasks, once arousal (together with its hedonic value) is constituted, only then emotion experience begins to arise. In other words, during emotion, first an interoceptive percept is formed, which constitutes the experience of arousal (together with its hedonic value). After that, once such an interoceptive percept has already been formed, then an emotion hypothesis needs to determine which emotion the situation makes more likely. That is, both factors seem to be processed separately. This suggest that it is unlikely that the perception of arousal states is shaped by emotion hypotheses. High-level, contextual knowledge plays the role of merely explaining or making sense of current bodily, interoceptive experience. That is, evidence suggest a two-factor, Schachterian sort of view, where contextual knowledge only explains (or makes sense of) an already formed percept, without shaping it or playing a role in the formation of such percept, as a PP view must claim. In a word, evidence suggest that this reading of IIE might be problematic to some degree.

4.4.4. The 'amalgam' objection

It could be argued that there is no need there to be regularities pertaining to emotion in physiological landscape, since interoceptive perceptual inferences integrate all sorts of information, not just interoceptive information, across all levels of the cortical hierarchy. Emotions are some sort of 'amalgam' of all that kind of information.

This reading of IIE hits the nail on the head in recognizing that in order for an emotion-hypothesis to be selected, the whole generative model that constitutes an agent needs

to be employed. That is, as I commented in Section 4.1. above, in order for an emotion-hypothesis to be selected among competing emotion-hypotheses, knowledge encoded in all sensory modalities, and in higher multimodal/amodal regions, across all levels of abstraction (or time-scales), needs to be recruited. However, contrary to this reading of IIE, I called attention to the fact that, in IIE, representations encoded in non-interoceptive sensory modalities, and amodal and multimodal knowledge, serve *interoceptive* ends: finding the better interoceptive hypothesis able to 'explain away' the incoming signals proprietary of the interoceptive 'channel' (Section 4.1).

The problem with the reading of IIE in question then is that it abandons the original idea put forward by Seth (and Hohwy) that emotions are a sort of interoceptive perception. IIE unambiguously identifies emotion with interoceptive perception of a certain sort. That is why Seth insists in claiming that emotions arise from interoceptive perception *in direct analogy* to the way in which visual percepts are formed (Seth, 2015; Seth & Friston, 2016). Thus, one manner in which 'the amalgam objection' can be evaluated is by seeing whether an analogous claim can be made in the visual domain. I think such a claim cannot be made, as it leads to a rather trivial tenet. In the case of vision, high-level amodal, and knowledge from non-visual modalities, contribute across all levels of the cortical hierarchy during visual percept formation. In this sense, we can then talk of an 'amalgam' involved during visual percept formation. However, this certainly does not make visual percepts identical to multimodal percepts. Of course, there is a rich interaction between modalities and contextual knowledge across all levels, but that does not make the visual percept that results from this process stop being a visual percept (a representation in the visual 'channel'), to put it this way. Knowledge from other modalities contribute by constraining the kind of visual activity that is generated from the top-down so as to successfully 'explain away' the incoming signals proprietary of the visual channel. That is, in the case of vision, forming a percept requires meeting incoming *visual* signals alone (to put it this way, one does not need to hear so as to see). Given the direct analogy between visual percept formation and emotion generation, just as we do not claim that, in the PP framework, visual percepts are identical to an 'amalgam',

there is no reason for us to read IIE as maintaining that this is what occurs in the emotion/interoception case.

There is another, more general worry, closely related to 'the amalgam objection'. Certainly, typical episodes of emotion generation and their regulation are complex, most of the time involving many aspects. For example, certain movements, facial expressions, thoughts, motivations, concomitant perceptual experiences in different modalities, action tendencies, etc. However, as I commented in Chapter 1, the goal of an emotion theory is to single out the minimal essential components of emotion, and to determine in which way these components interact so as to generate an emotion. That is, the goal of an emotion theory is *not* to determine which aspects typically contribute causally to emotion. In other words, the fact that emotions typically involve many aspects does not mean that all of them, even though they interact, are proper parts of the mechanism that constitutes emotion. At least, something must be said about why are all those components jointly required for emotion generation.

Anyway, I do not mean to imply that some sort of amalgam-like, multimodal view of emotion is implausible. It could be the case that a PP amalgam-like view is better suited to account for emotion than approaches which avoid putting all (or many) components associated with emotion as constituent parts of emotion. However, in this Thesis I am interested in discussing the interoceptive inference view insofar as it amounts to a PP version of the perceptual, interoceptive view of emotion, according to which emotions arise in direct analogy to vision.

4.5. Final remarks

Interoceptive, perceptual theories of emotion are indeed problematic—as there are no significant regularities pertaining to emotion in the physiological domain. So a PP account of emotion should avoid modelling emotion as interoceptive perceptions.

Contrary to IIE's claim, predicting interoceptive signals during perceptual inference cannot be then what is primary in emotion generation. Thus, Seth's (and Hohwy's) IIE lacks a compelling way to account for emotions *per se*. Interoceptive perceptual inferences are not in the driver's seat when it comes to emotion.

On the other hand, the interoceptive inference approach can also be taken to be an account of subjective feeling states in general, rather than as an account of emotions *per se*. That is, the interoceptive inference approach can also be taken to be an account of affect. Considering that valence is the essence of affect, in this respect, the claim put forward by the interoceptive inference approach is that affective *valence* arises by minimizing interoceptive PE via interoceptive *perceptual* inference. In the coming Chapters (Chapter 5 and 6), I defend this view from criticism and show that it can reply to objections.

As I argued in this Chapter, as it stands, IIE cannot account for emotion *per se*. Even though PP can indeed account for valence (affect)—as I argue in the coming Chapters—this leaves PP without an account of emotion. This is major drawback for PP ambitions. In Chapter 7, I will suggest that the PP approach to interoception can indeed be used to account for emotion *per se*. This requires amending IIE in a key respect. I agree with IIE's claim that emotions arise by minimizing interoceptive PE—after all, common sense (and also experimental research) indicates that interoception is part of emotion generation. I think that this more general claim is on track. However, I suggest that, contrary to IIE, emotions do *not* arise by minimizing interoceptive PE in the *specific* way proposed by IIE. It might be the case that emotions, instead of arising via interoceptive *perceptual inference*, arise via interoceptive *active inference*. In other words, emotions are not about forming an interoceptive percept of a certain sort. Rather, emotions are strategies for changing an interoceptive percept that has already been formed (via interoceptive perceptual inference). That is, emotions are specific strategies for regulating valence (affect). Now, interoceptive percepts (i.e., valence) inform about our homeostatic condition. Then, emotions are better seen as specific strategies for regulating homeostasis.

5. Against non-sensory theories of valence

According to IIE, emotions *per se* arise by minimizing interoceptive PE. This is the more general claim put forward by IIE. IIE's more specific claim is that emotions arise by minimizing interoceptive PE in a specific sort of way, namely, via *interoceptive perceptual inference*. That is, by attempting to find an interoceptive emotion-hypothesis that fits incoming interoceptive signals. In this respect, in direct analogy to the manner by which visual percepts are formed, IIE claims then that emotions are interoceptive percepts that result from a content-specifying emotion-hypothesis. As I just argued, this more specific claim is problematic.

However, as I commented in the Introduction, and as I will discuss in more detail in Chapter 6, the interoceptive inference approach can also be taken to be an account of affective valence, rather than a view on emotion *per se* exclusively. In this respect, the claim put forward by the interoceptive inference approach is that the affective aspect of subjective feeling states—i.e., affective valence—arise by minimizing interoceptive PE, in direct analogy to the way in which vision operates. That is, affective valence, not emotion *per se*, results from a perceptual, interoceptive process. This is the *interoceptive inference theory of valence* (ITV).

In ITV, perceptual hypotheses, insofar as they are *interoceptive* perceptual hypotheses, predict sensory, interoceptive activity. Thus, valence properties count as sensory, interoceptive representations. Also considering that everything in the PP perceptual machinery seems to be sensory, in the PP framework valance properties should also be taken to be constituted by sensory processing. In ITV, affective valence can be seen then as a sensory, perceptual state. The PP framework can be seen as committed then to the view that valence amounts to a sensory phenomenon.

However, the view that valence is a sensory phenomenon is controversial. What might be called *non-sensory signal theories of valence* (NSS) (Prinz, 2004, 2010; Carruthers, 2011) have compellingly showed that valence is not constituted by sensory processing. Valence cannot be understood as an interoceptive phenomenon. So any account that models valence on perceptual processing seems to be doomed to fail.

Taking into account all the above considerations, and the fact that valence is a necessary component of emotion (and the mark of the affective), discussing the nature of valence is key for my purposes. In this and the next Chapter, I will do exactly that.

In this Chapter, I focus in showing that NSS is not a compelling view on the nature of valence. Therefore, valence is likely not to be a non-sensory item in the furniture of the affective mind. Thus, the door is open for the view that valence is a sensory item in the furniture of the perceptual predictive machine posited by the PP framework.

In order to show that affective valence is likely to be a perceptual phenomenon, I begin by briefly characterizing valence (Section 5.1.), and introduce widely agreed desiderata for a theory of valence (Section 5.2.). Then, in Section 5.3., I present Prinz's and Carruthers's versions of the non-sensory signal theory of valence (NSS) (Sections 5.3.1. and 5.3.2., respectively), and their explanatory advantages (Section 5.3.3.). As I will show in Section 5.4., NSS faces decisive problems on its own (see above), which makes NSS a poor candidate for a compelling, plausible theory of valence. Thus, the door is open for the view that valence is a sensory item in the furniture of the perceptual predictive machine posited by the PP framework.

5.1. Characterizing valence

We strive to have certain kinds of emotions, and we strive to avoid having other kinds of emotions. Certain emotions are agreeable, while other emotions are disagreeable.

That is, there are positive emotions and negative emotions. For example, joy, pride, love, and amusement typically are positive emotions; while anger, fear, guilt, and contempt typically are negative emotions. Emotions are classified in this way in virtue of the character of its valence. Certain emotions are positive emotions since they have as a component positive valence; and certain emotions are negative emotions since they have as a component negative valence[25].

Valence is not only part of our folk psychological understanding of the nature of emotion, but it is also a construct that plays a fundamental role in the scientific study of emotion (see, e.g., Barrett, 2006a; Russell, 2003; Berridge & Kringelbach, 2015), to the point that, for some theorists, valence is one of the main building blocks of emotion (Barrett, 2006a; Russell, 2003). So note that the notion of valence in which I am interested in this Chapter is a non-normative notion that plays an explanatory role in psychology. Thus, contrary to what a few researchers have pointed out (e.g., Charchland, 2005; Picard, 1997; Solomon, 2003), when it is said, in the affective sciences, that an emotion is *positive* or *negative* (i.e., that it has positive or negative valence) it is *not* being said that such an emotion is positive or negative in the sense of being *good* or *bad* normatively, in any ethical or prudential sense. Valence is neither an ethical nor a prudential construct; it is a psychological construct that plays a role in affective sciences.

I am simply assuming then that there is such a thing as valence and that it does play a role in our best current theories about emotion. I am aware that Solomon (2003) has argued that valence is not a unified construct, and that, consequently, emotion research could just simply dispense with the notion of valence. However, if in the literature there are many different, orthogonal ways of using the term 'valence'—as Solomon critically remarks—it is simply because we have many competing theories of valence.

[25] Note that this way of characterizing valence leaves open the possibility that a certain emotion type *E* can have different valence value on different occasions.

Just as the fact that there are many different, orthogonal ways of using the term 'attention' (given that there are many competing theories of attention) does not imply that psychology should dispense with this construct, the fact that there are many competing theories of valence does not imply that there might not be a theory of valence that successfully captures the distinction between positive and negative emotions.

In fact, making sense of the evaluative polarity captured by the notion of valence appears to be mandatory precisely because "valence consistently comes out as the primary factor" in statistical analyses of emotion reports (Cochrane, 2009, p. 387). In other words, contrary to Solomon, given that valence is ubiquitously present on our thinking about affective phenomena, rather than abandon the notion altogether, "it would be more appropriate to specify the concept more exactly, so that possible differences in interpretation do not confound experimental results." (Cochrane, 2009, p. 387).

Not only emotions have valence as a component. Given that valence is the key construct in affect, all affective phenomena are valenced. For example, drives or motivations also exhibit such a positive and negative character as, for example, hunger and thirst, which are negatively valenced. Also moods are valenced, as depression and anxiety, which typically have negative valence as a component.

5.1.1. The evaluative and motivational role of valence

Why psychologists posit the construct of valence? What is the functional role of valence? One uncontroversial, straightforward answer is that the role of valence is to make things positively or negatively *matter* for the agent (i.e., endow things with positive or negative significance), thus facilitating and *impelling* behaviour relative to now relevant, valued aspects of the environment. Valence makes things matter to us,

and consequently urges action. That is, valence plays an evaluative and motivational role. Valence impels organisms to act in ways consistent with what they appraise as good or bad, as when, for example, negative emotions impel us to get away from stimuli appraised as having a negative impact, or to allocate attentional and memory resources towards negative events.

The claim that valence is mainly invoked as a construct that plays an evaluative and motivational role is evidenced by the way in which it has been characterized throughout its history in psychology[26]. For example, Tolman (1932) understood valence as the 'attracting or repulsive forces' that objects have for organisms. Lewin (1935) understood valence as 'imperative environmental facts' that guide behaviour and give rise to a world of significance for the organism. Schneirla (1959) understood valence in terms of the direction of behaviour toward relevant goals. More recently, Fridja (1986) held that valenced objects result attractive or aversive for organisms, defining thus the meaning of a situation. Davidson (1993) claimed that valence involves approach and avoidance behaviours, which are, according to him, intrinsically pleasant and unpleasant, respectively. Views that understand valence as evaluations (Ben-Ze'ev, 2000; Lazarus, 1991; Ortony et al., 1988) can also be seen as positing valence as a motivational construct, given that evaluative contents, contrary to mere indicative contents, recommend courses of action. Valence has also been characterized as involving representations of goals (Dyer, 1987; Izard, 1991; Panksepp, 2005; Rozin, 2003), which are arguably inherently motivational. Carruthers (2011) holds that valence consists in an inner non-sensory signal that confers value (good or bad) to attended stimuli, and motivates their pursuit or avoidance. Also Prinz (2004, 2010) holds that valence make affective states matter to us, and that it is in the business of shaping behaviour.

The evaluative and motivational role of valence is inherited by emotions, insofar as the latter contain a valence marker. In this way, the content represented by a certain

[26] See Colombetti (2005) for a historical review of the notion of valence.

emotion (i.e., its 'core relational theme') becomes something that matters for the agent, so that the latter is impelled by its emotional state to act in a relevant fashion. This evaluative/motivational aspect of emotion, inherited from the valence marker they contain, has been highlighted throughout the history of emotion research (see Frijda, 2008, p.72).

Now, it may be useful to briefly clarify what I mean here by 'evaluation' and 'motivation'. By 'evaluation' I simply refer to a mental item that, when it is associated or combined with a certain mental state, makes the latter something positive or negative for the agent. Now, since theories of valence are theories of what makes positive (negative) affective states *positive* (*negative*), the nature of such positivity/negativity will depend on the theory of valence that turns out to be true. I will argue for one of such theories. By 'motivational states' I mean mental states such as urges, drives, cravings, the readiness to act that can be felt before selecting a specific action in the external environment, etc. In other words, the *motivational oomph* that impel us to regulate affective internal states by changing them (of course, this usually take us to figure out how to act in the external environment). Thus, by 'motivation' I do not mean to refer to reasons for taking specific courses of action in the external environment, which result from practical reasoning (i.e., motives). On the other hand, considering that we can act not only physically but also mentally, and that it occurs previous to selecting an specific set of motor actions, by 'motivation' I do not mean to refer to motor states involved in the initiation of specific muscle and joints movements.

5.2. Desiderata for a theory of valence

There are certain general properties that it is widely agreed that valence must have. These properties are fundamental platitudes that constitute explanatory goals, or desiderata, for theories of valence.

A theory of valence must satisfy the following desiderata[27]. Firstly, a theory of valence must have sufficient scope so as to apply to all clear cases of emotion (*scope desideratum*). That is, all clear cases of emotion should exhibit the property or mechanism that is proposed as accounting for valence. Secondly, a theory of valence must accommodate our pre-theoretical taxonomies regarding the emotions that typically count as positive and the emotions that typically count as negative (*pre-theoretical taxonomy desideratum*). Thirdly, a theory of valence must explain the truism that positive emotions feel good and negative emotions feel bad (*feeling desideratum*). Finally, a theory of valence must account for the evaluative and motivational role of valence. That is, why valence makes things positive and negative to us (*evaluative desideratum*); and how the property or mechanism that is proposed as accounting for valence manages to account for its characteristic motivational role (*motivation desideratum*).

Intuitively, if a certain theory of valence *A* satisfies more of these desiderata than a certain theory *B*, then *A* is to be preferred over *B*.

5.3. Competing theories: non-sensory signal theories of valence

Partially following Prinz (2004, 2010), I distinguish four main families of theories of valence: (a) *approach/avoidance theories* (e.g., (Davidson, 1993; MacLean, 1993). These theories identify positive and negative valence with approach and avoidance behaviors (or action tendencies), respectively. (b) *Evaluative theories* (e.g., Ben-Ze'ev, 2000; Lazarus, 1991; Ortony et al., 1988). According to one version of this view, positive and negative valence are identified with the evaluation of a situation as goal-congruent or goal-incongruent, respectively (Lazarus, 1991). According to another version of this view, positive and negative valence are identified with the judgment that the intentional object of the relevant emotion is good or bad, respectively

[27] I do not think this list is exhaustive in any way. It is a tentative list. Nonetheless, these desiderata capture shared criteria for a theory of valence to count as satisfactory (see Prinz, 2010).

(Ben-Ze'ev, 2000; Ortony et al., 1988). (c) *Hedonic theories* (e.g., Barrett, 2006a; Damasio, 1994; Frijda, 1993; Schroeder, 2001), according to which positive and negative valence are identified with pleasure and displeasure, respectively. (d) What might called *non-sensory signal theories* (NSS) (e.g., Carruthers, 2011; Prinz, 2004, 2010), which identify valence with inner signals—not grounded in any perceptual system—that mark representations as good or bad (wanted or unwanted), so they are not representations themselves.

In this Section, I will only discuss NSS for the following reasons. Firstly, I take Prinz's (2004, 2010) arguments against *approach/avoidance* and *evaluative* theories to be successful in showing that they are no longer tenable (I will comment on this below). Secondly, the type of view that I will defend in this Chapter can be taken to be a version of a *hedonic theory*, so I will focus on this view later. Thirdly, NSS is the most recent philosophical proposal, it has several explanatory advantages, and for this kind of theories the most philosophically careful arguments have been developed. Finally, I consider NSS to be the main rival to the view I will be defending, for this kind of theories separate valence from sensory, interoceptive representations, contrary to the view I defend.

Non-sensory signal theories of valence

Non-sensory signal theories of valence (NSS) hold that valence amounts to a motivating inner signal that 'marks' mental states as good or bad, welcome or unwelcome. These views also share the claim that valence is *not* a sensory state, and, being just a 'signal', it neither amounts to an amodal (or multimodal) representation of any kind (e.g., a concept). Recently, Prinz (2004, 2010) and Carruthers (2011) have proposed versions of non-sensory signal theories. Since Prinz's view is much more developed, I will discuss Prinz's view more thoroughly.

Importantly, these views align with a certain tradition in affective sciences, which consists in regarding the affective, valenced aspect of sensory/perceptual experiences as something that "attaches" to sensory/perceptual representations. In other words, the valenced aspect of a sensory experience is regarded as something 'extra' to the sensory representations themselves. For example, according to the tradition in question, eating a sweet cake feels good because an affective mental item (valence) got attached to the sensory representation of sweetness, the latter being a distinct mental item from the former. That is, only when sensory/perceptual representations (e.g., sweetness, a landscape, music, etc.) have a "hedonic gloss" added by affect is that those representations become something that feels good (or bad). Such a "hedonic gloss" is considered to be a non-sensory, non-representational item in the furniture of the mind, distinct from any sort of sensory/perceptual representation or high-level knowledge (see, e.g., Berridge & Kringelbach, 2010, p.9).

5.3.1 Valence as inner reinforcers

Recently, Prinz (2004, 2010) has offered an account of the nature of valence that not only accounts for the intuitive plausibility of competing theories, but contrary to current proposals, it seems to satisfy all the above desiderata.

Prinz (2004) identifies emotions with perceptions of bodily changes. In Prinz's account, emotions, understood in this way, include a signal for its own cessation or continuation. Valence amounts to such a signal. More precisely, according to Prinz, valence amounts to what he calls *inner reinforcers*. Inner reinforces are inner devices that signal non-propositionally structured commands. Such commands specify whether the considered emotion (i.e., relevant perceived bodily state) should be maintained and had it more often in future occasions (positive valence), or whether the considered emotion should be cast aside and not had it in future occasions (negative valence).

Prinz illustrates the workings of inner reinforcers with imperatives that, regarding a certain emotion, say something like "More of this!" (positive valence) or "Less of this!" (negative valence). That is, inner reinforcers are signals which mediate behaviour relative to the maintenance or cessation of the relevant perceived bodily state (emotion). Thus, in Prinz's theory of valence, positive emotions are those which impel its own maintenance, and negative emotions are those which impel its own termination. In this sense, inner reinforcers make their target mental states something positive or negative: they make emotions something that *matter* for the agent (Prinz, 2004, p.178).

As signals with imperative content, inner reinforcers are inherently motivating: "Emotions exert motivating force by means of valence markers" (Prinz, 2004, p.242). Inner reinforcers impel agents to do something, namely, change an inner state by taking them to figure out how to act in the external environment, i.e., select candidate courses of action that could lead to the maintenance or cessation of a certain positive or negative emotion, respectively. Note that inner reinforcers do not motivate specific courses of action in the external environment, designed to achieve the goal of maintaining or terminating a certain emotion, such as, for example, deciding to go to a party to terminate a state of sadness. In other words, inner reinforcers do not command specific strategies of emotion regulation. Inner reinforcers just urge us to modify internal states, they command "Less of this inner state! Do whatever it takes to achieve that!", which can lead then to decide a certain specific regulatory action in the external environment, such as, for example, going to a party.

Importantly, inner reinforcers are reward and punishment signals. As such, they serve a major role in reinforcement learning and conditioning, besides their motivational role commented above. Inner reinforcers allow agents to learn which stimuli count as rewards and which stimuli count as punishments. External stimuli encountered in the past, and that triggered a certain emotion, will be sought out in the future, since a positive valence marker commanding the maintenance of that emotion got attached to it. This kind of associations are stored in memory. Thus, stimuli associated in memory

with emotions that contain a positive valence marker will increase the probability of an "appetitive" response—i.e., those stimuli count as a reward for the agent. The same can be said of negative valence markers, mutatis mutandis. In this sense, inner reinforcers are learning signals (i.e., learning signals with motivational force).

Crucially, inner reinforcers and perceived bodily changes are dissociable, separate components of emotion (Prinz, 2004, pp. 163-164). Thus, inner reinforcers can, in principle, be attached to non-emotional mental states. Now, since representations constituted by the bodily sensory modalities and inner reinforces are distinct, separate entities in the furniture of the mind, and inner reinforcers are neither grounded in exteroceptive modalities, valence is a *non-sensory* signal. Consequently, valence is not something that can be felt. However, inner reinforcers and bodily changes go usually together, making the latter seem good or bad to the agent.

Instead of presenting empirical evidence for this view (which still needs to be produced), Prinz supports IRV by arguing that it has more explanatory advantages than competing theories, and by showing that it can account for the most compelling aspects of competing theories. I will present these explanatory advantages below, in Section 5.3.3., and I will discuss his arguments against the hedonic theory of valence in the next Chapter.

5.3.2. Carruthers's view on valence

Recently, Carruthers (2011, p.126-135) has defended an account of the nature of valence along very similar lines to Prinz's view. According to Carruthers, valence consists in an inner non-sensory signal that confers value (good or bad) to attended stimuli. This non-sensory signal inherently motivates the pursuit or avoidance of such stimuli. Nonetheless, according to Carruthers, valence signals get generally attached

to representations of *external* events (e.g., your partner arriving home safe), rather than to inner bodily states, as in Prinz's view.

Carruthers (2011) adheres to the generally accepted view that affect consists in valence and arousal, and that emotions have affect as a component. He also adheres to the view that affect is to a major extent dependent on inner bodily perception (i.e., interoception). Nonetheless, as I mentioned above, in Carruthers's view, valence-signals are *non-sensory* signals. For, according to Carruthers, valence and arousal are separate causal mechanisms in the furniture of the affective mind, and represented physiological changes constitute arousal, not valence. When we experience physiological changes we do not then have the experience of valence, but of arousal. Valence only makes attended events good or bad. Furthermore, in Carruthers's view, valence is also a *non-conceptual* signal. For it confers value without deploying high-level abstract knowledge, such the concepts GOOD or BAD. In a word, valence is a non-sensory, non-conceptual indicator of value[28].

Carruthers emphasizes the role that valence plays in decision-making. In this account, valence drives decision-making. Closely following Damasio (1994), he claims that during decision-making valence signals gets attached to representations of considered options, making thus the latter attractive or repellent. That is why patients with lesions in regions associated with valence, such as the orbitofrontal cortex (OFC), show poor decision-making capacities (Damasio, 1994). However, contrary to Damasio (1994), Carruthers holds that valence does not amount to sensory, interoceptive representations.

Carruthers offers no positive evidence or argument for the view that valence is non-sensory. He does offers evidence for the claim that valence is non-conceptual, namely, the famous Bechara et al.'s (1994) Iowa Gambling Task studies, where subjects *see*

[28] Carruthers (2011, p.128) thinks that pleasure and pain consists in a positive or negative valence signal that got attached to bodily changes, respectively.

some decks as bad, without *judging* them to be bad. However, since this latter claim is not controversial, and I agree with it, I am not going to discuss such evidence. He does have arguments *against* the view that valence is sensory. I discuss these arguments below, in the sub-section where I reply to the objections against the view that valence amounts to interoceptive perceptions.

5.3.3. Explanatory advantages of NSS

Considering that Prinz, but not Carruthers, discusses how his view compares to competing theories in respect to the above desiderata, in this section I will mainly focus in IRV. However, since both Prinz's and Carruthers's views, being non-sensory theories, share all their key commitments, Prinz's points regarding IRV, in respect to the desiderata in question, directly apply to Carruthers's view.

NSS, if true, accommodates the desiderata presented in Section 5.2. Firstly, considering that non-sensory signals are motivating signals, NSS should have no problems accommodating the *motivation desideratum*. Secondly, it does not seem implausible to think that all emotions include a motivational component that impels us to look for ways to maintain or eliminate them. The single fact that there are emotion *regulation* and dysregulation phenomena speaks in favour of this assumption. Thus, IRV satisfies the *scope desideratum*. This also applies to Carruthers's view, for nothing in this view prevents that non-sensory signals can 'mark' emotion themselves, making thus the latter welcome or unwelcome. IRV also satisfies the *pre-theoretical taxonomy desideratum*. Clear cases of negative emotions, such as anger, guilt, and fear are emotions that we feel motivated to get rid of; and clear cases of positive emotions such as joy, elation, and love are emotions that we feel motivated to sustain. Finally, it seems that NSS also satisfy the *feeling* and *the evaluative desiderata*. Insofar as we are impelled to eliminate negative emotions and to maintain positive emotions, valence makes emotions something that matter to us, and in this sense we take them to be bad and good feelings, respectively.

Competing accounts fail to accommodate all the desiderata. Take, for example, the *approach/avoidance view* (Davidson, 1993; MacLean, 1993), which identifies positive and negative valence with approach and avoidance behaviours (or action tendencies), respectively. This view also fails to satisfy the *evaluative* and the *feeling desiderata*—as it not clear what is the link, if any, between approaching (avoiding) something and having it as good (bad), and feeling good (bad) about it. Moreover, the approach/avoidance view also fails to satisfy the *scope desideratum*—since not all emotions involve distinctive types of behaviour or action tendencies. It also faces problems regarding the *pre-theoretical taxonomy desideratum*, since clear cases of negative emotion (e.g., anger) typically involve approach behaviour.

Or consider the *evaluative view*. According to one version of this view, positive and negative valence are identified with the evaluation of a situation as goal-congruent or goal-incongruent, respectively (Lazarus, 1991). According to another version of this view, positive and negative valence are identified with the judgment that the intentional object of the relevant emotion is good or bad, respectively (Ben-Ze'ev, 2000; Ortony et al., 1988). The evaluative view does not accommodate the *scope desideratum*, since it is uncontroversial that non-human animals have valenced states. However, it is unlikely that, let's say, chaffinches have the conceptual apparatus required for evaluations and judgments to take place. It also fails to satisfy the *feeling desideratum*, as it is not clear how evaluations and judgments can feel like something.

IRV can also explain the intuitive plausibility of competing accounts, while avoiding their typical problems (Prinz, 2010). Just to take an example, IRV can easily explain why we tend to *avoid* the object of negative emotions. We tend to behave that manner because, frequently, an effective way to obey the imperative "Less of this emotion!" is to go away from the event that is causing the relevant emotion. At the same time, and contrary to the *approach/avoidance view*, IRV can accommodate the case of anger. We tend to approach the object of anger because putting an end to the situation that

causes the emotion by intervening on it is an effective way of getting rid of what is bringing about the emotion that needs to be eliminated. Finally, given that inner reinforcers are mere labels that function as command signals or indicators of value (i.e., they are *non-conceptual* signals), IRV avoids the problem of turning valence into something that is cognitively too demanding to be possessed by non-human animals and children, as the *evaluative view* does[29].

Considering that NSS can, if true, accommodate more desiderata than competing accounts, and it can explain the intuitiveness of the latter, it has an explanatory advantage over them.

5.4. Problems for non-sensory theories of valence

However, as I will show in this Section, NSS face decisive problems on their own, which also make NSS poor candidates for a compelling, plausible theory of valence.

Remember that this kind of view considers the affective, valenced aspect of sensory/perceptual experiences as something that "attaches" to sensory/perceptual representations. That is, sensory/perceptual representations (e.g., sweetness, a landscape, music, etc.) become something that feels good (or bad), only in case they have a "hedonic gloss" added by affect (valence). Such a "hedonic gloss" is considered to be a non-sensory, non-representational item in the furniture of the mind, distinct from any sort of sensory/perceptual representation (and also distinct from high-level knowledge or 'concepts'). Thus, by showing that NSS are problematic, I aim to show that the tradition in affective sciences that considers affect as a non-sensory mental item that "attaches" to sensory representations is misguided. This opens the door for

[29] Prinz also argues that the hedonic theory of valence cannot accommodate all the desiderata. In the next Chapter, I will show that a certain version of the hedonic theory can successfully reply to those objections.

the view that valence is a sensory representation on its own. I will defend such a view in the coming Chapter.

5.4.1. Problems for IRV: Valence cannot amount to inner reinforcers

Roughly, according to Prinz's (2004) account, emotions are constituted by two components, namely, perceived bodily changes and valence markers (inner reinforcers). These distinct components serve different functions. On the one hand, perceived bodily changes are supposed to serve two roles: (1) representing organism-environment relations (i.e., core relational themes), and (2) allocating physiological resources in order to facilitate adaptive action, independently of any kind of motivation. On the other hand, inner reinforcers signal an urgency to act to change inner bodily states, making such perceived bodily states matter for the agent. That is, inner reinforcers play a motivational role; while bodily changes serve a semantic function and facilitate action.

Crucially, as I mentioned above (Section 5.3.1), inner reinforcers and perceived bodily changes are dissociable components of emotion. That is, during emotion, IRV assumes a dissociation between motivation and interoception (i.e., the perception of the physiological condition of the entire body). This assumption is rather implausible. Let me explain.

Motivation, understood as a sense of urge, as an urgency to change an inner state, is a mental state that can be felt, and thus reported. That is why during certain emotional episodes we say things like: "I feel less motivated than before to get rid of this stomach ache" or "I'm really motivated to stop feeling guilty about my divorce". I take this to be uncontroversial.

Now, only perceptual representations can be felt. Remember that in this Thesis it is assumed that phenomenal consciousness is populated only by percepts, in the sense

that there are no phenomenal qualities beyond sensory representations. This assumption simply denies that there can be non-sensory representational vehicles with qualitative character (see Prinz, 2012).

If such motivational oomph is something that we usually feel, and only perceptual representations can be felt, then this sort of motivation is something which takes place in some perceptual modality. Obviously, the feeling of motivation in question is not something that we have due to olfaction, nor vision, nor any exteroceptive or proprioceptive modality—and there is no reason to claim that some mixture of those can do the job. The best candidate seems to be the interoceptive system (see, e.g., Berman et al., 2013; Naqvi & Bechara, 2010; Noel et al., 2013). That is, in this respect, the intuition is that the motivational oomph exhibited by valenced states consists in perceiving that our bodies are physiologically (un)prepared to engage in action. If the felt motivational component of emotion is grounded in interoception, and valence plays such a role of *impelling* us to act to change an inner state during emotional episodes, then there is likely to be no dissociation between interoception and valence, as IRV assumes.

The absence of this predicted dissociation gains intuitive support from affective states such as motivations or drives, as for example, hunger. Hunger is a felt valenced urge. In accordance with the claim that there is likely to be no dissociation between interoception and the motivational oomph in question, anatomical and functional evidence show that the experience of hunger is exhaustively grounded in interoceptive brain structures (see Craig, 2015).

Of course, it could still be argued that the motivational oomph characteristic of valenced states arises in the interoceptive system, but only in case interoceptive representations become the target of the modulatory action of inner reinforcers. This could explain the above intuition. In other words, such motivational oomph arises only in case a valence marker attaches to an inner bodily representation, the latter being the

only kind of representation that can be the proper target of valence markers. Thus, without interoceptive representations, the feeling of motivation in question cannot take place; however, interoceptive representations can take place without valence markers. Then, Prinz's view (i.e., dissociation between valence and bodily perception) predicts that there can be inner bodily states associated with a certain affective state which simply lack valence—i.e., they can be experienced without that state mattering for the agent—and that this abnormality is not dependent on abnormalities due to ill-formed interoceptive representation, but rather it depends on an abnormality in another component with a functional profile along the lines of inner reinforcers.

In the first place, this reply is question begging. The phenomenology of the motivational oomph characteristic of valenced states—plus the assumption that consciousness is exhausted by sensory/perceptual representations—indicates that the feeling of motivation in question is exhausted by interoceptive representations. Insisting that another component besides the latter representations needs to be posited seems then redundant, without phenomenological considerations or empirical evidence that justify speculating about the existence of this other component – inner reinforcers – that can modulate interoceptive representations in the required way. That is, without a case for the claim that interoceptive representations by themselves are motivationally inert, and without positive evidence for the dissociation of motivational oomph and interoceptive representations, the above objection begs the question.

There seem to be no cases of feelings grounded in interoception—such as drives/motivations—that dissociate from its valence. For example, there seem to be no such thing as non-valenced hunger or thirst. However, Prinz (2004) argues that surprise is a case in which bodily perception dissociates from valence. The (purported) emotion of surprise can be positive or negative. There are positive and negative surprises. However, in both cases, Prinz speculates, surprise is constituted by the same pattern of bodily changes. In such a case, the same pattern of bodily changes can then exhibit different valence. Therefore, so the argument goes, perceived bodily changes and valence dissociate.

I am not convinced by this line of reasoning. Assuming for the sake of argument that Prinz is right in claiming that surprise involves a distinctive pattern of bodily changes (though see Chapter 4), and that there are positive and negative surprises, there are still alternative explanations for this kind of case, in which the same pattern of bodily changes can exhibit different valence. Crucially, such alternative explanations can account for this kind of case without positing a separate component from bodily perception. Rather, they can account for this kind of case by showing that the same process by which percepts are standardly formed in the PP framework can operate in such a way that different percepts can arise from the same input stimuli. In fact, this is a typical phenomenon in the PP framework. There is no reason to deny this for the case of percepts of the inner condition of the body. Then, the PP view of percept formation can straightforwardly explain how the same pattern of bodily changes (the input stimuli) can exhibit different valence values, on the assumption that the latter are constituted by interoceptive perceptions.

As I discussed in previous Sections, in the PP framework, the assessment of the *precision* of PE, which is identified with attention, occupies a central role in the whole PP inferential machine. Attentional modulation, understood as the differential, context-sensitive assignment of precisions, can explain how different percepts can result from the same input stimuli.

Our expectations of precision depend on context. Among other things, given certain incoming sensory data, this implies that, depending on context, precisions differ within the same modality for different sensory features. During percept formation, attentional modulation determines which aspects of the incoming sensory signal are given more weight, and which aspects of the incoming sensory signal are (relatively) ignored. This occurs in a cascading fashion across all levels of the perceptual hierarchy. Then, different expected precisions, at different levels of the perceptual hierarchy, determine the relative influence that top-down sensory expectations have relative to the incoming

input. Importantly, several contextual factors influence such an assignment of weights, as for example, evaluations of the situation in light of previous experience. Thus, in different occasions and contexts, given different assignment of precisions, the percept that results will be composed of different interoceptive features, changing thus the configuration of the percept that will eventually be experienced (for example, in certain occasions the interoceptive representation in question will have heart activity represented and in other occasions not). Thus, by differentially weighting different aspects of the same input via attentional mechanisms, the kind of percept of the body that (purportedly) individuates positive surprise can differ from the percept of the body that (purportedly) individuates negative surprise (remember that I am assuming for the sake of argument that Prinz is right in claiming that surprise involves a distinctive pattern of bodily changes). Given certain incoming data, the idea is that different sensory attributes and precision regimes are expected for the positive-surprise-hypothesis than for the negative-surprise-hypothesis. For example, once the positive-surprise-hypothesis is considered to be the best sensory hypothesis given context (let's say that things typically deemed to be good are taking place in the external situation), the incoming interoceptive signals associated with negative emotions will be inferred to have low precision and will be taken to be noise. The inverse is the case for the negative-surprise-hypothesis. This determines that, from the same incoming signal, different bounded sensory attributes (i.e., percepts) are predicted by these distinct hypotheses. No additional, 'extra-sensory' mental item (e.g., some sort of non-sensory valence marker) is required to explain differences in valence given the same kind of input stimuli from the body.

In fact, different inferred precisions given context are likely to be responsible for observed differences in valence in cases such as the placebo effect (Büchel et al., 2014). When subjects are told that a certain infusion consists in a potent painkiller, placebo hypoalgesia is much stronger compared to the case when subjects are told ambiguous information about the infusion: it could be a painkiller or just a placebo infusion. These different pieces of contextual information regarding the properties of the infusion differentially bias the expected precision of the top-down sensory expectations regarding pain signals (negative valence). When subjects are told that the

infusion is a painkiller, the expected confidence in the hypothesis that the incoming signals are better explained by 'relief' drives percept formation (positive valence). Consequently, the 'portion' of the incoming signal that is compatible with the pain-hypothesis is inferred to be unreliable. It does not then have much influence in the process of finding a perceptual solution. However, when subjects are told that the infusion could be a painkiller or just a placebo, PE relative to the pain-hypothesis is given relatively more weight, having thus more influence in the process of finding a perceptual solution (see Büchel et al., 2014). This is a case in which the same input stimuli coming from the body can give rise to different valenced percepts given different inferred precisions. To put it this way, contextual cues drive attention in such a way that different 'portions' of the same incoming data can be given more or less attention, so that different configurations of features (percepts) can result. Importantly, different valenced states can arise from the same input from the body, *without* positing anything alien to the sensory machinery itself: the same stuff that is part of the typical process of percept formation accounts for states of different valence given the same bodily input.

IRV faces another challenge. Let me consider just the case of positive valence for ease of exposition. Positive inner reinforcers are inner devices that not only motivate, but also play a role in reinforcement learning. That is, IRV assumes that, in this respect, the same inner device that motivates also amounts to a reward learning signal. Now, the best candidate neural structures for realizing reward markers are dopamine rich structures in the midbrain. More precisely, reward learning signals are likely to be realized in the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNpc) (Schroeder, 2004). Not only other researchers have pointed out that these dopaminergic structures are fine candidates for realizing Prinz's positive inner reinforcers (Corns, 2014), but they have some properties that seem to nicely fit Prinz's construct: besides being responsible for reward learning signals (Schultz et al., 1997), they do not realize hedonic tone (Schroeder, 2004), and more controversially, they also

play a role in reward-based motivation (Berridge, 2007). It is not arbitrary then to take VTA and SNpc as realizing positive inner reinforcers[30].

IRV makes then the following key prediction: if positive valence amounts to an inner reward signal, then VTA and SNpc should be consistently active during positive emotions. Certainly, current neuroimaging studies on positive emotions are limited. However, they clearly show that these structures are not significantly and consistently involved during positive emotions (e.g., Lane et al., 1997; Murphy et al., 2003; Phan et al., 2002; Vytal & Hamann, 2010). Thus, it is hardly the case that positive valence amounts to an inner reward signal.

5.4.2. Problems for Carruthers's view

On the other hand, Carruthers's arguments to the effect that valence amounts to a non-sensory signal are misguided. As I commented above, according to Carruthers, valence and arousal are separate causal mechanisms in the furniture of the affective mind. While represented physiological changes constitute arousal, valence amounts to a non-conceptual signal, not grounded in any sensory modality. Let's call this the *independence claim*. One of the main reasons Carruthers puts forward to support the independence claim is that dimensional approaches to emotion arrange emotions in a circumplex graph, in which valence and arousal are represented as independent, *orthogonal* dimensions (e.g., Russell, 2003). Thus, "it is implausible that the former should reduce to the latter" (Carruthers, 2011, p.130). Considering that arousal is the dimension which is grounded in sensory, interoceptive representations—no argument nor evidence is offered for this claim!—valence should then be a non-sensory signal, not grounded in interoception.

[30] Prinz (2004, p. 161-162) mentions some regions that have been found to activate during some tasks that trigger valenced affective states in subjects. He does this in order to give some substance to the claim that there is indeed such thing as valence. Valence is a real phenomenon. However, Prinz is not explicitly committed to any neural realizers of valence markers.

I think this argument mistakenly takes reports of affective states to evince the nature of the causal mechanism responsible for affect. Probably because, in some dimensional approaches to emotion (e.g., Russell, 2003; Barrett, 2006a), linguistic analyses of questionnaires and extended *reports* of valence and arousal are graphically represented as orthogonal *descriptive* dimensions of affect in a circumplex graph (e.g., Russell, 2003), the above argument mistakenly takes valence and arousal to be independent *causal* components of affect, realized by separate causal mechanisms. That is, it is simply a mistake to infer that valence and arousal are realized by independent causal mechanisms of affect from the fact that linguistic analyses of reports of affect result in a circumplex graph, where valence an arousal figure as orthogonal descriptive dimensions. The circumplex model is designed to capture what people say about their own psychology, about their own subjective experience regarding affective states. The circumplex model does not then track the actual way in which the causal mechanisms responsible for those reports are related. Dimensional theorist of emotion explicitly recognize this point (e.g., Kuppens et al., 2013; Russell & Barrett, 1999). In fact, nothing in the circumplex model of affect, which Carruthers (2011) seems to endorse, prevents that valence and arousal are essentially unified at the *causal* level, both being grounded in bodily, interoceptive representations. In other words, nothing prevents that, for example, arousal is a component part of the mental state that constitutes valence, or that affect amounts to valence *at* certain levels of arousal, i.e., that the latter is not an independent aspect of affect, but it simply amounts to the intensity that valence can take. Certainly, as some dimensional theorists have found (Barrett et al., 2004), one can consciously focus more on one descriptive dimension than the other, as Carruthers notes, but nothing in the circumplex model of affect prevents that valence and arousal are both constituted by represented physiological changes at the causal level. In fact, this is what dimensional theorists seem to endorse (e.g., Russell & Barrett, 1999, p. 814-815; Barrett, 2006a; Barrett, 2015, p.45). Carruthers fails to consider the possibility that valence and arousal are essentially unified. As Barrett & Bliss-Moreau remark:

> "Core affect is a state of pleasure or displeasure with some degree of arousal (Barrett, 2006b; Russell, 2003; Russell & Barrett, 1999). Together, valence and arousal form a unified state, so 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." (Barrett & Bliss-Moreau, 2009, p. 171)

For example, it could the case that arousal, instead of being a separate construct than valence, simply corresponds to the "intensity" or 'volume' taken by the perceived physiological changes that constitute valence, so the former cannot take place without the latter.

Carruthers is impressed by the fact that people who are better at detecting their own heart-beats tend to exhibit more *arousal focus* than *valence focus* (Barrett et al., 2004). The notions of arousal focus and valence focus simply refer to the emphasis that subjects place on words related to arousal and valence during emotion reports, so that arousal and valence emerge as important aspects in the verbal descriptions of affect in an individual over time (Barrett & Bliss-Moreau, 2009). Then, considering that heart-beat detection is taken to be an indicator of *interoceptive accuracy* (see Garfinkel et al., 2015), one might be lead to conclude that arousal, rather than valence, is the aspect of affect that is grounded in interoceptive perception (see also Dunn et al., 2010).

However, notice that I am not disputing the claim that arousal consists in bodily perception. The above argument only suggests that that claim might be the case. That is, the above argument does not speak against the claim that valence could also consists in bodily perception. More precisely, the above evidence is consistent with the claim that valence consists in the perception of a *pattern* of inner bodily changes, which includes several changes besides changes in hear-rate, which are likely to be valence-neutral by themselves. In other words, such evidence is consistent with the claim that valence value is determined by the overall whole-body shape taken by the evolving inner bodily landscape, in which several physiological dimensions interact (e.g., Damasio, 1994, 2003); while arousal is determined only by heart-rate perception (or the perception of a small sub-set of physiological changes). Thus, this kind of evidence won't do the job for the defender of NSS. Of course, heart-rate seems to be a reliable

indicator of arousal, as I defended in Chapter 1. However, heart-rate is but one dimension of a pattern of changes in the inner physiological milieu. Thus, considering that changes in heart-rate might be critical for arousal but not much for valence value, it is certainly expected that people who are good at paying effortful attention to their heartbeats also exhibit an emphasis in the descriptive dimension of arousal during verbal reports. But this fact, let me insist, is silent with respect to whether valence is non-sensory.

It could be argued that the claim that valence consists in the perception of such pattern of inner bodily changes rest on a confusion, because arousal, but not valence, is the aspect of affect that consists in such a *pattern* of bodily changes. Carruthers endorses this view:

> "arousal is constitutive of the "fight or flight" preparations undertaken by the body in response to threat. […] It consists of a variety of autonomic changes in heart-rate, blood pressure, activity in the sweat glands, and levels of adrenaline and other chemicals in the bloodstream, as well as behavioral changes in posture, muscle tension, breathing rate, and so on." (Carruthers, 2011, p.127)

As I commented above, this claim is precisely what arousal theory proposed under the label 'general sympathetic arousal'. However, as I discussed in Chapter 1, even though this conception of the notion of 'arousal' it is still uncritically endorsed in some corners of psychology, that conception of arousal is not tenable anymore. Remember that according to arousal theory, general sympathetic arousal underlies "fight-or-flight" responses via a single mechanism that controls several measures of sympathetic/autonomic effectors. This conception of arousal emerged from mid-twenty century research on the brainstem reticular formation, which it was hypothesized that it realized the so-called ascending reticular activating system, basis of the sort of activation (arousal) responsible of "fight-or-flight" responses that Carruthers has in mind. It was thought that the brain structure in question was a functionally homogenous structure, and that it had activational (arousal) effects without any sort of specificity.

This conception of arousal is deeply problematic. A key prediction of this approach is that physiological measures of sympathetic activity (such as electro-dermal response and heartrate) should significantly co-vary within and across individuals. However, that turns out not to be the case (Bernston & Cacioppo, 2007; Cacioppo et al., 1991; Lacey, 1959, 1967). On the other hand, the many variables involved during autonomic control do not exhibit some sort of single continuum of activation or arousal properties that could be involved in a simple "fight or flight" mechanism (Bernston & Cacioppo, 2007). In other words, there is no *patterned* set of autonomic responses that constitutes a unified arousal system. In fact, the reticular formation, the supposedly key functionally homogeneous neural basis of the arousal system, is compose of several structures, each of them with its own functional profile (Sarter et al., 2003). The notion of 'arousal' on which Carruthers relies is simple outdated. In fact, it has been shown that certain autonomic measures do not reflect arousal at all, but rather they reflect valence properties. For example, cardiac activity, heart rate, blood pressure, and skin conductance duration reflect affective valence (see Cacioppo et al., 2000). Also the startle response and also facial EMG indicate valence rather than arousal (Mauss & Robinson, 2009).

On the other hand, remember that one of the reasons that Carruthers puts forward for the independence claim—i.e., the claim that valence and arousal are separate causal mechanisms, being arousal the one constituted by represented physiological changes—is that interoceptively *accurate* people (Garfinkel et al., 2015) tend to exhibit more arousal focus than valence focus (Barrett et al., 2004).

As I mentioned above, *interoceptive accuracy* is typically measured by heartbeat detection tasks. On one version of this task, subjects are asked to determine whether or not their own heartbeats are synchronized with a metronome (Barrett et al, 2004). Interoceptive accuracy is also measured by asking subjects to count their heartbeats, and then their responses are compared to the actual number heartbeats as measured by ECG (Ehlers & Breuer, 1992; Schandry, 1982). Thus, interoceptive accuracy only tells us how good an individual is in effortfully attending and keeping track of the

consciously accessible outputs of interoceptive processing (i.e., already formed interoceptive percepts). In this sense, this kind of task is meta-representational: subjects must form 'beliefs' about already formed interoceptive representations or percepts. Thus, interoceptive accuracy tells us nothing about the causal mechanism of interoceptive percept formation, or whether it is working properly or not (i.e., delivering proper interoceptive percepts), neither whether it is hyperfunctioning or hypofunctioning. In addition to that, arousal focus only tell us about linguistic emphases in questionnaires. Thus, arousal focus tell us nothing about the workings, nor the nature, of the *causal* mechanism responsible for arousal. It neither tells us whether the causal mechanism responsible for arousal is hyperfunctioning (nor hypofunctioning) in individuals high in arousal focus. These aspects are the relevant ones, if the goal is to determine whether either arousal or valence (or both) are constituted, at the causal level, by perception of the physiological inner milieu. In order to conclude that arousal, but not valence, is constituted by interoceptive perception, Carruthers needs to show, at least, that interventions in the functionings of the causal mechanism responsible for interoception give rise to modifications in arousal, but not in valence, and that triggering an state of arousal determines modifications is the interoceptive system, without altering valence properties. Considering what the construct of interoceptive accuracy really tell us, it is not much useful for the defender of the independence claim.

It is worth also considering that, insofar as it is operationalized by heartbeat detection tasks, interoceptive accuracy does not reflect *general* interoceptive accuracy. That is, accuracy not only with respect to heartbeat activity, but with respect to the activity of the whole pattern of physiological variables that constitute the physiological landscape of an organism. A more interesting correlation for the defender of the independence claim—even though not much useful for the reasons presented in the above paragraph—would be then that between general interoceptive accuracy and arousal focus. This is the case since, as I mentioned above, the claim that arousal is determined only by heart-rate perception is compatible with the claim that valence is grounded in interoceptive perception. The idea here is that heart-rate perception is critical for arousal, while valence is constituted by the perception of a whole pattern of

physiological changes, and not just the single variable of heart-rate. Thus, the fact that good heart-beat detectors exhibit high arousal focus does not points towards the independence claim. The defender of the latter would prefer to find a correlation between good whole-body perceivers and arousal focus. There is no such evidence. Moreover, it is unclear whether consciously monitoring heart-beats indicates *interoceptive* accuracy or a somatic, exteroceptive capacity: The 'beats' that are monitored by subjects during the heartbeat detection task could simply be reflecting the activation of somatic, non-interoceptive receptors on the chest wall.

Carruthers (2011, p.130) puts forward another argument for the *independence claim*. As it is well-known, after the lesion, OFC/vmPFC patients lose their capacity to respond appropriately to rewards and punishers, which has severe consequences for their personal lives and social interactions (Damasio, 2003). However, they retain their ability for cold reasoning. The standard explanation of the behaviour of OFC/vmPFC patients is that they lose the capacity to associate valenced responses with representations of behavioural options, which is key for normal decision-making. The upshot of this is that valenced responses are required for decision-making, and that malfunctioning of OFC/vmPFC compromises the ability to use such responses to guide decision-making. Carruthers thinks that if valence were grounded in representations of bodily changes, people who tend not to explicitly focus attention on their heart-rate would show aberrant decision-making abilities in the way shown by patients with OFC/vmPFC lesions. But they do not. Then, so the argument goes, valence is not grounded in representations of bodily changes. Valence should then amount to a non-sensory signal, while arousal is the aspect of affect that is grounded in patterns of bodily changes.

There are several problems with this argument. Let me point to two of the most worrying of them. In the first place, the fact that OFC/vmPFC patients show poor decision-making abilities seems to favour the view that valence is grounded in representations of the body. Let me briefly explain. Determining that something is positive or negative, beneficial or harmful, is key for decision-making. It is generally

assumed that bodily responses inform about whether something is beneficial or harmful: positively valenced bodily responses assign positive value to considered behavioural options, while negatively valenced bodily responses assign negative value to them. Given its rich connections with regions involved in visceral representation and control, the OFC/vmPFC, during decision-making, plays the role of linking representations of external situations with representations of bodily responses. Insofar as OFC/vmPFC patients fail to link the input from the body that informs about value with their considered options, they fail to behave appropriately. That is, the standard explanation of the pattern of behaviour shown by these patients assumes that bodily responses determine valence value (Damasio, 1994).

In the second place, it is simply a mistake to infer that a certain mechanism is malfunctioning from the fact that its outputs tend not to be explicitly attended. Then, just as the fact that some people do not tend to explicitly focus on the phonetic properties of their spoken language does not imply that their mechanism of language production is malfunctioning, the fact that some people fail to explicitly focus on their heart-rate does not implies that their mechanism responsible for bodily representation is malfunctioning. Then, it not surprising that such people do not behave as OFC/vmPFC patients.

Carruthers also argues for the *independence claim* by remarking that when the so called *affective* (valenced) component of pain is 'switch off' by administering morphine to a subject, the so called *sensory* aspect of pain, supposedly grounded in the somatosensory cortex (so not grounded in interoceptive modalities), still remains intact, so that subjects still sense pain. From this Carruthers infers that valence is not grounded in a sensory modality. However, this argument does not get off the ground. Without stating why the so called affective component of pain is not also grounded in sensory representations, distinct from the ones anchored in the somatosensory cortex (e.g., interoceptive representations), Carruthers's claim that the affective component

is not sensory simply begs the question. Carruthers's view that valence is non-sensory is then not well motivated[31].

Finally, I would like to call attention to a more general worry regarding NSS. It is certainly the case that NSS can meet the feeling desideratum, as I mentioned above (Section 5.3.3.). According to the feeling desideratum, a satisfactory theory of valence must explain the truism that positive emotions *feel good* and negative emotions *feel bad*. There is something, valence, which endows emotions with a positive or negative qualitative character. NSS account for this platitude by claiming that positive (negative) emotions are *good (bad) feelings*. That is, for NSS positive emotions feel good only in the sense that they are good feelings, i.e., feelings that we want to have (or are impelled to maintain). In other words, NSS accounts for the feeling desideratum rather indirectly, by claiming that we *take* positive emotions to feel good because they are good experiences, experiences that we are motivated to maintain. In this account, once a bodily experience is painted by the "hedonic gloss" of a non-sensory, motivating valence marker, it becomes desired, and thus a positive (negative) feeling emerges.

I think this move is not a satisfactory way to meet the feeling desideratum. It makes little sense to hold that because valence markers "gloss" percepts is that the latter feel good (bad). If there are no phenomenal qualities beyond the features that configure sensory representations (i.e., there is no cognitive phenomenology), then if valence markers are not themselves something sensory, it is hard to see how the *positiveness* of a feeling could be part of the phenomenology of emotion, *without* an account of how such a "gloss" becomes part of the sensory features that configure percepts.

[31] Closely following Schroeder (2004), Carruthers also argues that valence cannot be equated with interoceptive, bodily representations, based on cases such as skydiving, where supposedly the same physiological changes can give rise to both positive and negative experiences. This cases are covered by the discussion of the case of positive and negative surprise and placebo analgesia (Section 5.4.1.).

NSS exhibits then several difficulties. The positive claim made by NSS could still be the case. However, as I attempted to show, their negative arguments to the effect that valence is not grounded in bodily perception are not compelling. Considering that their view rests upon such negative arguments, the case for NSS seems to be undermined. Thus, the door is open for the view that valence is a sensory item in the furniture of the perceptual predictive machine posited by the PP framework. In the coming Chapter, I argue that once affective valence is taken to arise via interoceptive inference, the view that valence is a sensory phenomenon can reply to the objections made by defenders of NSS to the view that valence is a sensory representation. Considering that NSS is problematic on their own, as I argued in this Chapter, affective valence is then likely to be a sensory item in the furniture of the mind.

6. Valence and interoceptive perceptual inference

As we saw in the previous Chapter, the view that valence itself cannot be a sensory phenomenon, in both its versions, exhibits several problems. Thus, the door is open for the view that valence is a sensory phenomenon. Nonetheless, it could still be the case that the view that valence itself is indeed a sensory phenomenon is also untenable. In fact, defenders of NSS have compellingly argued that, independently of the truth of their view, assuming that valence is a sensory item in furniture of the mind is implausible. This poses a major challenge to the interoceptive inference theory of valence (ITV). That is, the view that identifies valence, the defining construct of affect (i.e., subjective feeling states), with interoceptive perceptions formed via interoceptive inference.

In this Chapter, I defend the view that valence is indeed a sensory, interoceptive phenomenon. More precisely, valence can be seen as an interoceptive percept that represents positive and negative homeostatic changes, which is formed via interoceptive perceptual inference. Once affective valence is taken to arise via interoceptive inference, the view that valence is a sensory phenomenon can reply to the objections made by defenders of NSS. Thus, affective valence is likely to be a perceptual phenomenon. ITV seems to hold.

In what follows, I begin by critically discussing some of the main possible ways in which valence can be understood in the PP framework (6.1.). Even though each of these views exhibits some difficulties, they all point towards a more global PP view on the nature of valence. All these views point towards the claim that valence amounts to an interoceptive percept, formed via interoceptive perceptual inference, that informs about positive and negative homeostatic changes. I present this claim in Section 6.2. Then, in Section 6.3., I discuss some theoretical and empirical considerations that suggest that the different aspects involved in this claim might be the case. In Section 6.4., I show that the view in question can reply to the objections typically faced by

views that, as the one defended here, identify valence with mental states that can be felt—e.g., (dis)pleasure. Finally, in Section 6.5, I show that this view on the nature of valence satisfies desiderata for a theory of valence.

6.1. Valence and predictive processing

There are various ways in which valence can be taken to be part of the perceptual PP machinery. With the exception of the view proposed by Joffily and Coricelli (Joffily & Coricelli, 2013), these approaches to the nature of valence have not been put forward in a completely explicit manner *as such* in the literature (i.e., as careful, specially dedicated treatments of how valence *per se* should be understood in the PP framework). However, as we will see, these approaches to the nature of valence can be clearly discerned in the PP literature relative to the affective domain. Interestingly, as soon as some of the more problematic aspects of these views are left aside, all these views point towards the same claim. This claim is that valence amounts to an interoceptive percept, formed via interoceptive perceptual inference, which informs about positive and negative homeostatic changes.

6.1.1. Valence as valuable, '(un)familiar' states

One possible way of understanding valence in the PP framework is to equate affective valence with valuable states for the organism. In fact, the move of equating affective valence with that which is valuable for the organism is not arbitrary at all. For example, the close link between affective valence and value is emphasized by views such those of Carruthers (2011) and Panksepp (1998). Remember that, according to Carruthers, valence consists in an (non-sensory/non-conceptual) indicator of value. In this view, valence confers value to attended stimuli, making the latter attractive or repellent. Valence indicates that which is good or bad for the organism. Similarly, for Panksepp, mammals have different emotion systems, each of which attributes value to external

events. In this sense, in this view valence can also be seen as mental item that indicates that which is valuable for the organism[32].

Now, remember that, in the PP framework, minimizing PE is a strategy for complying with the free energy principle (Friston 2005, 2009, 2010) (see Chapter 2). According to the latter, organisms must ensure that they maintain themselves within expected bounds of 'surprisal'. Organisms avoid 'surprising' states, and seek 'non-surprising' events. An organism is in a 'surprising' state in case it is outside the subset of possible states that are most probable for such organism to be in, given its constitution. Then, "the consequences of minimizing free energy are that some states are occupied more frequently than others—*and these states can be labelled as valuable*." (Friston et al., 2013, p.2) (Italics are mine). In other words, 'familiar states' are valuable states. 'Surprising' states reflect conditions incompatible with homeostasis. Therefore, "the long-term (distal) imperative—of maintaining states within physiological bounds—translates into a short-term (proximal) avoidance of surprise" (Friston, 2010, p.2), so that "low free-energy systems will look like they are responding adaptively to changes in the external or internal milieu, to maintain a homeostatic exchange with the environment" (Friston & Stephan, 2007, p. 48). As Joffily and Coricelli (2013) remark,

> "In the free-energy principle, value is the complement of free-energy in the sense that minimizing free-energy corresponds to maximizing the probability that an agent will visit valuable states, where the evolutionary value of a phenotype is the negative surprise averaged over all the (interoceptive and exteroceptive) sensory states it experiences [2]. This formulation parallels a recently proposed reinforcement learning theory for homeostatic regulation [44], which attempts to integrate reward (valence) maximization with the minimization of departures from homeostasis (free-energy)." (Joffily & Coricelli, 2013, p. 10)

Then, one way in which valence can be understood in the PP framework is to equate affective valence with valuable states of the organism. In this view, valence indicates that which is good or bad for the organism. In other words, that which is familiar or unfamiliar, respectively. Familiar states feel good, while unfamiliar states feel bad. In

[32] For the close relation between valence and value, see Higgins (2007).

that sense, positive feelings indicate that the organism finds itself in a familiar state, while negative feelings indicate that the organism finds itself in an unfamiliar state.

At first sight, this seems quite plausible. Given our constitution, states that are not frequently occupied by organisms like us typically feel bad. For example, being at five thousand meters high feels awfully, breaking a bone feels terribly bad, and so on. In these sort of cases, as this view expects, our built-in expectation of low-surprisal is not satisfied.

However, this view cannot be quite right. There are many cases of familiar states that do not exhibit their corresponding positive valence. Just to take an example, one can find oneself sitting on one's desk, without any sort of needs (hunger, thirst, etc.), simply daydreaming about some random stuff. In this sort of case, one finds oneself in a familiar, low-surprisal state. Nonetheless, intuitively, in this sort of cases the valence value is neutral, rather than positive. Certainly, this argument does not show this view to be wrong. In fact, I think there is a lot of truth in it. However, this view can be refined so as to become more compelling (Section 6.2).

6.1.2. Valence as high-level perceptual hypotheses

Another possible way of understanding valence in the PP framework is to hold that valence arises from high-level perceptual hypotheses. More precisely, from high-level *interoceptive* perceptual hypotheses. Seth's view on subjective feeling states can be read in this way (Seth, 2013).

Before presenting this way of reading Seth's 'interoceptive inference' approach to interoception, let me make a few remarks. The notion of 'interoceptive inference' is most of the time put forward by Seth (and Hohwy) as the claim that *emotions* arise

from successfully minimizing interoceptive PE via interoceptive perceptual inference. However, it is clear that Seth does not exclusively use 'emotion' in the restricted sense in which 'emotion' is understood in this Thesis, and in the philosophical literature more generally (see Chapter 1). By 'emotion' he mainly refers to the subjective feeling states that are typically taken to result from bodily perception (e.g., Damasio, 2003). Then, Seth's 'interoceptive inference' approach to interoception can also be understood as the claim that subjective feeling states (or 'interoceptive feelings') arise by minimizing interoceptive PE; rather than the claim that *emotions per se* arise by minimizing interoceptive PE[33]. Now, remember that subjective feeling states amount to affective states, i.e., mental states that feel good or bad at a certain intensity level. Moods, emotions, 'homeostatic motivations', etc. are cases of subjective feeling states. As I discussed in Chapter 1, valence (at a certain intensity level) is what makes affective states affective in the first place. To put it this way, valence is the essence of affect. Thus, Seth's 'interoceptive inference' view on subjective feeling states can be understood as a view on the way in which valenced states arise. This is confirmed by the work of Quattrocki and Friston on the role of interoception and the neurotransmitter oxytocin in the generation of some of the features of autism (Quattrocki & Friston, 2014). In this work, they make use of Seth's 'interoceptive inference' approach to interoceptive processing and subjective feeling states. Quattrocki and Friston propose that, in the interoceptive system, precisions are realized by the neurotransmitter oxytocin. Thus, oxytocin permits proper learning (i.e., the proper building of a generative model) of those aspect of social communication which are impaired in autism (see, Quattrocki & Friston, 2014). They describe an aspect of Seth's view as one in which valence is seen as identical to an interoceptive feeling ('gut feeling') that is constituted by high-level descending interoceptive predictions.

[33] Seth also claims that what he calls the 'sense of presence' arises by minimizing interoceptive PE (Seth, 2013; Seth et al, 2012; Seth & Critchley, 2013). However, Seth (2015) seems to have changed his view in this latter respect: the sense of presence, instead of arising be minimizing interoceptive PE, arises when the generative models that underlie prediction are *counterfactually rich*. Considering that the sense of presence, i.e., "the sense of the subjective reality of the contents of perception" (Seth, 2015, p.2), is arguably a sort of subjective feeling state, Seth's 'interoceptive inference' approach to interoception can also be understood as the claim that 'subjective feeling states' arise from minimizing interoceptive PE via interoceptive perceptual inference.

> "The expectations from these higher level representations then furnish the basis for top-down predictions – that endow perceptual inferences with an affective valence – literally 'gut' feelings – that are thought to be a key component of emotional salience and self-awareness (Critchley and Seth, 2012; Seth, 2013; Seth et al., 2011)." (Quattrocki & Friston, 2014, p. 413)

Let's get down to business. In the PP framework, valenced states could also be understood then as arising from high-level interoceptive perceptual hypotheses. That is, just in the way subjective feeling states arise as described in Chapter 3. As we saw, in direct analogy to the way in which exteroceptive percepts are formed, subjective feeling states arise by way of interoceptive perceptual inference. Then, in this view, valenced states arise via interoceptive inferences of the likely causes of incoming interoceptive signals. That is, via interoceptive perceptual inference. The latter consists in updating hypotheses so that the generated predictions fit the incoming signal. The task here consists then in forming an interoceptive percept. In this view, valence amounts to a high-level interoceptive expectations about which valence value is more likely given incoming interoceptive signals and contextual aspects of the situation. The interoceptive expectations in question are encoded at the top of the interoceptive hierarchy. Then, they generate descending interoceptive predictions that constrain activity in levels of the interoceptive hierarchy that encode both relatively variant and invariant aspects of interoception, such as the physiological condition of the whole body and the activity of single organs and part of the body, respectively (see Chapter 3). The idea is roughly the following. Let's say that a certain situation triggers certain physiological changes. Contextual factors, such as for example, your partner suddenly entering the room, make the interoceptive expectation that something positive is occurring to the organism the most likely hypothesis. The interoceptive activity expected for that hypothesis is generated from the top-down. Let's say that these descending predictions achieve to match and thus suppress incoming interoceptive activity. An interoceptive percept is thus formed: A positive feeling arises.

Note that the problem faced by IIE relative to the fact that there are no regularities pertaining to emotion in the physiological domain (Chapter 4) does not apply to the

view of *valence* in question. There *are* regularities pertaining to valence in the physiological domain (Caccioppo, 1991, 2000) (more on this below).

Importantly, the view presented above (Section 6.1.1.) identified valence with 'familiar states', i.e., states that are more frequently occupied by organisms. Note that 'familiar states' become expected states or 'beliefs' about which sensory data the organism should be encountering. That is, organisms have a built-in high-level expectation of low-surprisal. The satisfaction of the latter determines valence value. Then, both in that view and in the view presented in this subsection valence is identified with interoceptive expectations. However, there are two key differences between this view and the one presented above (Section 6.1.1.). In the first place, there is a difference in direction of fit. On the one hand, the view of valence as familiar states has world-to-mind direction of fit: positive (negative) valence arises as current states conform to the 'goal state' of low-surprisal. On the other hand, as long as valenced states arise via interoceptive *perceptual inference*, Seth's view on valence has mind-to-world direction of fit. In the second place, the view of valence as 'familiar states' does involve high-level interoceptive expectations, but these expectations are long-term expectations. That is, in the long run it is expected that organisms occupy low-surprisal states. Note that this, however, is compatible with, for example, transient states of thirst or hunger. The latter can be expected, for example, considering that it is already 1pm, and one has not had lunch yet. But as long as such transient states are incompatible with the long term expectation of low-surprisal, such states are not valuable, even though transiently expected. Transient expectations are not relevant in this account. On the contrary, the view that valence amounts to a high-level interoceptive perceptual hypotheses can involve the expectation that states incompatible with the 'goal' of low-surprisal should take place, given context. This is the case since perceptual hypotheses attempt to determine the nature of the causes of current, transient incoming signals. On occasions, the expected states that should be taking place (selected perceptual hypotheses) can be incompatible with the goal of keeping low surprisal states, as the organism needs to be informed whether such goal is being satisfied or not. That is precisely the task of perceptual inferences.

I think this view on valence is on the right track. However, it leaves an open question: what makes the resulting interoceptive percept a percept that constitutes a *positive* (*negative*) feeling? Certainly, in the PP framework, the content of the hypothesis determines the content of the resulting percept/experience. Thus, as the standard answer goes, the resulting percept constitutes a feeling of positiveness/negativeness because it was formed driven by a positiveness/negativeness hypothesis, to put it this way. However, this answer obscures a more fundamental question, namely, why such a hypothesis, and consequently the resulting percept, counts as a hypothesis of something positive/negative in the first place. In other words, what makes such a representation (hypothesis) a representation of a positive/negative state?

6.1.3. Valence as a sign of homeostatic states

Another possible way of understanding valence in the PP framework is by making a little twist on the above account. The above account can be extended so as to better emphasize the role of interoceptive inferences in homeostatic processes and, relatedly, in decision-making. Valence properties could be understood as interoceptive percepts that signal positive and negative homeostatic states, conveying thus information about what is homeostatically positive or negative to other systems involved in decision making. Gu and Fitzgerald have proposed such an account (Gu & Fitzgerald, 2014). They have used Seth's 'interoceptive inference' approach to interoception so as to account for the role of bodily perception in informing decision-making in this manner. More precisely, they place Damasio's 'somatic marker hypothesis' (Damasio, 1996) within the framework provided by the 'interoceptive inference' approach to interoception. Very roughly, according to the 'somatic markers hypothesis', during decision-making, representations of the inner milieu constitute valence properties. That is, good and bad 'gut feelings'. These representations are combined (or "integrated") in the vmPFC with exteroceptive representations. In this process, the former 'tag' or 'mark' the latter as positive or negative, guiding thus value-based choices. In Gu and Fitzgerald's extension of this view, interoceptive percepts formed via interoceptive perceptual inference inform about the homeostatic, physiological

condition of the body. These interoceptive percepts inform decision-making by signalling that which is homeostatically good or bad for the organism as in Damasio's account. That is, interoceptive percepts realized in the insular cortex,

> "[…] encode and represent interoceptive information and, in so doing, acts to contextualise choice behaviour by informing other neural systems about the internal state of the body. In other words, the insula computes a 'state' variable of the internal world of the agent and passes it to other neural systems that carry out other computations in decision-making." (Gu & Fitzgerald, 2014, p.269)

I think that this way of understanding valence in the PP framework is on track. It successfully deals with the above problem of what makes an interoceptive perceptual hypothesis a representation of a positive/negative state. It deals with this problem by highlighting the fact that such hypothesis has a positive/negative character given the positivity/negativity of the homeostatic process that it tracks. Before unfolding the philosophical commitments of this view in more detail in Section 6.2., and motivate such commitments (Section 6.3.), let me briefly discuss one last way in which valence has been understood in the PP framework. This, in order to highlight a key aspect of the view on valence defended in this Chapter.

6.1.4. Valence as the rate of change of free-energy over time

In this Section, I discuss the view on the nature of valence proposed by Joffily and Coricelli (Joffily & Coricelli, 2013). To date, this view is the only view on the nature of valence that has been put forward in a completely explicit manner *as such* in the literature (i.e., as careful, specially dedicated treatments of how valence per se should be understood in the PP framework).

According to Joffily and Coricelli, valence can be identified with the rate of change of free-energy (or PE) over time. A shift from a high free-energy to a lower free-energy state gives rise to positive valence, which is identified with pleasure. While a shift

from a low free-energy to a higher-free energy state gives rise to negative valence, which is identified with displeasure. In other words, going from a high surprisal state—so a less familiar state—to a lower surprisal state—so to a more familiar state—constitutes positive valence. A shift in the other direction constitutes negative valence. Interestingly, note that instead of identifying valence with (un)familiar states *simpliciter*, as the view presented in Section 6.1.1., the view in question identifies valence with a *shift* in (un)familiar states. That is, positive and negative valence are identified with a decrease and increase of PE (free-energy) *over time*, respectively.

> According to our definition of emotional valence, when $F_i'(t)$ is positive (i.e., free-energy is increasing over time at level *i* of the hierarchy) the valence of the state at this level *i* is negative at time *t*. When $F_i'(t)$ is negative (i.e., free-energy is decreasing over time at level *i*) the valence of the state at this level *i* is positive at time *t*. (Joffily & Coricelli, 2013, p.3)

Now, remember that PE (roughly, free-energy) is minimized separately at each level of the perceptual hierarchy. Note that in this quote such hierarchical nature of the PP regime is nicely emphasized. This means that "positive and negative valence can be independently attributed to each state in the model" (Joffily & Coricelli, 2013, p.3). It is implied here then that, at a certain time, an organism can entertain different, opposing valence values. While a certain situation can result 'unfamiliar' at lower levels of the perceptual hierarchy, it can be 'familiar' at higher levels, and vice-versa. To take an example from Joffily and Coricelli (2013), think of the case of an old friend who suddenly steps in your door. While the event of such a friend suddenly stepping through the door violates the rather variant expectation about the visual causes of sensations, it fulfils the more invariant expectation of being close to our friends. As Joffily and Coricelli remark, in the view in question, in this kind of case such shift in the familiarity of that visual state should involve unpleasantness (negative valence) at lower, more variant levels. However, the event in question should trigger pleasantness (positive valence) at higher, more invariant levels (Joffily & Coricelli, 2013, p.11). It must be noted that this example makes patent that this view assumes that valence properties take place at higher levels of the perceptual hierarchy (Joffily & Coricelli, 2013, p.11): The old friend example is a case where the affective state that intuitively

takes place is pleasure (positive valence), even though low-level variant expectations were violated; but, higher-level, relatively invariant expectations were satisfied.

As it can be inferred from the above, in this view, valence properties play the role of informing the agent whether the hard-wired goal of keeping surprise levels low is being satisfied or not. Positive valence signals that the agent is on her way towards that goal, while negative valence signals that the agent is deviating from that goal. However, even though the interoceptive domain is important in this respect (as I showed in the previous Chapter), Joffily and Coricelli stress that surprisal can be kept low by minimizing PE in any modality. Then, their proposal "treats interoceptive and exteroceptive predictions (and their uncertainty) on an equal footing" (Joffily & Coricelli, 2013, p.10) in respect to valence. A shift from a familiar state to a more unfamiliar state *in any modality* gives rise to negative valence, while a shift from an unfamiliar state to a more familiar state in any modality gives rise to positive valence.

This latter aspect is another respect in which this view differs from the one presented in 6.1.1., which identifies valence with (un)familiar states simpliciter. The latter view emphasizes the close link between surprising states and their incompatibility with homeostasis/interoception, as "the long-term (distal) imperative—of maintaining states within physiological bounds—translates into a short-term (proximal) avoidance of surprise" (Friston, 2010, p.2). That is, valence emerges in the interoceptive system. On the other hand, the view presented in this Section holds that a shift in the rate of change of free-energy (or PE) over time *in any modality* is what makes affective valence emerge.

The view presented in this Section also differs from the view presented in Section 6.1.2. The latter view understands valence as a high-level perceptual hypothesis about the causes of interoceptive input. However, in the present account, valence is not represented explicitly, but rather it simply emerges as the shift in the 'familiarity' of a state takes place as described above. As Joffily and Coricelli remark,

> "According to our scheme, emotional valence is not estimated itself by the agent but emerges naturally from the process of estimating hidden states by means of free-energy minimization. One could eventually hypothesize that some living organisms, such as humans, explicitly represent valence as one of the causes of their sensations. This means that these agents should also estimate valence (and its uncertainty) like any other hidden state in their generative model. Nevertheless, the explicit representation of valence is not a requirement for emotional valence to exist in our scheme and to play an important role in the adaptation of biological agents to unexpected changes in their world." (Joffily & Coricelli, 2013, p.8)

As an account of valence, Joffily and Coricelli's account faces a key problem. As we saw, according to their view, a shift in the 'familiarity' of a state gives rise to valence. Such a shift can occur in any modality, as this view "treats interoceptive and exteroceptive predictions (and their uncertainty) on an equal footing" (Joffily & Coricelli, 2013, p.10). Thus, to take the case of vision, a change from an unexpected visual state to a more expected visual state gives rise to positive valence, independently of what occurs in other modalities. For example, visually perceiving what initially looks (illusorily) in the mirror as a grey piece of your hair rapidly shifting to its typical brown shades should give rise to a good feeling, independently of the states occurring in other modalities. This is rather implausible. Think of the case of the anhedonia involved in major depression. A depressed person can certainly undergo such a visual shift in visual expectations and have such a visual experience. However, insofar as she exhibits anhedonia, that experience likely feels hedonically indifferent to the depressed person. It does not involve positive valence. A shift in expectations simpliciter cannot be the whole story of how valence emerges. Interestingly, anhedonia has being linked to aberrant functioning of the interoceptive system (see, e.g., Dunn et al., 2009; Furman et al., 2013; Harshaw, 2015; Naqvi and Bechara, 2010). This supports the idea that affectivity is a matter of interoceptive processing, rather than to processing in exteroceptive modalities in isolation of the interoceptive system (more on this below). This is why Friston prefers to present Joffily and Coricelli's account on affective valence as a process that occurs in the *interoceptive* system, rather than as a process that occurs indistinctively in any modality (Ondobaka et al., 2017, p.65).

Joffily and Coricelli's account faces another potential problem. As I have been insisting throughout this Thesis, phenomenal consciousness is populated by nothing over and above formed percepts. In the PP framework, percepts are formed once, during perceptual inference, PE is successfully minimized. The content of such an experience is given by the content of the perceptual hypothesis that achieved to successfully minimize PE. PE as such is not then something that can be felt (Hohwy, 2013). The rate of change of PE is, obviously, a property of PE as such. Valence, the positive or negative character of feelings, is part of the content of affective experience. Intuitively, valence is then something that can be felt, particularly if valence is identified with (dis)pleasure, as Joffily and Coricelli do (though, see discussion below, Section 6.4.). It is hard to see then that valence can be constituted by something that is not part of the content of a perceptual hypothesis, but rather of PE as such, which cannot be felt. Or to put it in a slightly different way, if the content of formed percepts, and thus of experience, is given by top-down perceptual hypotheses, it is hard to see how valence could emerge from something distinct than a perceptual hypothesis and its content.

However, Joffily and Coricelli's account is quite successful in pointing towards a key aspect. If we read their view as Friston does, i.e., as equating valence with the rate of change of *interoceptive* PE, such an emphasis on change is key, because, as we saw in Section 3.1.8., interoceptors track *changes* in the physiological, homeostatic condition of the organism, rather that physiological states as such.

Let me digress. Joffily and Coricelli also offer an account of some aspects of emotion *per se*, along the same lines of their view on the nature of valence (Joffily & Coricelli, 2013). However, there are several reasons to circumvent this view during my exposition. So, in this Thesis, I will not consider it as a competing view on the nature of emotion *per se*. Let me explain.

Joffily and Coricelli hold that their view on valence can be extended so as to account for some aspects of the dynamics of the, so called, 'factive' and 'epistemic' emotions (Gordon, 1987). While the latter are related to uncertain beliefs, the former are related to certain beliefs. More precisely, epistemic emotions are directed at outcomes which are significantly uncertain for the individual. For example, one can fear or hope for events that might or might not occur, such as the development of a treatment for eternal youth. According to Joffily and Coricelli, hope and fear are typical cases of epistemic emotions. On the other hand, factive emotions are directed at events which the individual has a significant degree of certainty that such events are the case. For example, happiness towards the arrival of a loved one. According to Joffily and Coricelli, typical cases of factive emotions are happiness, unhappiness, relief, and disappointment.

Importantly, Joffily and Coricelli replace the notion of (un)certainty in play in the characterization of epistemic and factive emotions made by Gordon (1987). They replace it by the notion, commented above, of the rate of change of PE (or the dynamics of free-energy). Thus, they individuate, for example, the emotions of hope, happiness, fear, and unhappiness in the following way:

> "Our proposal stands on the assumption that, when both $F_i'(t)$ and $F_i''(t)$ are negative (i.e., free-energy $F_i(t)$ is decreasing 'faster and faster' over time) the agent *hopes* to be visiting a state of lower free-energy in the near future at this level *i*. However, when $F_i'(t)$ is negative and $F_i''(t)$ is positive (i.e., free-energy is decreasing 'slower and slower' over time) the agent is *happy* to be currently visiting a state of lower free-energy than the previous one at this level *i*. Equivalently, when $F_i'(t)$ and $F_i''(t)$ are positive (i.e., free-energy is increasing 'faster and faster' over time) the agent *fears* to be visiting a state of greater free-energy in the near future at this level *i*. However, when $F_i'(t)$ is positive and $F_i''(t)$ is negative (i.e., free-energy is increasing 'slower and slower' over time) the agent is *unhappy* to be currently visiting a state of higher free-energy than the previous one at this level *i*." (Joffily & Coricelli, 2013, p. 3-4)

As it is patent from the quote above, in this view, emotions are conceived as directed at 'free-energy states', which are *internal* states to the organism. In this sense, according to this view, emotions inform organisms about changes in the rate of PE minimization. However, as we saw in Chapter 1, emotions are uncontroversially

directed at *external* events (i.e., core relational themes). Insofar as they take emotions to track internal states, their view might be better suited to account for other kinds of affective phenomena, not directed at external events, rather than to account for emotion *per se* (such as, e.g., valence).

However, insofar as changes in the rate of PE minimization at higher-levels of the cortical hierarchy typically result from the satisfaction or violation of high-level exteroceptive expectations about external events, the claim put forward by Joffily and Coricelli has the resources to avoid the above worry. It could be argued, for example, that the high-level exteroceptive expectation of being away from snakes, once violated, results in a 'faster-and-faster' increase of free-energy over time. Thus, when an organism is close to snakes, such an exteroceptively represented mental state is about something *dangerous* for the organism (remember that *dangerousness* is the core relational theme of fear), because it triggers the dynamics of free-energy that individuate fear. In other words, given that 'faster-and-faster' increases of free-energy over time are consistently caused by dangerousness, when a high-level expectation about an external event triggers such free-energy dynamics, the latter makes that which is represented by the exteroceptive high-level expectation in question count as dangerous. In this manner, a state of fear can be constituted.

Nonetheless, the proposal on emotions *per se* made by Joffily and Coricelli still exhibits a couple of pitfalls. In the first place, as we just saw above, the free-energy dynamics that they consider are the free-energy dynamics which take place in any modality (Joffily & Coricelli, 2013, p.10). Thus, to keep the example from above, once the exteroceptive expectation of staying away from snakes is violated, the free-energy dynamics that individuate fear occur, independently of what occurs in other modalities. Just as we saw above in respect to valence, the problem with this view is that the interoceptive modality is privileged in respect to affective states. Thus, imagine an anhedonic, depressed person (or someone with some sort of limbic lesion) who find herself in front of a snake, violating thus her high-level exteroceptive expectation of being away from snakes. As the individual in question finds herself close to a snake,

in this case, the free-energy dynamics that individuate fear take place: free-energy begins to increase 'faster-and-faster' over time. According to the view put forward by Joffily and Coricelli, as long as the individual in question visually (exteroceptively) represents being close to a snake—which violates her high-level expectation of being away from snakes—she should then experience fear. The anhedonic, depressed subject in question (or someone with some sort of limbic lesion, e.g., in insular regions) can certainly have the stored high expectation of being away from snakes, and she can certainly visually represent being close to a snake. However, considering that her interoceptive processing regions are compromised, it is unlikely that she undergoes fear (or a non-aberrant experience of fear), as limbic regions are privileged in respect to affective states. This might be why, for example, patients with lesions in the anterior insula exhibit diminished fear (see, e.g., Devinsky et al., 1995). This also supports the idea that affectivity is a matter of interoceptive processing, rather than to processing in exteroceptive modalities in isolation of the interoceptive system. Therefore, it seems unlikely that dynamics of free-energy *simpliciter* suffice for an emotion to arise. Moreover, emotions are bodily felt. As long as the free-energy dynamics in, for example, the visual modality (or in amodal levels of the cortical hierarchy) cannot account for this non-negotiable aspect of emotion, the view on emotion *per se* advanced by Joffily and Coricelli is not thoroughly compelling.

Certainly, the view in question could be amended by holding that the relevant dynamics of free-energy are the ones which occur in the interoceptive system. Nonetheless, this view still faces a key obstacle. The proposed dynamics of free-energy by themselves cannot account for emotion differentiation. Different emotions can involve the *same* free-energy dynamics. For example, not only fear can involve a 'faster-and-faster' increase of free-energy, but also anger. Intuitively, in typical cases of fear, the agent does not expect ('desires') to find herself in a certain situation (e.g., being close to a snake), but she, actually or imaginatively, finds herself in that situation. This triggers an increase in free-energy at higher-levels of the cortical hierarchy. The non-expected situation is still there, in fact it is more imminent (e.g., the snake is even closer), so free-energy keeps increasing. The agent now expects, at lower levels of the cortical hierarchy, that the situation will remain (the snake is not

going anywhere). Free-energy keeps increasing. In this manner, fear takes place. Now, the same free-energy story can straightforwardly apply to the case of anger. Intuitively, in typical cases of anger, the agent does not expect ('desire') to find herself in a certain situation (e.g., in prison for her political views), but she, actually or imaginatively, finds herself in that situation. This triggers an increase in free-energy at higher-levels of the cortical hierarchy. The non-expected situation is still there, in fact it is now even worst (e.g., the police officer tells her that she is going to be in prison for a really long time), so free-energy keeps increasing. The agent now expects, at lower levels of the cortical hierarchy, that the situation will remain (she is still behind bars). Free-energy keeps increasing. In fact, this sort of story in which free energy increases 'faster-and-faster' over time can be applied to several negative emotions. Considering that accounting for emotion differentiation is a non-negotiable aspect of any philosophical view on the nature of emotion, this is a major issue which makes this free-energy dynamics view on emotion uncompelling. End of digression.

6.1.5. Tacking stock

The first view commented above (6.1.1.) identifies valence with *(un)familiar states* simpliciter. Familiar states feel good, while unfamiliar states feel bad. In this view, valence *indicates* that which is 'familiar' or 'unfamiliar', indicating thus that which is good or bad for the organism, respectively. Moreover, in this view, it is emphasized that 'unfamiliar' states reflect conditions incompatible with *homeostasis*. In one way or another, all views commented above concur in holding that (in)compatibility with homeostasis—or the (un)familiarity of a state—is key for valence. Now, 'familiar states' become expected states or 'beliefs' about which sensory data the organism should be encountering. That is, organisms have a built-in high-level expectation of low-surprisal. The satisfaction of the latter determines valence value. This insight is also shared among the views commented above. Expectations about maintaining the organism within 'familiarity' limits, or limit of homeostatic viability, seems to be key for a PP account of valence. However, unlike the view of Joffily and Coricelli, the view of valence as '(un)familiar' states failed to highlight the fact that *shifts* in the

'familiarity' of a state is what matters for valence, rather than '(un)familiarity' *simpliciter*. This indicates that shifts in homeostatic states—'(un)familiar states'—are key for understanding valence in the PP framework. The view of valence as '(un)familiar' states also fails, however, in remarking that, given that phenomenal consciousness is populated by nothing over and above percepts, valence must result then via *perceptual inference*. Perceptual inference must be then part of the PP story on valence. The second view holds that valence arises from high-level *interoceptive perceptual hypotheses* about incoming, transient physiological sates. This view has the problem of being silent about what makes the interoceptive percepts that constitute valenced states percepts that constitute *positive* (*negative*) feelings. However, when this view is complemented with the insights of the view above, we can begin to understand *valence properties as interoceptive percepts that signal positive and negative homeostatic states*. The latter is that which confers the positive and negative character to the content of the interoceptive percepts that constitute valence. This is precisely the view commented in 6.1.3. Finally, the view presented in 6.1.4., which identifies valence with the rate of change of PE over time, is successful in emphasizing that shifts in the (un)familiarity of states are key. However, it failed to recognize that *interoception* is special in respect to valence. The interoceptive system seems to be the sensory system responsible for the generation of valenced states.

Even though each of these views exhibits some difficulties, it is clear that they all point towards a more global PP view on the nature of valence. From the above considerations we have then that a PP view on valence must hold that valence arises from *perceptual inference*. More precisely, from *interoceptive* perceptual inference. Furthermore, the content of the percepts that result from such a process is given by *shifts* in physiological, *homeostatic* states (i.e., *(un)familiar states*). Valence *indicates* those shifts in physiological states which are good or bad for the organism. In other words, all these views point towards the claim that valence amounts to an interoceptive percept, formed via interoceptive perceptual inference, that informs about positive and negative homeostatic changes. Below I put forward this view.

6.2. Valence as an interoceptive percept formed via interoceptive perceptual inference

The view on the nature of valence that emerges from the PP framework holds then that valence is a representation in a sensory system, namely the interoceptive system. Insofar as valence is taken to be a sensory representation, it counts as a mental state that can be felt. Now, the claim that valence can be felt is typically put forward by views that identify valence with (dis)pleasure (e.g., Barrett, 2006a; Frijda, 1993; Russell, 2003; Schroeder, 2006). In this sense, this view is somehow committed to the view that valence is (dis)pleasure, at least in that respect. In fact, Joffly and Coricelli (2013) describe their view on valence as a view on (dis)pleasure: "Our suggestion is that pleasure is elicited in the transition from a state of high to low surprise." (Joffly & Coricelli, 2013, p.8). In this sense, the view on valence that emerges from the PP framework can be taken to be a version of the *hedonic theory of valence*, which identifies valence with (dis)pleasure.

What is the nature of valence/(dis)pleasure? As I mentioned above, an answer worth exploring is that certain represented bundles of bounded interoceptive features constitute positive valence (pleasure) and other represented bundles of bounded interoceptive features constitute negative valence (displeasure). In other words, certain interoceptive representations or percepts constitute positive valence and others negative valence. This kind of view, as Schroeder (2004, p. 84) puts it, "has some distinguished advocates". For example, an influential version of this view was put forward by Damasio (1994). According to Damasio, (dis)pleasure, or valence, is a "particular body landscape that our brains are perceiving" (Damasio, 1994, p.263), which is triggered by cascading and mutually constraining physiological changes, as described in Chapter 3. This kind of approach can be called an *interoceptive theory of valence*, since valence is seen as constituted by certain interoceptive representations or features.

Now, here is the bold claim defended in this Chapter. Valence arises when interoceptive PE is successfully minimized via *interoceptive perceptual inference*. That is, valence, or (dis)pleasure, amounts to an interoceptive percept. Note that the claim is *not* that valence arises when interoceptive PE is minimized *tout court*. Remember that there are three ways by which interoceptive PE can be minimized: interoceptive perceptual inference, and internal and external interoceptive actions (Chapter 3). The claim is that valence arises only when interoceptive PE is minimized via interoceptive perceptual inference: valence is an interoceptive percept. Let's call this view the 'interoceptive theory of valence' (ITV) (which, as I mentioned above, can be taken to be a version of the hedonic theory of valence).

Put more precisely, ITV's claim—valence/(dis)pleasure amounts to an interoceptive percept—is the following. When an interoceptive percept is formed, it can either constitute pleasure or displeasure, i.e., positive or negative valence, respectively. As we saw above (Section 3.1.7.), interoceptive percepts have a certain kind of content. Now, it is widely accepted that perceptual mental states can be individuated by their content (e.g., Prinz, 2002). Then, the claim in question amounts to the following: when an interoceptive percept is formed, it constitutes positive valence in case it *represents* positive physiological changes, and it constitutes negative valence in case it *represents* negative physiological changes. Remember that physiological changes can be taken to be positive or negative, according to whether they tend to *shift* in a certain way. Physiological changes can be taken to be positive or negative, according to whether they tend to *approach* or *deviate* from the aimed-at regulatory level of homeostasis maintenance, respectively. Thus, (dis)pleasure *informs* us about how we are faring in dealing with the hard-wired goal of maintaining homeostasis (keeping 'surprisal' low).

Note that this is an individuation claim, not a claim about what accounts for the qualitative character of (dis)pleasure. That is, the claim is that when an interoceptive percept is formed, it counts as positive valence/pleasure (negative valence/displeasure) in case it represents physiological changes that approach (deviates from) homeostasis. The claim in question is *not* a version of a representationalist approach to phenomenal

consciousness (e.g., Byrne, 2001; Dretske, 1995; Harman, 1990; Lycan, 1987; Tye, 1995) as applied to valence/(dis)pleasure—i.e., valence/(dis)pleasure feels the way it feels because it has homeostatically relevant physiological changes as its content[34]. Now, contrary to the way the notion of 'content' (as applied to (dis)pleasure) is understood in some representationalist quarters (e.g., Bain, 2013), I do not take the notion of 'content' as demanding that the subject has "personal-level" access to it. Very roughly, by 'content' I refer to the familiar notion that plays an explanatory role in the cognitive sciences, where mental states consist in computational states which are said to possess certain representational properties or content. That is, by 'content' I mean "sub-personal" informational content.

6.3. Motivating the claim

The claim in question is then committed to three main tenets. Firstly, it holds that valence amounts to an interoceptive percept. This might be called *the interoception tenet*. Secondly, ITV's claim holds that, consequently, valence has content. Let's call this *the content tenet*. Thirdly, it holds that such content amounts to physiological changes than can be either positive or negative, given homeostatic standards. That is, valence informs us about our homeostatic, physiological condition. Let's call this *the homeostasis tenet*. Let me motivate these tenets.

6.3.1. Motivating 'the interoception tenet'

The claim that (dis)pleasure amounts to an interoceptive percept is motivated by the following platitude: (dis)pleasure can be felt. That is, (dis)pleasure can take place in phenomenal consciousness. As I commented above, phenomenal consciousness is populated by nothing over and above percepts. Therefore, (dis)pleasure must be a percept of some sort. Intuitively, (dis)pleasure is not grounded in vision, audition, nor

[34] For a representationalist theory of the phenomenal character of (dis)pleasure, see Bain (2003).

in any exteroceptive modality. It is also not grounded in motor/proprioceptive representations. The best candidate seems to be interoception, i.e., the perception of our own bodily changes.

This intuition is not only supported by common sense—every time we undergo a distinctively clear experience of pleasure our body is found to be reacting in different ways—but it is also shared by scientific practice. For example, as the psychologist and cognitive scientists L.F. Barrett remarks: "affective properties such as pleasure, displeasure and arousal — *which are thought to be rooted in interoception* — are fundamental properties of conscious experience" (Italics are mine) (Barrett, 2015, p.425). This intuition is also indirectly supported by various established facts. For example, in the literature on the psychophysics of music experience, piloerection, which consistently occurs driven by certain physiological changes, is used as a standard, reliable measure of pleasure (Guhn et al., 2007; Konečni, Wanic, & Brown, 2007). Moreover, meta-analytic studies show that patterns of physiological changes consistently configure for pleasure and displeasure under certain experimental conditions (Cacioppo et al., 2000; Lang et al., 1993), which supports the idea that *representations* of such physiological changes (i.e., interoceptive percepts) are suitable candidates for constituting positive and negative valence. Then, by observing (under certain conditions) the pattern of physiological changes that a subject is undergoing, it is possible, in principle, to predict whether she is having pleasure or displeasure. Also supporting the intuition in question, using fMRI, Critchley and colleagues (Critchley et al., 2004) measured brain activity during interoceptive performance—as measured by a heartbeat detection task—and rated subject's affective experiences. They found that the right anterior insula is not only key for the experience of bodily changes, but it also grounds a representation of physiological changes that realize pleasant and unpleasant feelings:

> "The observed interrelationship among right anterior insula activity, interoceptive accuracy and subjective negative emotional experience supports the proposal that affective feeling states reflect information concerning bodily responses represented in right anterior insula" (Craig et al., 2004, p. 193).

Considering that the insula is key for realizing interoceptive percepts, the hypothesis that (dis)pleasure is constituted by interoceptive percepts predicts that insula activation, and consequently interoceptive performance, should be found to be good a predictor of whether a subject is undergoing a valenced state. Critchley and colleagues' study support this prediction.

On the other hand, after insula damage, tobacco addicts become tobacco-anhedonics (Navqi & Bechara, 2010). Considering that the insula is a key region for the realization of interoceptive percepts, this fact supports the idea that (dis)pleasure is constituted by percepts in the interoceptive system**.** Relatedly, if valence is constituted by an interoceptive percept, and valence is an essential component of emotion, then, the anterior insula, which is key in realizing interoceptive percepts (Section 3.1.5. and 3.3.4), should be consistently activated during emotions. This is precisely what evidence suggests (e.g., Murphy et al., 2003).

Moreover, it has been shown that autonomic measures such cardiac activity, heart rate, blood pressure, and skin conductance duration reflect affective valence, rather than simply sympathetic activation (Cacioppo et al., 2000) (see Chapter 1). Also the startle response indicates valence (Mauss & Robinson, 2009).

On the other hand, the so-called 'homeostatic motivations' or 'drives', such as for example, hunger, thirst and orgasm, are intrinsically (un)pleasant experiences. Indeed, they are the most primitive, fundamental cases of (dis)pleasure. Crucially, they are exhaustively constituted by dedicated perceptual representations in the interoceptive system. So the pleasantness of those experiences does not seem to be dependent on other kind of perceptual representations. Having already excluded the possibility that valence is something extra-bodily, some sort of non-sensory signal that "attaches" to interoceptive representations (Section 5.4. above)—making thus the latter something that matters for the agent (something that feels good or bad)—it seems then that (dis)pleasure must be grounded in interoception alone.

6.3.2. Motivating 'the content tenet'

The claim that (dis)pleasure is a content-involving mental state might strike some as rather unintuitive. This is because some might have the impression that there is an important respect in which standard perceptual states—which clearly have content—and hedonic states differ[35]. That is, contrary to, for example, visual and auditive perceptual states, hedonic experiences seem to be *responses* to representations of objects and events, rather than themselves representations that inform us about the properties of objects and events. In other words, whereas vision has obvious objective contents (e.g., colours, shapes, textures, faces, etc.), it is very hard to see what is being represented by hedonic experiences—see Block (1995, p. 234); McGinn (1982, p. 8), O'Shaughnessy (1980, pp. 169-70). After all, when someone has an orgasm, it does not seem that she is representing things external to the mind to be in a certain way, in the same way as it occurs in the case when someone activates ORGASM or sees someone having an orgasm. The pleasure we undergo during an orgasm does not seem to be informing us about anything, it seems to simply occur as a free-floating experience that merely "glosses" the content-involving experiences that it accompanies, such as the experience of muscle contractions.

However, as I discussed above (Sections 3.1.2. and 3.1.7.), perceptual systems represent. Therefore, the claim that (dis)pleasure has content directly follows from the claim that (dis)pleasure is a perceptual representation in a dedicated input system, namely, the interoceptive system. Then, *if* the latter claim is the case (i.e., *the interoception tenet*), the claim in question should look more than tempting, even though the content of hedonic experiences is not obvious in the way that the content of other mental states are.

[35] By '*standard* perceptual states' I simply refer to the (online) perceptual states that take place in the sensory modalities typically discussed in the philosophical literature, e.g., vision.

Certainly, this is an assumption that still needs to be more thoroughly justified. However, the idea that (dis)pleasure is a content-involving mental state looks also tempting for reasons independent from the claim that (dis)pleasure amounts to an interoceptive representation.

Firstly, all experiences have associated contents, think of vision, audition, olfaction, etc. Then, it seems arbitrary to exclude (dis)pleasure from the class of content-involving mental states that can be experienced[36].

Secondly, the view that (dis)pleasure has content straightforwardly makes sense of the platitude that pleasure and displeasure are opposites: they are opposites since they have opposite contents. On the account on offer, such contents amount to positive and negative physiological changes. Then the opposition in question is due to the uncontroversial opposition between the latter polarities: the positivity and negativity that physiological changes can take given homeostatic standards. In other words, the polarity characteristic of pleasure and displeasure is inherited from the more fundamental polarity between positive and negative physiological changes. On the other hand, the view that takes (dis)pleasure to be a "free-floating" mental state seems to lack a way of straightforwardly making sense of this phenomenon.

Thirdly, related to the point above, as Schroeder (2004) remarks, if we take (dis)pleasure to be a content-involving mental state, we can readily make sense of the fact that the opposite states of pleasure and displeasure do not occur simultaneously (or that it is extremely rare that that occurs). They generally cancel each other out: the displeasure of being hungry stops as the pleasure of eating takes place; the pleasure of live music stops as the displeasure of having to go to the toilet begins. The representational view readily makes sense of this: Given that pleasure and displeasure inform us about contradictory events (i.e., they have contradictory contents), and that

[36] See Schroeder (2001, pp. 510-511) for a similar line of reasoning.

the mind avoids simultaneously entertaining mutually inconsistent pairs of representations, it is expected that simultaneously having pleasure and displeasure should not occur (or that that should be extremely rare). In turn, the view that (dis)pleasure is a "free-floating" mental state that merely results as a response to mental states that do have content has no resources to account for this sort of mutual inhibition (more on this below).

Fourthly, common sense acknowledges that certain instances of (dis)pleasure are not proper instances of (dis)pleasure, but rather "false (dis)pleasures". For example, the pleasure induced by recreational drugs such as MDMA; or pain in a phantom limb; or the pleasure induced by a neurosurgeon that electrically stimulates the brain of a patient; or the case of a hungry, hypnotized subject that, automatically responding to the instructions of the hypnotist, experiences displeasure after tasting her favourite food. That is, common sense recognizes the existence of hedonic illusions. Considering that illusions are uncontroversially characterized as involving *misrepresentation*, the view that (dis)pleasure is a content-involving mental state is especially well-suited to account for the existence of hedonic illusions (or to put it more strongly, it seems to be the only type of view able to make sense of hedonic illusions).

Finally, if we take (dis)pleasure to be a content-involving mental state, we can readily explain the rationality typically involved in hedonic experiences. Hedonic experiences can be said to be rational in the sense that they provide us with reasons for action. More precisely, given that it is good to have pleasure and bad to have displeasure, those hedonic states provide us with good reasons for seeking and avoidance behaviours, respectively. Consequently, you can properly justify your trying yet another bite of pizza in terms of the pleasure that it brings you, i.e., in terms of the reason you had to act in such a way. If (dis)pleasure is a mental state that lacks content, it is hard to see how it can provide a reason that informs action by interacting in a logically relevant way with other metal states that we take to have contents, such as beliefs, desires and intentions. To put it in other words, experience puts epistemic

constrains on belief and action (or intentions): if you see a dog on the tree, you are justified in forming the belief that there is a dog on the tree, but you are not justified in forming the belief that there is a cat on the tree. In the latter case, you are not so justified since the *content* of your visual experience logically differs in relevant respects from the *content* of your belief. Similarly, in the pizza case above, the pleasure that it causes you does *not* provide you with a reason for avoiding another slice (or intending to avoid another slice)—but it does provide with a reason for having another slice (or intending to have another slice)—since pleasure represents certain events as good, and you intent (we can safely assume) what you take to be good. Certainly, the view that (dis)pleasure is a "free-floating" mental item that "attaches" to content-involving mental states can easily accommodate the sort of rationality exhibited by (dis)pleasure—e.g., by holding that (dis)pleasure provides us with reasons only insofar as it "attaches" to truly content involving mental states. However, taken together, the above considerations strongly suggest that, as any other mental state that can populate phenomenal consciousness, (dis)pleasure also has content.

6.3.3. Motivating the 'homeostasis tenet'

So what is being represented by valence/(dis)pleasure? Considering that I identify valence/(dis)pleasure with interoceptive percepts, I am committed to the view that valence/(dis)pleasure represents what interoceptive percepts represent, namely, positive and negative physiological changes (see Section 3.1. above). This could sound a bit odd to some ears, after all when we introspect during an episode of pleasure, we do not seem to be aware that our physiology is changing in a homeostatically positive direction. We just feel good[37]. Compare this to visual experience. When we introspect during vision, we are not only aware of the content of our percepts, but their content

[37] Certainly, we feel good *about* the external trigger of our pleasure. You are stressed and suddenly run into a beautiful landscape, so you now feel good, in a certain sense, *about* that. However, note that this *aboutness* does not imply that the content of your experience of pleasure amounts to the external stimulus that contributed to trigger your hedonic state, i.e., the beautiful landscape. In this case, the beautiful landscape is the content of the visual experience that *accompanies* your hedonic state and that contributed to trigger the latter.

can be immediately and readily spelled out (at least roughly): if you introspect during a visual experience of a dog, you are readily aware that there is a dog in front of you. Even though there is no need for informational content to be accessible in that way, from a pre-theoretical point of view it does look introspectively odd to claim that, unlike other kinds of experience, (dis)pleasure has as its content such an unfamiliar and abstract-sounding content. Perhaps, the worry goes, (dis)pleasure, if any, exhibits then a less odd sort of content.

However, leaving pre-theoretical intuitions aside, interoceptive percepts only represent homeostatically positive and negative physiological changes (Section 3.1.). Thus, if the view that (dis)pleasure amounts to an interoceptive percept holds, (dis)pleasure should have this sort of "abstract-sounding" content. Interestingly, as we will see below, several reasons point towards the claim that valence/(dis)pleasure does represents homeostatically positive and negative physiological changes. In fact, contrary to the pre-theoretical worry above, this claim turns out to be quite intuitive.

The *homeostasis tenet* is motivated by four main considerations. Firstly, at least since Plato and Aristotle, this claim has been a default view in the affective sciences, not yet truly dismissed by naturalistically inclined approaches. For example, Plato, in the *Philebus,* held that pleasure arises when one perceives (*aisthēsis*) the replenishment of a lack to a state of natural balance. Think of, for instance, the pleasure that you get when you are thirsty and drink water. Even though Aristotle (*De Sensu*) thought that Plato's view did not generalize—since cases such as the pleasure of smelling a flower do not seem to be cases that depend on any sort of replenishment in Plato's sense—he did recognized the existence of pleasures that are indeed contingent on inner physiological states. In these cases, just as in Plato's view, pleasure arises when the cause of a pleasant sensation achieves to restore physiological balance. Fechner (1873) also exploited the insight that physiological balance is key for understanding (dis)pleasure. Taking into account the Mayer-Helmholtz principle of conservation of energy, Fechner held that when an organism's physiological system excessively increases its energy levels (given a certain set-point), a physiological imbalance takes

place, which needs to be rectified so as to keep the organism within its limits of viability. According to Fechner, the physiological imbalance implied by an excess of energy gives rise to displeasure, and the return to balance implied by the dissipation of energy gives rise to pleasure. Most famously, these insights are further developed by the psychophysiologist Michel Cabanac (1971, 1979). Aware of the dependency relation between pleasure and the physiological usefulness of the stimulus in maintaining homeostatic balance, Cabanac put forward the notion of *alliesthesia*. The latter describes the common-sensical, and scientifically well-established phenomenon that, given a certain type of stimulus that is held constant, the polarity of affect changes as a function of the previous physiological condition of the organism's body: the same degree of heat can give rise to pleasure or displeasure as a function of the organism's current core temperature (see Cabanac, 1971, 1979; Panskepp, 1998) (more on this below). According to these views, (dis)pleasure is constitutively dependent on (represented) homeostatic usefulness.

Considering (dis)pleasure as constitutively dependent on (represented) homeostatic usefulness has been a default view in biological approaches to (dis)pleasure (see, e.g., Panskepp, 1998), and in early philosophical approaches to affective phenomena. This view has proven to be experimentally productive in its most recent versions. This, and the fact that it is a default view, suggests that the link between (dis)pleasure and homeostatic usefulness is by no means arbitrary, and that the close connection between (dis)pleasure and homeostatic usefulness might turn out to show that homeostatic usefulness is that to which (dis)pleasure responds, if philosophical inquiry develops further. This kind of view is then worth exploring, and careful philosophical discussion is certainly needed. However, the kind of view in question has practically not taken part in philosophical discussion (outside ancient philosophy), probably because it has not yet been put forward explicitly in the philosophy of mind, where naturalistic approaches to its affective dimension are still rare. I think this kind of view can be brought back to life in philosophy, and become illuminating for the current debate.

Secondly, as I mentioned above, *basic (dis)pleasures* strongly suggest that there is a direct causal relation between homeostatic usefulness and (dis)pleasure, a phenomenon known as *alliesthesia*. By 'basic (dis)pleasures' I simply refer to those (dis)pleasures that are typically assumed to be shared with other mammals, such as the (dis)pleasures of eating, drinking, copulating, pain, hunger, thirst, etc. Basic displeasures are the kind of (dis)pleasures for which most scientific research has been conducted (likely because they can be operationalized without much difficulty, and they can be studied in some detail by carrying out invasive interventions in non-human animals). As I mentioned above, alliesthesia consists in the scientifically well-established phenomenon that, given a certain type of stimulus that is held constant, the polarity of affect changes as a function of the evolving physiological condition of the organism's body. In other words, alliesthesia amounts to the process by which the current homeostatic condition of the body determines the (un)pleasantness of incoming stimuli, insofar as such stimuli modify the organism's physiological landscape (Cabanac, 1971, 1979). As Cabanac puts it: "when the *milieu intérieur* varies, pleasure changes according to the stimulus usefulness for the body." (Cabanac, 2010, p.117). By 'usefulness' it is meant here physiological homeostatic utility. So the basic idea is that certain stimuli cause pleasure in case they re-establish homeostasis, while certain stimuli cause displeasure in case they diminish homeostatic balance. Alliesthesia is a robust phenomenon for all sorts of basic pleasures, including sex and affective touch (see e.g., Beauchamp & Cowart, 1985; Berridge et al., 1984; Cabanac, 1971; Gottfried et al., 2003; Jacobs et al., 1988; Johnson, 2007; LaBar et al., 2001; McCaughey & Tordoff, 2002; McCaughey et al., 2005; Panksepp, 1998; Rolls et al., 1981). For example, induced hypoglycaemia and induced hypoglucy (via insulin and 2-deoxyglucose administration, respectively) increases pleasure for sweet food; while after consuming sugar, sweet stimuli become less pleasant (Cabanac et al., 1968; Cabanac & Duclaux, 1970). Moreover, after immersing a hypothermic subject in a warm bath, and thus changing her condition to a hyperthermic state, warm stimuli previously experienced by the subject as pleasant become unpleasant, and cold stimuli previously experienced by the subject as unpleasant become pleasant. On the other hand, when a subject is immersed in a cool bath, and consequently deep body temperature drops, cold stimuli are progressively experienced by the subject as

unpleasant, while warm stimuli are experienced as pleasant (Bleichert et al., 1973; Cabanac, 1971; Cunningham & Cabanac, 1971). Interestingly, once a physiological disturbance that gives rise to pain begins to be rectified, so that the pain experience vanishes, pleasure takes place. That is, relief from pain, just as the relief from cold, itch, hunger and thirst, involves pleasure. This is exactly what is expected if pleasure arises as a physiological imbalance is rectified (or, more precisely, as a representation of such physiological replenishment is activated) (Andreatta et al. 2010; Grill & Coghill 2002; Leknes & Bastian, 2014; Seymour et al. 2005). Examples like these abound in the literature on 'basic pleasures'. In fact, as Panksepp puts it, "one can readily predict the affective consequences of various external stimuli in humans from a knowledge of bodily imbalances" (Panksepp, 1998, p.181). In a word, the phenomenon of alliesthesia strongly suggests that (dis)pleasure informs us about the homeostatic usefulness of a stimulus (see also Veldhuizen et al., 2010)—i.e., about the homeostatic impact that some stimulus has in the inner milieu (the nature of the stimulus thus being irrelevant).

Thirdly, at least in the case of 'basic (dis)pleasures', (dis)pleasure's evolutionary function is likely to be the fostering of homeostatically useful behaviour by monitoring homeostatically relevant physiological changes. Considering that, according to naturalistic approaches to content (e.g., Dretske, 1981, 1986), the function of a mental state is critical for fixing its content, this also suggests that (dis)pleasure signals homeostatic usefulness (more on this below).

(Dis)pleasure is likely to have a function, rather than being a mere evolutionary by-product. As K.C. Berridge remarks:

> "Brain evolution cannot afford to wastefully dispense the massive amounts of neural machinery that process pleasure on major psychological processes that have no fitness benefit. It is difficult to imagine an evolutionary scenario that would have led to such prominent and similar limbic brains in so many species if pleasure were not adaptive." (Berridge, 2010, p. 13)

So (dis)pleasure is likely to have a function. Now, taking into account the discussion on 'basic pleasures' above, the function of (dis)pleasure seems to be guiding behaviour by signalling which stimuli and actions promote or threaten the viability of the organism (and the organism's group, in the case of social animals). In order to keep itself within its limits of viability, the most fundamental task an organism needs to solve consists in optimally maintaining its physiological functioning. In this regard, during decision-making, (dis)pleasure likely informs about how the organism is faring in achieving this hard-wired goal. That is, (dis)pleasure informs about the impact that stimuli and actions have for survival, i.e., the impact that they have for the physiological, homeostatic condition of the body. In this sense, pleasure can be seen as providing a sense of relief or accomplishment relative to a certain homeostatic goal, allowing thus the re-allocation of attention, and cognitive and motivational resources to other goals: once you have satisfied your hunger, you can keep writing that paper (see also Carver & Scheier 1990; Carver 2003). This view is certainly hard to establish, as it is the case with many evolutionary hypotheses. However, it is worth noting that this claim, namely, that the evolutionary function of pleasure consists in signalling progress relative to homeostatic goals, so as to guide survival behaviour, is a default view in affective sciences (see Kringelbach & Berridge (eds)., 2010, p. 13-14). This claim should not then come as a controversial tenet.

Finally, naturalistic theories of content suggest that, if any, homeostatic, physiological changes should be the content of paradigmatic cases of (dis)pleasure (i.e., 'basic (dis)pleasures). According to naturalistic theories of content (e.g., Dretske, 1981, 1986), a mental state *C* represents *c* in virtue of the fact that *c* causally co-varies with *C* in a reliable fashion, and has the function (by evolutionary or learning history) of causally co-varying in that way. Therefore, in order to show that paradigmatic cases of (dis)pleasure have as content homeostatic, physiological changes it needs to be shown (a) that the latter causally co-vary with the former, and (b) that (dis)pleasure has the function of doing so. The phenomenon of alliesthesia discussed above shows that (a) obtains, and the considerations above regarding the evolutionary functions of (dis)pleasure—i.e., tracking homeostatic states—suggests that (b) also obtains.

On the other hand, let me insist on this point, if (dis)pleasure amounts to an interoceptive percept (*the interoception tenet*), then (dis)pleasure has the content that interoceptive percepts have. As we saw above (Section 6.3.2.), such content amounts to physiological changes that can be either positive or negative, given homeostatic standards. Now, it could still be the case that interoceptive percepts, besides representing positive and negative homeostatic changes, represent also distal, external properties and events (see also Section 4.4.1.). Prinz (2004) argues for this view: interoceptive percepts besides proximally representing physiological changes, they also distally represent core relational themes. However, as we saw in Chapters 1 and 4, that view is rather problematic. Therefore, if (dis)pleasure is constituted by an interoceptive percepts, it represents only positive and negative physiological changes.

Let me briefly digress. Note that this does *not* imply that interoceptive percepts, understood in this way, cannot be associated with representations of external events via learning. Thus, it might be the case that when a represented external event is associated with an interoceptive percept that informs about a negative (positive) homeostatic state, such an external event is taken to be negative (positive) (decision-making likely operates along these lines). In this sense, interoceptive percepts can say to agents "this external thing is bad (good)". It could be argued, then, that interoceptive percepts, in some sense, inform the agent about those aspect of the external world which are good or bad. However, note that 'good' and 'bad' fall short of constituting core relations themes ('demeaning offenses', 'dangers', 'transgressions', 'losses', etc.). Thus, interoceptive percepts do not constitute by themselves emotions *per se*, which inform about core relational themes (Chapter 1). End of digression.

Finally, it is worth also taking into account that even though the claim that (dis)pleasure represents, or informs about homeostatic changes is rather a philosophical claim, it is usually explicitly maintained outside philosophy. For example, Panksepp holds that "pleasure *indicates* something is biologically useful" (Panksepp, 1998, p.182) (italics are mine); and Cabanac claims that "sensory pleasure is both the *sign* of the presence of a useful stimulus and also the motivation (or drive)

[…]" (Cabanac, 1979, p. 2) (italics are mine). Also Craig remarks that "feelings from the body in humans reflect its homeostatic condition" (Craig, 2015, p. 177).

Corns (2014) claims that the homeostatic view on the nature of pleasure cannot account for all sorts of pleasures, because there are types of pleasures that do not seem to have homeostatically specified 'desired' levels. Insofar as there are, let's say, 'desired' glucose levels, the homeostatic restitution of glucose by eating can certainly account for the pleasure of eating. However, the argument goes, insofar as there are no music levels, the homeostatic view cannot account for the pleasure of music (and the like).

Even though this argument looks intuitive at first sight, it is not very compelling. Remember that the homeostatic system has a network-like nature (Section 3.1.9.). That is, homeostasis does not operate in a "thermostat-like" fashion: single, segregated variables (e.g., glucose) each being separately regulated to a certain static built-in value. Rather, during homeostasis, once a local physiological change relative to a certain variable takes place, cascading changes are triggered in the activity of many variables and effectors all over the body landscape, which mutually constrain each other in a network-like fashion. This, in order to maintain an adaptive whole-body physiological state. Now, music causes physiological changes. For example, changes in heart-rate in line with the tempo of music (e.g., Edworthy & Waring, 2010; Karageorghis et al., 2006; Trappe, 2010). Given the network-like nature of homeostasis, those physiological changes can trigger compensatory changes in several physiological variables and effectors (e.g., blood circulation and temperature). In certain contexts, the resulting body landscape can result adaptive or maladaptive, having thus a homeostatic impact. For example, in certain kinds of contexts, achieving certain heart-rate levels is adaptive, given the task at hand: When an agent needs to be particularly active, in need to engage in physical action, increased heart-rate levels result adaptive. Think of the case of running or cleaning the flat. In these cases, people enjoy listening to up-tempo music, and slow-tempo music results annoying. The opposite also seems to be the case. When an agent needs to rest after excessive activity,

slow-tempo music is enjoyable, but up-tempo music is annoying. The pleasure of music, at least in these cases, does seem to have its basis on homeostatic processes (see Edworthy & Waring, 2010; Karageorghis et al., 2006). Even more, as Habibi and Damasio have compelling showed, the pleasure of music seems to have its basis on homeostasis in general, not only in the above cases (Habibi & Damasio, 2014). Even in the less intuitive case, namely, the pleasure of sad music, the pleasure of music seems to have its basis on homeostasis (Sachs et al., 2015). Music does seem to play a homeostatic role. The point raised by Corns certainly requires much more discussion. Taking into account that the topic of this Thesis is not the nature of pleasure, I leave that discussion for another occasion, as it demands a longer, dedicated treatment.

It has also been argued that (dis)pleasure cannot amount to bodily perception, because cases such as skydiving suggest that the same physiological changes (those that skydiving supposedly typically triggers) occur in both, someone who enjoys the experience, and in someone who dislikes the experience (Schroeder, 2004).

As we saw in Section 5.4.1., regarding the PP account of some aspects of placebo hypoalgesia, the PP framework has the resources to straightforwardly reply to this sort of objection. As we saw, in the PP framework, the same input from the body can give rise to different kinds of percepts, depending on different assignments of precision-weighting. Thus, given different high-level expectations about valence value, different 'portions' of the incoming interoceptive PE are attended and ignored, given the different assignment of precisions which such different expectations determine. This results in different interoceptive percepts of the physiological, homeostatic condition of the organism, just as we saw regarding placebo hypoalgesia (see Section 5.4.1.). Thus, cases such skydiving are no threat for a PP-based interoceptive view on the nature of valence.

Finally, it could also be argued that subjects can represent their bodily, physiological state without having a valenced state. Thus, the argument goes, interoceptive percepts

do not constitute valence. For example, someone can perceive her heart raising, but feel neither good nor bad. Or, to give a more typical example, someone can perceive tissue damage without feeling bad (pain asymbolia). This is certainly the case. However, this fact does not imply that interoceptive percepts formed via interoceptive perceptual inference do not constitute valence. In the view suggested in this Chapter, interoceptive percepts represent whole-body positive/negative physiological changes. Remember that *whole-body* positive/negative physiological changes are represented at higher levels of the interoceptive hierarchy (Section 3.1.5.). A key interoceptive high-level region is the anterior insula. As I mentioned in Section 3.3.4., it is hypothesized that the anterior insula functions as a comparator between expected and incoming interoceptive signals, during interoceptive percept formation, among other functions (Seth et al., 2011). In other words, the anterior insula functions as an interoceptive PE calculator module. To put it colloquially, this region is key then for telling the system "this whole-body physiological landscape deviates from (approaches) the expectation of homeostasis". Thus, it is only after this comparison has taken place that interoceptive percepts get to represent bodily changes which are *positive* or *negative*, given homeostatic standards. As I discussed in this Chapter, this is the sort of content that matters for constituting valence. Then, it is certainly the case that bodily changes can get to be represented without the latter being represented as positive/negative in that sense. This occurs at lower levels of the interoceptive hierarchy (e.g., in the mid-insula, see Craig, 2015). In this case, the comparison between expected and incoming interoceptive signals has not taken place at the relevant hierarchical level, which occurs in the anterior insula. However, the claim defended in this Chapter is that represented bodily changes constitute valence, only insofar as they are represented as positive or negative, given the high-level expectation of homeostasis maintenance.

To sum up, several reasons motivate the three main commitments of the claim that (dis)pleasure consists in an interoceptive percept that *informs* us about how we are faring in dealing with the hard-wired goal of maintaining homeostasis (keeping 'surprisal' low). Firstly, the fact that (dis)pleasure is a mental state that can be felt strongly suggests that (dis)pleasure must be a percept of some sort, and the best candidate seems to be a percept in the interoceptive system (*the interoception tenet*).

This intuition is also supported by certain strands of empirical evidence. Secondly, contrary to what might be initially thought, that fact also suggests that (dis)pleasure is a content-involving mental state (*the content tenet*), since percepts typically have content. The content tenet is also suggested from the fact that all experiences have content, and from the fact that this tenet makes sense of key platitudes regarding (dis)pleasure: the platitude that pleasure and displeasure are opposites, that there are "false" (dis)pleasures, and that hedonic states exhibit a certain rationality. Finally, the tenet that the content of (dis)pleasure amounts to homeostatically positive or negative physiological changes (*the homeostasis tenet*) is motivated by the following reasons. In the first place, historically speaking, holding that there is a direct causal relation between homeostatic usefulness and (dis)pleasure has been a default view in the affective sciences. This shows that endorsing such a link is by no means arbitrary. In the second place, the phenomenon of alliesthesia strongly suggests that there is indeed such a direct causal relation between homeostatic usefulness and (dis)pleasure, at least in the case of 'basic (dis)pleasures'. In the third place, there are reasons that lead us to speculate that the evolutionary function of (dis)pleasure is precisely monitoring homeostatically relevant physiological changes so as to guide behaviour. Finally, naturalistic theories of content suggest that, if any, homeostatic, physiological changes should be the content of paradigmatic cases of (dis)pleasure.

Taken together, these considerations strongly suggest the plausibility of the claim in question: when an interoceptive percept is formed, it constitutes positive valence/pleasure in case it represents homeostatically positive physiological changes, and it constitutes negative valence/displeasure in case it represents homeostatically negative physiological changes. Moreover, as we saw in Section 3.2., several strands of evidence points towards the claim that, during homeostatic regulation, interoceptive percepts are formed by a process that operates under the principles of PP. Certainly, I do not take these considerations as conclusive in any case. But they do show that the view defended in this Chapter is a plausible, tenable view, which can result theoretically and experimentally fruitful. More research is certainly needed.

Then, in the view defended in this Chapter—i.e., ITV— the valence properties which (partially) constitute emotion count as sensory, interoceptive representations of homeostatic changes. The above suggest that this view might be the case. However, as Prinz (2004, 2010) and others have argued, theories which identify valence with mental states that can be felt—e.g., (dis)pleasure/perceptual representations—face several problems. In the coming Section I will argue that such criticism is misguided. I will show that ITV can reply to the objections typically faced by this kind of views. Consequently, ITV constitutes itself as a promising view.

6.4. Replying to objections

The view on the nature of valence that emerges from the PP framework claims then that valence is a representation in a sensory system. According to ITV, the valence properties which (partially) constitute emotion count as sensory, interoceptive representations of homeostatic changes. Insofar as valence is taken to be a sensory representation, it is a mental state that can be felt. As I mentioned above, the claim that valence can be felt is typically put forward by views that identify valence with (dis)pleasure. In this sense, ITV is somehow committed to the view that valence is (dis)pleasure, at least in that respect. This is emphasized in the view on valence put forward by Joffly and Coricelli (2013, p.8). Thus, the view on valence that emerges from the PP framework can be taken to be a version of the *hedonic theory of valence*, which identifies valence with (dis)pleasure. However, as Prinz (2004, 2010) and others have argued, theories which identify valence with mental states that can be felt—e.g., (dis)pleasure/perceptual representations—face several problems. I discuss them below.

6.4.1. Not all emotions are pleasant or unpleasant

Remember that ITV is somehow committed to the view that valence is (dis)pleasure, at least in the sense that it holds that valence can be felt, insofar as valence amounts to a sensory, interoceptive representation. Thus, ITV also counts as a version of the hedonic theory of valence, so ITV is also target of the objections raised against this kind of view.

Firstly, it has been argued that hedonic theories of valence, or theories that hold that valence is a felt component part of emotion, do not satisfy the *scope desideratum*. This desideratum states that all clear cases of emotion should exhibit the property or mechanism that is proposed as accounting for valence. It has been argued that these views do not satisfy the *scope desideratum*, since not all positive or negative emotions are pleasant or unpleasant, feel good or bad, respectively. For example, it has been claimed that it is not clear in what way a negative emotion like anger involves displeasure (Deonna & Teroni, 2012, p. 15).

I do not think that such an intuition poses a real challenge to ITV. I am not convinced by it. Try as I may, I cannot remember any time I felt anger without also feeling hedonically bad at the same time. A bit of experimental philosophy could solve this point. Nonetheless, intuitions about cases won't let us go far. The issue is not about the folk *concept* of positive and negative emotions; it is about valence as a *natural kind*. Anyway, ITV does not need to commit to the strong claim that all *emotion types* have a distinctive 'hedonic tone', either pleasure or displeasure, a good or bad feeling. ITV only needs to be committed to the weaker claim that every time an emotion takes place it has valence, be it positive (pleasure), negative (displeasure), or mixed (both pleasure and displeasure). Then, it could be the case that sometimes anger is negatively valenced and other times positively valenced (or has a mixed valence), which could motivate the intuition that it is not clear that anger always involves displeasure.

6.4.2. Not all negative (or positive) emotions feel the same

Secondly, it has been argued on phenomenological grounds that ITV fails satisfy the *scope desideratum* in another way. It has been argued that negative valence cannot be identical to the feeling of displeasure, because not all negative emotions feel the same.[38] Negative emotions, it is argued, feel bad in different ways, so they cannot share a common felt aspect, namely, the specific 'hedonic tone' of displeasure (Prinz, 2010; Solomon, 2003). There seems that negative emotions do not have "a common phenomenological denominator" (Prinz, 2010, p.7). Therefore, the feeling of displeasure cannot be present in the many different negatively valenced emotions. Displeasure cannot be identical to negative valence. In other words, the wide variation in the way in which negative (or positive) feelings feel indicates that they do not share a common perceived felt aspect, namely, negative (or positive) valence. Thus, valence is not something perceptual, i.e., something that can be felt.

This objection is quite compelling. However, it could only work if ITV were committed to the following two claims:

> **(*a*)** The phenomenology of an emotion is exhausted by its valence component.

> **(*b*)** Displeasure consists in a single type of distinctive, narrowly circumscribed feeling, which is the exact same type of narrowly circumscribed feeling across every instance of any negative emotion.

[38] This argument also applies to positive valence. For ease of exposition I put it in terms of negative valence.

If the phenomenology of an emotion is exhausted by its valence component (which is identified with an interoceptive percept or (dis)pleasure by ITV)—**(*a*)**—and displeasure amounts to just one type of narrowly circumscribed feeling—**(*b*)—**one can certainly conclude that variations in the phenomenology of negative emotions undermines the hypothesis that negative valence is the feeling of displeasure. However, ITV does not need to be committed neither to **(*a*)** nor to **(*b*)**. Let me begin with **(*b*)**.

6.4.2.1. Displeasure is a circumscribed, uniform feeling

The claim that negative valence consists in an interoceptive percept *tout court* or to displeasure *tout court*, and that all negative emotions have such a percept or displeasure as a component, is distinct from the much stronger claim ***(b)*** above. That is, the claim that negative valence consists in a single, narrowly circumscribed type of interoceptive percept or displeasure, and that such single, narrowly circumscribed type of percept or displeasure is shared by all instances of negative emotions. ITV needs only be committed to the first, weaker claim.

To put it this way, **(*b*)** is analogous to the quite implausible claim, **(*b*')**, that the experience of red that is a component part of our visual experiences of mammalian blood is the exact same type of narrowly circumscribed experience across every instance of the visual experience of mammalian blood, let's say, the experience of red_{24}: when one visually experiences dog blood in contexts c_1, c_2, and c_n, the experience of red involved in such an experience is the experience of red_{24}; when one experiences cat blood in contexts c_1, c_2, and c_n, the experience of red involved in such an experience is also the experience of red_{24}, and so on. Such claim differs from the much weaker, plausible claim that, in normal conditions, every instance of the visual experience of mammalian blood has as a component the experience of red *tout court*. This allows that when one visually experiences blood samples from different mammals the experience of red involved in those experiences varies: in one case it could be red_{24},

in another case red_{27}, in another case red_{22}, and so on. In this latter case, even if it is the case that experience varies, the claim that the visual experience of mammalian blood has as a component the experience of red still holds. This analogy could be extended to other kinds of experience and to other modalities besides vision, *mutatis mutandis*. I see no reason why it should not also be extended to the case of the experience of displeasure, especially considering that, as ITV holds, displeasure or negative valence is grounded in a perceptual modality (interoception). Thus, even if it is the case that displeasure experiences varies across instances of negative emotions, the claim that negative valence consists in displeasure *tout court* still holds.

On the other hand, if displeasure or negative valence amounts to an interoceptive representation, as ITV holds, variability in the experience of displeasure is actually to be expected. As I discussed in Section 3.1., interoceptive percepts bind together several physiological/interoceptive features. In other words, interoceptive percepts can be seen as an 'amalgam' of different physiological features. It is not implausible to suppose that in different occasions and contexts such an 'amalgam' will be composed of different homeostatic, physiological features (analogously to vision), changing thus the configuration of the percept that will eventually be experienced (for example, in certain occasions the interoceptive representation in question will have heart activity represented and in other occasions not). Thus, it is unlikely that the experience of displeasure is uniform in any way.

In fact, this follows naturally from the PP framework. As I mentioned above, in the PP framework, the assessment of the precision of PE, which is identified with attention, occupies a central role in the whole PP inferential machine. Now, our expectations of precision depend on context. This implies that, during percept formation, and depending on context, the assignment of precision-weighting can differ within the same modality for different sensory features. Thus, ITV can straightforwardly explain the rich variation in our experience of positive (and negative) feelings by appealing to the context-dependent variation of precision-weighting for different sensory attributes during interoceptive percept formation. For example, let's take the case of hunger, as

it is an uncontroversial case of an affective state grounded in interoception. Hunger involves a distinctive set of physiological changes. Among them, a decrease in blood sugar levels, a decrease in blood levels of amino acids, and the fluctuation of leptin and ghrelin hormone levels. Now, let's say that you just played an especially tough football game. So, besides the physiological changes characteristic of hunger, a set of physiological changes relative to the usual injuries of sport are also taking place at the same time. For example, pain in your legs, increased heart-rate, and inflammation. However, given the time of the day, and all that energy expenditure during the game, the expectation that hungry should be taking place drives interoceptive percept formation. This implies that, on the one hand, precisions of interoceptive PE for the interoceptive attributes expected for the perceptual hypothesis of hunger are highly weighted. On the other hand, precisions of interoceptive PE relative to the features characteristic of pain in your legs, increased heart-rate, and inflammation are ignored to a certain extent. In other words, the latter interoceptive signals are not much attended; while the interoceptive signals relative to hunger are highly attended. However, let's say that now you need to walk to the kitchen to make you a sandwich, even though you are hurt. Then, precisions of interoceptive PE relative to the features characteristic of pain in your legs, increased heart-rate, and inflammation it is inferred now to have high precision-weighting. These interoceptive attributes are now also highly attended and become salient. That is, your conscious perception of your inner milieu includes now such attributes in a more distinct fashion. Thus, during interoceptive perception, expectations of precisions determine which features become more salient. Thus, on the assumption that valence is constituted by an interoceptive percept, the wide variation in the way in which negative (or positive) feelings feel can also be accounted for by way of the context-sensitive, differentially weighted assignment of precisions during interoceptive percept formation.

Furthermore, the PP framework offers another reply to the objection in question. Remember that PP proposes a hierarchical architecture of the mind. Perceptual systems are constituted by several levels of processing, in which one level attempts to predict the activity of the level below so as to minimize precision-weighted PE. Lower levels of the hierarchy encode regularities that operate at fast time-scales, which capture

variant aspects of experience. On the other hand, higher levels encode increasingly more complex regularities that operate at slow time-scales, which capture relatively more invariant aspects of experience. Interestingly, precisions can be differentially assigned at each level of the perceptual hierarchy. Thus, the same bounded interoceptive features that compose an interoceptive percept can be, depending on context, differentially attended. In this manner, bodily experience can exhibit some variation.

6.4.2.2. The phenomenology of an emotion is exhausted by its valence component

Let's now consider **(*a*)**, the claim that the phenomenology of an emotion is exhausted by its valence component. One might still not be convinced that ITV is not committed to **(*b*)**. However, generally speaking, when one component of experience remains the same, while the other components vary, experience as a whole varies. Insofar as emotion is built out of components, being valence one of them, it is not implausible to suppose that the same goes for the experience of emotion. Thus, even if negative valence or displeasure is a uniform, narrowly circumscribed experience, it could still be the case that we experience negative emotions differently simply because the context in which they tend to arise varies, and, during emotions, represented aspects of such context are also experienced along with valence, which supposedly remains uniform. Among such contextual aspects one could include, for example, imagery associated to the thoughts that take place during emotions, imagery associated to episodic memories relative to previous emotional episodes, motor imagery associated to the simulation of actions for coping, etc. There is no reason that supports the claim that all that we experience during a certain emotion is only one of its many components, namely, valence (and certainly ITV does not need to commit to such a claim). Thus, negative emotions feel different because the experienced context in which they tend to arise varies, while their hedonic tone remains the same (if we have to assume that). If this is the case, variation in the phenomenology of negative emotions does not count against the claim that all negative emotions share one common felt aspect, namely, displeasure, as ITV claims. Let me make the following

analogy. Consider the following chords: *Em11b5*, *A7*, and *Dm7*. *A7* and *Dm7* are composed of four pitches and *Em11b5* is composed of five pitches. However, these three chords only share a single, narrowly circumscribed exact pitch, namely, *A*; while they differ with respect to the other pitches. Even though they share one exact common felt aspect, namely, the pitch *A*, they certainly feel quite differently. Generally speaking, when one component of experience remains the same, while the other components vary, experience as a whole varies. Insofar as emotion is built out of components, being valence one of them, it is not implausible to suppose that the same goes for the experience of emotion.

To sum up, ITV does not need to be committed neither to **(*a*)** nor to **(*b*)**. Therefore, the fact that the phenomenology of negative emotions varies is no objection to ITV. The latter fact can be easily accommodated by ITV.

6.4.3. Valence can be non-conscious

Thirdly, Prinz (2004) has argued that valence cannot be identified with pleasure and displeasure, because pleasure and displeasure are necessarily always conscious (in the phenomenal sense of 'conscious'): there seems to be no such thing as non-conscious pleasures or displeasures. However, emotions can be non-conscious (Ledoux, 1996; Winkielman & Berridge, 2004; Winkielman et al., 2005). So, if emotions are always valenced, and emotions can be non-conscious, the construct of valence needs to be able to occur outside consciousness. Therefore, Prinz concludes, we cannot identify valence with pleasure and displeasure. Remember that the claim that valence can be felt is typically put forward by views that identify valence with (dis)pleasure. In this sense, ITV is somehow committed to the view that valence is (dis)pleasure (Joffly & Coricelli, 2013). Then, ITV is also susceptible to this criticism.

This objection only works if the claim endorsed by ITV were that positive and negative valence amount to the *folk concepts* of pleasure and displeasure, respectively. However, scientific enquiry deals with natural kinds, not with our concepts for those kinds. Given that ITV is a scientific hypothesis about the nature of one of the components of the mechanism of emotion, the claim in question cannot concern the folk *concept* of pleasure and displeasure, but rather the *natural kinds* to which 'pleasure' and 'displeasure' refer. Certainly, according to the current folk conception, 'pleasure' and 'displeasure' refer exclusively to conscious states, but the natural kinds *pleasure* and *displeasure* do not need to be as the folk conception dictates. It is already common knowledge that usually our folk conception about the properties of a certain target phenomenon *C* ends up differing significantly from what scientific research tell us about the nature of *C*. I see no reason why the case of pleasure and displeasure should escape this trend. In fact, in the scientific literature on valence, one can already begin to see scientist willing to talk of pleasure as not necessarily conscious (see, e.g., Berridge & Kringelbach, 2010, p.7-8).

This makes plenty of sense. As I argued above, under the view defended in this Chapter, representations of negative valence or displeasure are perceptual representations. Now, perceptual states do not become immediately conscious. Some additional process must intervene for a perceptual representation to become a *conscious* perceptual representation. For example, some claim that such a perceptual state must be represented by another, higher level state (Lycan, 1996; Rosenthal, 2005), or that it must be globally broadcast (Baars, 2002), or that it must be modulated by attention (Prinz, 2012).

Hohwy (2012) has proposed a compelling view on the relation between attention (precisions), model accuracy, and conscious perception in the PP framework that is relevant for this matter, and that it can be used so as to deal with the objection in question. The idea is that perceptual states get to be conscious when the exhibit a relatively high degree of accuracy and precision. Let me briefly explain. In statistical terms, 'accuracy' refers to the inverse amplitude of PE *per se*. That is, under some

simplifications, models from which perceptual hypotheses are put forward exhibit more *accuracy* in case they better represent the causal structure of the world, as the more PE is minimized. Remember that *precisions* refer to the variability of, or uncertainty about PE. In the PP framework, precision is attention. Now, accuracy and precision double dissociate:

> "It is a trivial point that precision and accuracy can come apart: a measurement can be accurate but imprecise, as in feeling the child's fever with a hand on the forehead or it can be very precise but inaccurate, as when using an ill calibrated thermometer. This yields two broad dimensions for perceptual inference in terms of predictive coding: accuracy (via expectation of sensory input) and precision (via expectation of variability of sensory input). These can also come apart. Some of the states and parameters of an internal model can be inaccurate and yet precise (being confident that the sound comes from in front of you when it really comes from behind, Jack and Thurlow, 1973). Or they can be accurate and yet, imprecise (correctly detecting a faint sound but being uncertain about what to conclude given a noisy background)." (Hohwy, 2012, p. 4).

Given this dissociation between precisions and accuracy, and depending on context, a certain perceptual state can exhibit relatively high accuracy and high precision. However, they can also be accurate but imprecise (at different degrees of accuracy and precision), and *vice versa*.

The PP notions of accuracy and precisions offer a nice solution to the objection that valence should be able to occur outside consciousness. Sensory states can be non-conscious in cases were their accuracy and precision are both relatively low (Hohwy, 2012). Insofar as, under the hypothesis of ITV, negative valence or displeasure is represented in the brain as any other piece of sensory or perceptual information is represented, there mere fact that the brain entertains a representation of negative valence or displeasure does not make such a representation a conscious representation. It should tend to have, under this view, relatively high accuracy and precision. If this is the case, displeasure or negative valence can be non-conscious, just as any perceptual representation. Accordingly, non-conscious displeasure amounts to a representation of displeasure that does not exhibit the required degree of accuracy and

precision; while conscious displeasure amount to a representation of displeasure that does exhibits a relatively significant degree of accuracy and precision.

6.5. Satisfying desiderata

ITV seems to satisfy all the desiderata for a theory of valence. Firstly, as we saw in Section 6.4.1 and 6.4.2, contrary to what has been claimed, ITV can successfully accommodate the *scope desideratum*. On the one hand, intuitively, all clear cases of positive (negative) emotions feel pleasant (unpleasant), and counterexamples to this view fail, as ITV does not need to commit to the strong claim that all *emotion types* have a distinctive 'hedonic tone', either pleasure or displeasure, a good or bad feeling. On the other hand, the fact that not all negative (positive) emotions feel the same fails to show that all clear cases of positive (negative) emotions feel pleasant (unpleasant). This is the case since the phenomenology of an emotion is not exhausted by its valence component, and ITV does not need to hold that valence or (dis)pleasure consists in a single type of distinctive, narrowly circumscribed feeling. Moreover, if hedonic feelings are grounded in interoceptive perception, as ITV clams, emotions should have an interoceptive component. As we saw in Section 6.3.1., this is precisely what certain strands of evidence suggest.

Secondly, ITV can easily accommodate the *pre-theoretical taxonomy desideratum*. Clear cases of positive emotions, such as joy are pleasant or feel good, and also clear cases of negative emotions, such as fear and guilt, are certainly unpleasant or feel bad. Furthermore, during clear cases of emotion we certainly feel our bodies fluctuating in pleasant and unpleasant ways. This suggest that clear cases of emotion do involve a valenced interoceptive component.

Thirdly, ITV straightforwardly accounts for the *feeling* and *evaluative desiderata*. Organisms have the hard-wired expectation of maintaining homeostasis. Positive

emotions feel good and negative emotions feel bad (*feeling desideratum*) since their valence component represents positive and negative homeostatic changes, respectively (Section 6.3.3). Interoceptive percepts make things positive and negative to us (they make things matter to us) since they indicate the objects and events which are likely to promote or threaten homeostasis (*evaluative desideratum*).

Finally, remember that the interoceptive system grounds inherently motivational states by way of what I called internal and external actions—think of hunger and thirst—(Sections 3.1.9.1. and 3.3.2.2.). The idea is that, as an interoceptive percept informs about the homeostatic condition of the organism, interoceptive actions are automatically, and by default, motivated so as to rectify the homeostatic state about which such a percept informs. This seems to be a characteristic property of the interoceptive system (Craig, 2014). This explains the fact that neuroimaging studies show that the anterior insula and the anterior cingulate activate in an intertwined manner during emotion (Murphy et al., 2003). The anterior cingulate plays the role of triggering interoceptive and proprioceptive 'policies' during homeostasis maintenance (Craig, 2014). Thus, the view that identifies valence with interoceptive perceptions straightforwardly accounts for the *motivation desideratum*.

In this Chapter, I argued that, contrary to the arguments of defenders of the view that valence cannot be a sensory phenomenon (as the 'interoceptive inference' approach must claim), ITV can successfully reply to these pressing objections. Moreover, as I argued in Chapter 5, the view that valence amounts to a non-sensory signal faces decisive problems. Thus, the view that valence arises by way of interoceptive perceptual inferences of the causes of interoceptive afferents emerges as a promising view. However, as we saw in Chapter 4, the interoceptive inference view of *emotion per se* is problematic. This leaves the PP framework without an account of emotion. This is major drawback for PP ambitions, as the principles of PP promise to give us a unifying account of all the seemingly disparate variety of mental phenomena, ranging from perception to action, and everything in between. This includes emotion. However, in the coming Chapter, I will argue that the PP approach to interoception

can indeed be used to account for emotion *per se*. This requires amending IIE. I will argue that, rather than via interoceptive perceptual inference, as IIE holds, emotions arise by minimizing interoceptive PE via *external interoceptive actions*. If this view is on track, PP's ambitions are safe: interoceptive PE minimization can account for affect and emotion, the core of our mental life.

7. Concluding Chapter: Emotion as active interoceptive inference

In the previous Chapter, I defended the view that affective valence can be taken to be constituted by an interoceptive percept, formed via interoceptive perceptual inference, which represents positive and negative homeostatic changes (ITV). This view, which maintains that valence amounts to a sensory phenomenon, can reply to the objections made by defenders of non-sensory theories of valence (and satisfies agreed desiderata for a theory of valence). ITV constitutes itself then as a promising view.

However, as I discussed in Chapter 4, interoceptive, perceptual theories of emotion are indeed doomed to fail—as there are no emotions configured in the physiological landscape. Perceiving/feeling our physiology cannot be then the *whole* story about emotion *per se*. Contrary to IIE's claim, predicting interoceptive signals during perceptual inference cannot be what is primary in emotion generation. Thus, Seth's (and Hohwy's) IIE lacks a thoroughly compelling way to account for emotion *per se*.

Then, even though PP can account for affective valence, the PP framework is left without a compelling view on the nature of emotion. This is a major drawback for PP ambitions. Remember that the PP framework promises to give us a unifying account of *all* the seemingly disparate variety of mental phenomena, ranging from perception to action, and everything in between. This includes emotion.

However, allegedly in line with IIE, common sense (and also experimental research) tells us that every time we experience an emotion, this experience is accompanied by interoceptive feelings. This suggests that, even though having an emotion does not simply consist in perceiving interoceptive changes, as IIE claims, having an emotion does involve some type of process that must be intertwined with interoception. In other words, something more that interoceptive percept formation is needed so as to account

for emotion, and it must be something closely intertwined with interoceptive, bodily perception.

In this concluding Chapter, I suggest that, by amending IIE in a key respect, the PP approach to interoception can indeed be used to account for emotion *per se*. As I argued in Chapter 4, interoceptive *perceptual* inference cannot be the primary driving component of emotion. However, as I mentioned above, interoceptive perception must be a central part of the story about emotion generation. Therefore, assuming that the PP framework is on track, emotions must arise by minimizing interoceptive prediction error, and in another fashion than as IIE proposes. If it is not simply via *perceptual* inference, we are left then with *active* inference. Emotions must arise then via interoceptive *active* inference, instead of via interoceptive perceptual inference.

In this Chapter, I will explore this insight. Emotions are not about forming an interoceptive percept of a certain sort. Rather, emotions are strategies for changing an interoceptive percept that has already been formed (via interoceptive perceptual inference). This percept constitutes affective valence (Chapter 6). That is, emotions are specific strategies for regulating affective valence. Now, interoceptive percepts (i.e., valence) inform about our homeostatic condition. Then, emotions are better seen as specific strategies for regulating homeostasis.

More precisely, the idea is that emotions arise by minimizing interoceptive PE via *external interoceptive actions*. Here the task consists in minimizing the discrepancy between already formed interoceptive percepts and the hard-wired expectation (or 'goal') of stable homeostasis. As we saw in Chapter 3, *external interoceptive actions* require stored knowledge of 'sensorimotor contingencies'. In this Chapter, I propose the view that a certain emotion *E* amounts to a strategy for minimizing interoceptive PE by way of a specific set of representations of 'sensorimotor contingencies'. That is, by way of stored knowledge of the counterfactual relations that obtain between (possible) actions and its prospective interoceptive, sensory consequences. An emotion

arises when such knowledge is applied in order to regulate valence. In this sense, emotions are specific strategies for regulating affect by way of specific forms of action-guiding stored knowledge. According to this view, emotions are then individuated by the kind of stored knowledge of 'sensorimotor contingencies' that is brought to bear in the control of valence—plus by what they represent. As long as emotions are not identified here with interoceptive perceptions, this view does not require there to be regularities pertaining to emotion in the inner milieu. Therefore, if the proposed view holds, PP's ambitions are safe: interoceptive PE minimization can account for affect and emotion, the core of our mental life.

I begin this Chapter by briefly motivating the view that emotions might be forms of action (rather than forms of perception) (Section, 7.1.). Then, after presenting the claim that emotions arise by minimizing interoceptive PE via external interoceptive actions (Section 7.2.), I discuss the idea that the kind of stored knowledge of 'sensorimotor contingencies' required for external interoceptive actions can be taken to consist in emotion-specific action-oriented representations, encoded at higher levels of the cortical hierarchy. Insofar as these representations (or 'chunks' of the generative model) encode abstract (Section 7.3.) sets of knowledge about the same category in the world, they can be taken to consist in emotion 'concepts' (Section 7.2.2). The latter have both, mind-to-world and world-to-mind direction of fit (Section 7.4.). In Section 7.5., I highlight the way in which the suggested expansion of IIE can account for those aspects that should be taken into account when developing a satisfactory view on the nature of emotion.

7.1. Emotion as 'active' interoceptive inference?

Emotions have motivational force. Emotions are motivational states that urge us to act in different ways. I take this to be rather uncontroversial. Many common sense phenomena point towards the centrality of motivated action in emotion. Let me mention just a couple of such phenomena. In the first place, the very existence of

virtues speaks of the quintessentially motivational character of emotion. Virtues such as self-discipline, resilience, prudence, and temperance, amount precisely to the ability to control the motivational force of emotions. These character traits would not be virtues in the first place, if emotions lack motivational power in their very constitution. In the second place, we commonly appeal to the motivational force of emotions for the sake of explanation: "Christine rapidly hid her bottle of gin because she was scared of the police", "Michelle made loud noises in the middle of the night because she was secretly angry at her husband John". People have the urgency to retaliate in burst of anger, to kiss out of love, and to repair damage out of guilt. As I mentioned in Chapter 1, sometimes these urges also impel us to do little or nothing at all—which are also things we do. For example, freezing out of fear, napping out of sadness, lying hours on the couch out of shame. All this sorts of action exhibit a sense of urge, a 'motivational oomph', which is accompanied by the expectation that such an urge will vanish after action completion.

In a word, during emotion, motivated action turns out to be quite fundamental. That is why, as we saw in Chapter 1, throughout the history of emotion research, the motivational, action-oriented aspect of emotion has been regarded by some researchers to be the primary aspect of emotion. Other kinds of families of emotion theory identify emotion (or take as the primary, driving aspect of emotion) with other aspects typically involved during emotion, such as feelings/perceptions and judgments. As we also saw in Chapter 1, both these kinds of theories (and also hybrid theories) struggle with accounting for key phenomena involved during emotion. Thus, action looks as a more than promising place at which to look in order to better understand emotion. It is worth then exploring whether the action-oriented aspect of emotion might be the 'essence' of emotion, to put it this way.

On the other hand, there is a strong folk-psychological intuition that emotion *does* consist in feeling/perceiving our physiology reacting in some sort of way. This intuition is suggestively supported by scientific research, as meta-analytic studies show that brain regions involved in interoception are consistently active during

emotion, and, besides that, other measures that reliably indicate interoceptive activity are found to occur during emotion (e.g., Cacioppo et al., 2000; Murphy et al., 2003).

Thus, we have that both, action and physiological perception look as central aspects of the generation of emotion episodes. This suggests that both aspects are somehow intertwined during emotion generation.

Interestingly, as I discussed in Chapter 3, interoception takes place as the organism attempts to regulate homeostasis. Interoceptive percepts inform the organism about its current homeostatic condition. However, given the fact that inner physiological 'policies' (e.g., releasing vasopressin in the case of thirst) can rarely rectify homeostatic imbalances by triggering inner physiological resources alone (i.e., we simply lack the physiological resources to re-hydrate ourselves by producing water), the interoceptive system engages actions in the external environment in order to rectify homeostatic imbalances (e.g., looking for some water). Motivating action is part of what the interceptive system does, to put it this way (see Craig, 2015; Devinsky et al., 1995). This is what I called in Chapter 3 *external interoceptive actions* (allostatic actions). In other words, the interoceptive system amounts to a perceptual system, which is also inherently motivational. Interoception is the action-oriented perceptual system *par excellence*.

The above suggests that external *interoceptive active* inference might be the primary aspect during the generation of emotion episodes. Furthermore, on the assumption that interoception is key during emotion, this is also suggested by the fact that interoceptive *perceptual* inference cannot be the whole story as to how emotions arise (Chapter 4). Consequently, and as I mentioned above, assuming that the PP framework is on track, and that there are only two ways of minimizing interoceptive PE, we must embrace the other horn, namely, interoceptive *active* inference. It is worth then considering the hypothesis that emotions arise as interoceptive PE is minimized via *external interoceptive active inference*.

I suggest that emotions arise via *external interoceptive active inference*: by sampling and modifying the external environment in order to change an already formed interoceptive percept (which has been formed via interoceptive perceptual inference). This percept constitutes valence, and informs about homeostatic imbalances (Chapter 6). Thus, emotions are specific strategies for regulating affective valence, and consequently, homeostasis. More precisely, a certain emotion *E* amounts to a strategy for minimizing high-level interoceptive PE by way of a specific set of stored knowledge of the counterfactual relations that obtain between (possible) actions and its prospective interoceptive, sensory consequences ("if I act in this manner, interoceptive signals should evolve in such-and-such way"). An emotion is generated when such knowledge is applied in order to regulate valence (i.e., affect). When high-level interoceptive PE is minimized via the set of 'sensorimotor contingencies' that corresponds to stored knowledge about emotion *E*, the emotion *E* is generated. Emotions are specific strategies for regulating affect by way of specific forms of action-oriented stored knowledge.

The idea is that, initially, a certain event triggers physiological changes in the organism. This event is typically triggered by an exteroceptively perceived external event. For example, a letter stating that your landlord needs to take back the property. However, the event in question can also be internally triggered. For example, the physiological changes that result from a bad posture, or a poor night of sleep. The physiological changes that have been triggered by some external (or internal) event are interoceptively perceived as positive or negative physiological changes, given homeostatic standards—i.e., given the hard-wired expectation of stable homeostasis (Chapters 3 and 6). These percepts are formed via interoceptive *perceptual* inference, as described in Chapters 3 and 6. Remember that physiological changes can be taken to be *good* or *bad*, *positive* or *negative*, according to whether they tend to *approach* or *deviate* from the aimed-at (flexible) regulatory level of homeostasis maintenance, respectively (Chapter 3). In case an interoceptive percept represents homeostatically

positive physiological changes, positive valence (or pleasure) takes place. In case an interoceptive percept represents homeostatically negative physiological changes, negative valence (or displeasure) takes place (Chapter 6).

The discrepancy between an already formed interoceptive percept that informs about current homeostatic condition and the hard-wired expectation (or 'goal') of stable homeostasis constitutes *high-level interoceptive PE*. In other words, negative valence reflects states which are incompatible with the high-level expectation (or 'goal') of maintaining homeostasis. Note that, in a certain sense, *positive* valence also reflects states which are incompatible with the high-level expectation of homeostasis maintenance. This is the case since positive physiological changes amount to changes which are *approaching* the 'goal' set by homeostatic standards. That is, such physiological changes are not yet quite in line with the standard in question. The phenomenon of *allostasis* shows that this is the case (Cabanac, 1971). Pleasure typically takes place as a homeostatic imbalance begins to be rectified. However, pleasure stops as such an imbalance has already being rectified (Cabanac, 1971, 1979). Think of the 'homeostatic motivation' of hunger, and its corresponding process of satiation. When an organism is hungry and eats something nutritious, the pleasure obtained from that stimulus is significant. However, as the organism in question already begins to be satiated, the hedonic value of food decreases, to the point that, as the organism is already satiated, food tends to become aversive (Cabanac, 1979). In this sense, pleasure is a form of 'ongoing relief'.

Now, we have then that high-level interoceptive PE is triggered in case that, after comparison, current perceived physiological changes differ from the expected 'goal' state of physiological balance. Such perceived physiological changes constitute valence (i.e., affect). As I discussed in Chapter 3, the main task of the interoceptive system is *not* now forming a percept, but rather bringing physiological variables to their expected state by minimizing such high-level interoceptive PE. In the PP framework, this means that active inference needs to be engaged: actions must be bring forth so as to fulfil predictions. Taking into account the fact that the organism cannot

minimize high-level interoceptive PE via internal interoceptive actions (Chapter 3), external interoceptive actions are motivated. As I discussed in Chapter 3, external interoceptive actions consists in changing the external environment in order to change an interoceptive percept that constitutes valence. Insofar as external interoceptive actions are a form of active inference, they require representations of 'sensorimotor contingencies': counterfactual knowledge of the way in which interoceptive signals would evolve, if certain actions ensue.

As I am proposing that emotions are driven by external interoceptive actions, the models for emotions need to store such kind of knowledge. The view I am suggesting is that when high-level interoceptive PE is triggered by any sort of event, and it is minimized via the set of 'sensorimotor contingencies' which is stored in the, let' say, 'anger-model', the emotion of anger arises and it is experienced—it is experienced in case there is a significant degree for accuracy and precision, as discussed in Chapter 6. The 'anger-model' can be taken to be that 'chunk' of the generative model of the organism that stores expectations about anger (more on this below). Then, when such knowledge is activated in order to minimize high-level interoceptive PE, the emotion of anger unfolds.

7.2.1. Emotions as 'homeostatic motivations'

Note that this view sees emotion in analogy with 'homeostatic motivations' or drives, such as for example, hunger. Hunger amounts to a mental state that is constituted by both, the negatively valenced state of an empty stomach, plus the motivation to act in the world in such a way so as to change such a bad feeling. Analogously, emotions, if the suggested view is on track, are also constituted by both, a valenced state, plus the motivation to act in such a way as to change such state. However, there is an important difference between emotions *per se* and homeostatic motivations. While 'homeostatic motivations' are directed at inner states: the negatively valenced lack of nutrients, emotions are directed at external events, namely, core relational themes (Chapter 1).

More precisely, emotions exhibit two layers of content. On the one hand, their valence component represents inner states, namely, positive or negative homeostatic changes (Chapters 3 and 6). On the other hand, the emotion-specific action-oriented knowledge that is brought to change valenced states, and thus minimize high-level interoceptive PE, represents core relational themes.

7.2.2. Representing core relational themes

As I have been claiming, such emotion-specific action-oriented knowledge—knowledge of 'sensorimotor contingencies'—specifies actions through which interoceptive PE can be minimized. Active inference has world-to-mind direction of fit. How is it, then, that such knowledge can get to represent the world, i.e., core relational themes? In other words, how is it that emotions, understood as forms of action, can inform about states of the world, having thus mind-to-world direction of fit?

As I mentioned above, emotion models ('anger-model', 'fear-model', 'guilt-model', etc.) can be seen as different 'chunks' of the organism's generative model. These 'chunks' encode expectations about emotion. Different emotions encode different expectations relative to emotion-specific 'sensorimotor contingencies'. At higher levels of the cortical hierarchy, such emotion-specific expectations are relatively abstract, encoding slow time-scale expectations (Chapter 2). Insofar as such knowledge encodes expectations about distinct categories ('anger, 'fear', 'guilt', etc.), and such expectations are abstract, it is not arbitrary then to consider such 'chunks' of knowledge as concepts, i.e., emotion concepts (see Hohwy, 2013, pp. 72-73). Concepts have mind-to-world direction of fit. Emotion concepts should be no exception to this (see, e.g., Barrett, 2006a). Thus, the distinct 'chunks' of knowledge which encode emotion-specific expectations relative to 'sensorimotor contingencies' can also be taken to have, besides world-to-mind direction of fit, also mind-to-world direction of

fit. Let me expand this point below, and discuss the key aspects of the claim presented in this subsection which have not been discussed in previous Chapters.

7.2.2.1. Knowledge of 'sensorimotor contingencies' and emotion 'concepts'

The view suggested in this Chapter can then be put in the following way. Emotions arise as emotion-specific action-oriented representations, encoded at higher levels of the cortical hierarchy, are used to minimize high-level interoceptive PE. In other words, a certain emotion *E* arises as the action-oriented emotion 'concept' for *E* is used to control valence.

In this view, emotion concepts can be understood as those representations of 'sensorimotor contingencies' which are encoded at higher levels of the cortical hierarchy. Remember from Chapter 2 that, in the PP framework, lower levels of the hierarchy encode regularities that operate at fast time-scales, which capture variant aspects of experience. On the other hand, higher levels encode increasingly more complex regularities that operate at slow time-scales, which capture relatively more invariant aspects of experience. The higher-levels of the hierarchy are not modality specific (i.e., amodal and multimodal), and encode even more abstract, slow time-scale regularities, not directly related to regularities pertaining to the domain of just one modality. In a word, higher levels encode relatively abstract expectations. This suggests that, in the PP framework, the traditional distinction between perceptual representations and conceptual representation dissolves into the distinctions regarding the levels at which expectations are encoded (see Hohwy, 2013, pp. 72-73).

The high-level 'chunks' of stored knowledge that encode expectations relative to the same category, for example, the category *dog*, can then be taken to constitute the 'concept' DOG. Now, as I remarked in Chapters 2 and 3, implicit in the PP framework is the idea that, particularly at higher levels of processing, stored representations

relative to a certain category must encode knowledge of 'sensorimotor contingencies' relevant to that category (see also Seth, 2015). Thus, the 'chunks' of stored knowledge that constitute a 'concept' for a certain category must then encode knowledge of 'sensorimotor contingencies'. In other words, the 'concept' for a certain category specify actions relevant for that category. For example, DOG includes knowledge relative to ways of actively interacting with dogs in appropriate ways, given that category. More precisely, it includes knowledge about which evolving sensory states to expect, across all levels of the perceptual hierarchies, if one would interact with a dog in such-and-such a way. 'Concepts' for emotions should be no exception to this (see, e.g., Barrett, 2006a; Wilson-Mendenhall et al., 2011).

This approach to 'concepts' in the PP framework is by no means arbitrary. This view is directly in line with the situated view on the nature of concepts (e.g., Barsalou, 1999, 2009). In fact, the latter has been articulated in terms of the PP framework (Barsalou, 2009). Roughly, the view on categorical inference put forward by Barsalou (Barsalou, 2009) can be put in the following way. The categorical inferences that a certain concept permits take place by way of the top-down activation of the stored sensory/perceptual and motor states which constitute the concept. Let's call these perceptual and motor states 'sensorimotor states'. Such sensorimotor states can be taken to be the 'features' of the concept. These sensorimotor states get to be constituent 'features' of a concept through exposure. That is, as the agent encounters instances of a certain category *c* in the world, the relevant sensorimotor states that consistently occur during the worldly interaction with *c* are stored. In this manner, such sensorimotor states become stored priors about which sensory states to expect, given the hypothesis that an instance of *c* is taking place. The concept *C* is thus formed. Then, during categorical inference, the sensorimotor states in question are generated from the top-down in order to guide adaptive behaviour. Let me briefly illustrate an important aspect of this way of understanding categorical inference.

Considering that a concept consists here in a set of sensorimotor states that consistently occur during interaction with instances of the relevant category, once activated, such

concepts allow the generation of predictions or "educated guesses about what might occur next" (Barsalou, 2009, p. 1284). This takes place via pattern completion. For example, let's imagine an agent, Sophie, who sees a dog from her window, and, at t_1, activates the visual 'parts' or 'features' of the model/concept for dogs. Given that certain kinds of sounds—i.e., barks—tended to consistently co-occur with instances of dog previously experienced by Sophie, when the 'auditive part' of her model/concept for dog activates at t_2 via pattern completion, her brain can now predict the kind of sound that this dog will likely make. Considering that certain kinds of canine behaviours tended to co-occur with certain food odours, when the motor, visual, and auditive 'parts' of her model/concept for dog activates at t_3, her brain can predict how this dog will likely react to the sausage odour that comes out of her flat. Even more, given that certain kinds of human-dog interaction tended to consistently co-occur with instances of dog previously experienced by Sophie, her brain can predict how to properly interact with this dog. Categorical inferences based on stored expectations about sensorimotor states, once activated, control subsequent actions via pattern-completion. This should also apply to emotion concepts (and concepts for any category) (see, e.g., Barrett, 2006a; Wilson-Mendenhall et al., 2011).

In the case of emotion concepts, they can be taken to be constituted by the sensorimotor states that consistently occur during the encounters of the agent with instances of an emotion type. By 'instances of an emotion type' I mean the kinds of situations which the culture to which the agent belongs typically refers as characteristic of the emotion type in question (see Barrett, 2006a). For example, let's take the case of anger. In western culture, anger typically involves a demeaning offense. This is the core relational theme of anger (Chapter 1) (Lazarus, 1991). In situations in which demeaning offenses take place, people tend to do things that eliminate the origin of the offense in question. For example, people tend to raise their tone of voice, threaten other people with physical violence, move their faces in certain ways to express disproval, etc. The high-level sensorimotor states that consistently tend to occur during the kind of situation that we recognize as characteristic of anger become then stored priors. In this manner, ANGER— i.e., the high-level 'chunk' of stored knowledge that encodes expectations relative to anger—is formed. Then, during categorical inference, the

sensorimotor states in question are generated from the top-down in order to guide adaptive behaviour. This means that such knowledge is used in order to minimize high-level interoceptive PE, so as to maintain the organism within viability limits. In other words, the system knows that, in a situation characteristic of anger, if certain actions are performed, interoceptive signals should evolve in such a way as to reduce the difference between the expectation of stable homeostasis and the current perceived physiological state. Emotion 'concepts'—i.e., that 'chunk' of the generative model which encodes high-level knowledge relative to a certain emotion category—encode then knowledge of 'sensorimotor contingencies'.

7.3. 'Sensorimotor contingencies' and the cortical hierarchy: active inference and its levels of granularity

The architecture posited by the PP framework is inherently hierarchical. Thus, knowledge of 'sensorimotor contingencies' must also be found across all levels of the cortical hierarchy (Chapter 2). As we saw in Chapter 2, representations of 'sensorimotor contingencies' can be low-level or high-level depending on how variant or invariant are the regularities that they encode, respectively. For example, low-level, fast-changing actions include movements such as microsaccades. Slower time-scale actions include arm movements, or walking. Even more 'abstract', slower-timescale actions can include actions such as waiting for the night to fall, doing a PhD, or working as a Lecturer.

In the PP framework, higher-level precision-weighted expectations of action constrain and modulate lower-level proprioceptive predictions. If the system has the high-level expectation (or 'goal') of eating, this can be achieved, depending on context, by several different cascades of lower-level precision-weighted proprioceptive predictions. For example, and depending on context, the system can achieve the expectation of eating by extending the arm, walking to the fridge, cycling to the supermarket, etc. In turn, these lower-level predictions (or 'sub-goals') can be fulfilled in several different ways

depending on context. In fact, the lower in the proprioceptive hierarchy, the more the context-dependent variability of the precision-weighted predictions in question: the relatively low-level expectation (or 'goal') of grasping your mug, can be fulfilled via very distinct predictions about shoulder and wrist micro-movements, depending on what the context affords—e.g., your initial position, room temperature, metabolic resources, etc. Higher levels constrain and modulate lower levels. Higher levels encode expectations of action which are coarse-grained, while lower levels encode expectations of action which are fine-grained (the latter seem to be rather automatic, while he former seem to be more 'intentional').

It follows from the above paragraphs that the expectations of action that each emotion model/concept encodes so as to minimize high-level interoceptive PE must also be seen as represented at different time-scales or levels of abstraction. That is, the actions specified by the sensorimotor contingencies encoded by emotion concepts/models exhibit different degrees of granularity. There are expectations of action relative to emotion which are very abstract. For example, and to keep the example of anger above, the expectation (or 'goal') of eliminating the origin of a demeaning offense. There are also expectations of action which are relatively lower level. The latter amount to context-sensitive ways of fulfilling the high-level expectation (or 'goal') in question. In this case, the abstract prediction in question can be fulfilled by several distinct lower-level expectations (or 'sub-goals'). For example, attacking, making a phone call, making an ironic joke, sighing, etc. In turn, these lower-level expectations can be fulfilled by an even richer array of relatively lower-level proprioceptive predictions. For example, attacking can be fulfilled by running towards the offender, or by slowly walking towards the offender while expanding the chest, etc. In turn, these latter expectations of action can be fulfilled by several lower-level proprioceptive predictions, and so on and so forth. In a word, knowledge of sensorimotor contingencies exhibit different degrees of granularity. The same high-level expectation of action can be fulfilled by several distinct lower-level predictions. Such high-level expectations constrain and modulate lower-level proprioceptive predictions in a context-sensitive manner.

This distinction between levels of abstraction relative to the expectations (or 'goals') that emotion models/concepts encode is thoroughly compatible, *mutatis mutandis*, with the distinction between *relational goals* and *situated goals* made by Scarantino (2014):

> "[…] *relational goals* are *abstract goals* that need to be situated in a *concrete context* in order to guide bodily changes. This is typical of most goal-oriented processes, including non-emotional intentional actions. When we decide to get to school by 10am in order to attend a talk, the *overarching action goal* of getting to school by 10am can be achieved through a variety of *situated goals* (e.g., taking a bus at 9:20am, taking the subway at 9:30am) (cf. Pacherie 2008). Each of these situated goals can in turn be achieved by a variety of *motor goals* that directly guide bodily changes. For simplicity of reference, I will distinguish between the *relational goal* of an emotion and its *relational sub-goals*, understood as the collection of *situated* and *motoric* goals by which the relational goal can be achieved." (Scarantino, 2014, p. 169)

In this Chapter I suggest the view that emotions arise as emotion-specific action-oriented representations (i.e., emotion 'concepts'), encoded at higher levels of the cortical hierarchy, are used to minimize high-level interoceptive PE. These representations encode knowledge of 'sensorimotor contingencies'. When high-level interoceptive PE is minimized via the set of sensorimotor contingencies that corresponds to stored knowledge about emotion *E*, the emotion *E* is generated. The claim is that the kind of knowledge in question is high-level. That is, it specifies action expectations which are abstract (i.e., 'relational goals'). For example, in the case of anger, the expectation of eliminating the origin of a demeaning offense. In this view, emotions are individuated by those emotion-specific abstract expectations.

7.4. Emotion 'concepts' as 'pushmi-pullyu' expectations

Importantly, emotion concepts/models, understood as the high-level 'chunks' of stored knowledge that encode action expectations relative to emotion categories, exhibit two intertwined representational aspects. Remember that emotion concepts track the external conditions that consistently occur during the encounters of the agent with instances of an emotion type. Such external conditions consist in situations which the culture to which the agent belongs typically refers as characteristic of the emotion type in question (e.g., "that's anger", "you are feeling shame", etc.) (see Barrett, 2006a). For example, anger typically occurs when a demeaning offense takes place. This is the core relational theme of anger (Chapter 1) (Lazarus, 1991). However, in situations in which demeaning offenses take place, people also tend to *do* things that eliminate the origin of the offense in question. It is likely then that such regularities also get to be encoded by emotion models. Thus, emotion concepts/models get to represent core relational themes as they jointly encode "educated guesses" as to which actions correspond to that kind of situation (core relational theme), in order to deal with latter (Section 7.3.).

In other words, emotion concepts get to represent core relational themes as they encode action expectations relative to which action should take place in order to obtain the evolving interoceptive signals that would minimize high-level interoceptive PE, given such core relational theme. In this sense, emotion concepts have both, mind-to-world and world-to-mind directions of fit. They represent descriptively the way in which the world is (i.e., core relational themes), and how the world is to be. That is, the high-level 'chunks' of knowledge that encode emotion-specific 'sensorimotor contingencies' count as *pushmi-pullyu representations*: representations that jointly have facts and 'goals' as their content (Millikan, 1994, 2004) (see Scarantino, 2014).

7.5. Explanatory advantages

In this Section, I highlight the way in which the expansion of IIE suggested in this Chapter can account for those aspects that should be taken into account when developing a satisfactory view on the nature of emotion, as stated in Chapter 1. That is, in this Section, I show how the view in question can satisfy the discussed desiderata for a theory of emotion (Chapter 1). Remember that there are three main general approaches to emotion theories, namely, perceptual theories, cognitive theories, and agential theories. The view suggested in this Chapter has the resources to account for those aspects in which the main three general approaches to emotion theories result satisfactory, and those aspect in which they result satisfactory. As I commented in Chapter 1, insofar as the PP framework dissolves to an important extent the division of perception, cognition, and action, the expansion of IIE suggested in this Chapter is then particularly well positioned to integrate the insights of these three approaches, while avoiding their typical problems.

Let's get down to business. As we saw in Chapter 1, a theory of emotion must explain how emotions are meaningful. That is, it must explain how emotions can represent core relational themes. Emotions exhibit intentionality. While cognitive theories straightforwardly account for the meaningfulness of emotion, as they identify emotions with judgments, perceptual and agential theories have difficulties in accounting for this phenomenon. Interestingly, being an agential theory, the view suggested in this Chapter accounts for the meaningfulness of emotion in a similar way as cognitive theories, without turning into a version of this kind of theory. In the view suggested in this Chapter, emotions are about core relational themes, since emotions are constituted by emotions concepts/models. The latter, as any other kind of concept, have mind-to-world direction of fit. That is, they represent the world as having certain properties. Given that emotion concepts/models track the external conditions that consistently occur during the encounters of the agent with instances of an emotion type, and that core relational themes consistently occur during emotions, emotion concepts represent core relational themes.

Secondly, as discussed in Chapter 1, a theory of emotion must explain how emotions can arise without there being emotions configured in the physiological landscape, taking into account the fact that there typically is a bodily aspect to the phenomenology of emotion, and that phenomenal consciousness is populated by nothing over and above percepts.

The expansion of IIE suggested in this Chapter accounts for the bodily aspect characteristic of the phenomenology of emotion by appealing to its valence component. Remember that, in the view suggested in this Chapter, emotions arise as high-level, emotion-specific knowledge of sensorimotor contingencies is used in order to minimize high-level interoceptive PE. The latter amounts to the difference between the expectation of stable homeostasis and the perceived physiological condition of the organism. Valence is the construct that informs organisms about their current physiological condition. As it was discussed in Chapters 5 and 6, valence arises by way of interoceptive perceptual inferences of the internal causes of interoceptive signals. Valence amounts to the perception of such physiological changes. As valence is a different affective construct than emotion, this view on the nature of valence does not require there to be emotions configured in the physiological landscape. This view only requires there to be physiological changes which are positive or negative in some sense. As it was discussed in Chapters 3, 5, and 6, as valence is coarser-grained than emotion, and that there are patterns of physiological changes that distinguish between valence properties, this is an assumption which should not be taken to be controversial. Then, the interoceptive system can extract regularities regarding valence properties from the physiological landscape, so as to use them later for the sake of perceptual inference. Thus, the expansion of IIE suggested in this Chapter satisfies the *desideratum* in question: valence is an interceptive percept, and it is a component part of emotion. Valence understood in this way does not require there to be physiological regularities relative to emotion. Moreover, given that valence is a component part of emotion, and that, being a percept, it can be felt, the view proposed in this Chapter can account for another desideratum discussed in Chapter 1: A theory of emotion must account for the fact that emotions are intrinsically bodily felt.

Thirdly, as discussed in Chapter 1, a theory of emotion must explain the fact that emotions can occur outside consciousness. As it has been claimed in number of occasions throughout this Thesis, I am assuming that phenomenal consciousness is only populated by percepts. Then, emotions must be a perceptual phenomenon of some sort, as long as they can be felt (and this is certainly the case). Now, as we saw in Chapter 6, Hohwy (2012) suggests that perceptual states get to be conscious when the exhibit a relatively high degree of accuracy and precision. Remember that, roughly, models from which perceptual hypotheses are put forward exhibit more *accuracy* in case they better represent the causal structure of the world, as the more PE is minimized. Remember that *precisions* refer to the variability of, or uncertainty about PE. Depending on context, a certain perceptual state can exhibit relatively high accuracy and high precision. However, they can also be accurate but imprecise (at different degrees of accuracy and precision), and *vice versa*. Importantly, as Hohwy (2012) suggests, perceptual states can be non-conscious in cases were their accuracy and precision are both relatively low (Hohwy, 2012). Now, in Chapters 5 and 6, I argued for the view that valence arises as interoceptive PE is minimized via interoceptive *perceptual* inference. Valence consists in an interoceptive percept. If this view is on track, valence can be non-conscious, just as any perceptual representation. Non-conscious valence is a sensory representation of the inner milieu, which does not exhibit the required degree of accuracy and precision.

However, in the expansion of IIE suggested in this Chapter, besides valence, emotions are constituted by high-level knowledge of 'sensorimotor contingencies'. An emotion arises when such knowledge is applied in order to regulate valence. In fact, in this view, the driving component of emotion generation is that kind of emotion-specific action expectations. Interoceptive *active* inference is the driving component of emotion. Thus, considering that only formed percepts populate phenomenal consciousness, this component itself cannot be felt. Only its interoceptive, sensory consequences can be felt, once they are 'explained away' via interoceptive *perceptual* inference. Valence is the component of emotion that can be felt. This means that, in the suggested expansion of IIE, emotions can take place outside consciousness, only insofar as their valence component can be non-conscious.

Now, haven't I argued in Section 6.4.2. that the phenomenology of an emotion is *not* exhausted by its valence component? In a certain sense, this is the case, as I argued in that Section. However, in another sense, all that which is felt during an emotion episode is its valence component. Certainly, as an episode of emotion generation takes place, many different sorts of mental states and processes *accompany* such an episode, and many of them can be felt. For example, during an episode of emotion, besides the perception of some physiological changes, we also seem to think and reason in certain kinds of ways. We also move our limbs in some ways, and say things to ourselves in inner speech. We also focus our visual and auditive attention in certain features of the environment more than in others features of the environment, among other accompanying mental states and processes. Insofar as these aspects, which do not constitute proper component parts of emotion (Chapter 1), occur together with an emotion, the experience as a whole goes beyond the sole experience of valence (perceived bodily changes). In this sense, the phenomenology of an emotion is *not* exhausted by its valence component (Section 6.4.2.) However, as it is suggested in this Chapter, if emotion has only two constituent components, namely, valence and expectations of action, and valence is the only component which is inherently perceptual, then only valence is the proper component part of emotion that can be felt. Certainly, the emotion-specific expectations of action which are used to minimize high-level interoceptive PE can have associated imagery. However, the latter is not inherent to the former.

Fourthly, as we saw in Chapter 1, a theory of emotion must account for the fact that emotions have motivational force. As a version of the *agential* theory of emotion, in which emotion generation is driven by action expectations dependent on the interoceptive system, the view suggested in this Chapter straightforwardly satisfies this *desideratum*. As we saw in Section 6.5., the interoceptive system grounds inherently motivational states by way of what I called external actions (Sections 3.1.9.1. and 3.3.2.2.). The idea is that, as an interoceptive percept informs about the homeostatic condition of the organism, interoceptive actions are automatically, and by default, motivated. This, in order to rectify the homeostatic state about which such a percept informs—think of hunger and thirst. This seems to be a characteristic property of the

interoceptive system (Craig, 2014). This explains the fact that neuroimaging studies show that the anterior insula and the anterior cingulate activate in an intertwined manner during emotion (Murphy et al., 2003). The anterior cingulate plays the role of triggering interoceptive and proprioceptive 'policies' during homeostasis maintenance (Craig, 2014). The motivational character of the interoceptive system also explains the fact that lesions to the anterior insula significantly impact motivation (e.g., Devinsky et al., 1995). Thus, the view that identifies emotion with the external interoceptive actions driven by the perceived homeostatic condition of the organism directly accounts for the 'motivational oomph' characteristic of emotion.

Finally, as discussed in Chapter 1, a theory of emotion must explain the close connection between emotion and action, without identifying emotion types with specific sets of instrumental behaviours. As I just mentioned above, the link between emotion and action is straightforward in the view suggested in this Chapter. Now, as we saw in Chapter 1, a typical problem of agential theories of emotion is that, in emphasizing the action-oriented, motivational aspect of emotion, agential theories tend to be committed to the claim that each emotion type has a characteristic set of instrumental behaviours. However, as it was discussed in Chapter 1, emotions cannot be individuated by sets of instrumental behaviours. Different sets of instrumental behaviours are involved in the same emotion type, and the same type of behaviour is involved in different types of emotion.

Closely following Frijda (1986, 2010), the view suggested in this Chapter avoids this problem by holding that the expectations of action which individuate emotion types are encoded at higher levels of the cortical hierarchy. At these levels, models encode slow time-scale regularities, which exhibit a rather abstract level of granularity. That is, these levels do *not* encode specific sorts of instrumental behaviour and motor 'policies'. The latter are situationally driven, as I commented above.

The emotion-specific knowledge of 'sensorimotor contingencies' that individuate emotions encodes expectations ('goals') relative to the types of problem with which a certain emotion type consistently needs to deal. Following Frijda (1986, 2010), the emotion-specific knowledge of 'sensorimotor contingencies' that individuate *anger* can be taken to consist in expectations relative to the task of *regaining control of action to remove obstruction* (see Frijda, 1986, p. 88). In the case of *fear*, the emotion-specific knowledge of 'sensorimotor contingencies' that individuate it can be taken to consist in expectations relative to the task of *making oneself inaccessible to the relevant stimulus so as to avoid it*. These actions are engaged since the system predicts that, given the core relational theme which is considered to be taking place, the actions in question will achieve to trigger interoceptive signals compatible with the expectation of homeostatic balance.

The expansion of IIE suggested in this Chapter exhibits another explanatory advantage. Remember from Chapter 4 the reading of IIE according to which the claim that IIE puts forward is that basically *all* emotion hypotheses expect the *same* interoceptive states. In this reading of IIE, perceptual interoceptive hypotheses about which emotion might be taking place encode expectations about which affective properties should occur given such a considered emotion. The affective properties in question can be taken to consist in an arousal state together with its hedonic value (i.e., valence), which are assumed to be states in the interoceptive system. All emotions involve arousal, and the task of interoceptive inferences is to infer which emotion is causing such an arousal state. The reason for doubting this reading of IIE is that evidence suggests that, during perceptual tasks relative to emotion processing, first an interoceptive percept is formed, which constitutes the experience of affect (i.e., arousal together with its hedonic value, or valence at a certain level of intensity), and only then the processes involved in the generation of emotion *per se* begin to occur (see Barrett et al., 2007). The expansion of IIE suggested in this Chapter directly makes sense of this key piece of evidence. In this view, once the organism is informed about a homeostatic change by an *already* formed interoceptive percept, then action-oriented emotion knowledge is brought to do the further processing required for fulfilling the expectation of stable homeostasis.

On the other hand, the agential view suggested in this Chapter straightforwardly accounts for those aspects which are left unexplained by the current philosophically more developed agential theory of emotion, namely, the motivational theory of emotion (MTE) (Scarantino, 2014) (Chapter 1).

In the first place, as we saw in Chapter 1, MTE leaves unexplained an aspect that any agential theory must explain, insofar as it gives to action a primary role in the generation of emotion episodes. Agential theories, insofar as they are action theories, should say something about why emotions have the motivational force that they have, by appealing to the resources that the proposed theory itself provides. MTE is silent in this respect. The agential theory that, as it is suggested in this concluding Chapter, emerges out of the PP framework has a straightforward answer. Emotions have the motivational force that they have, because they are grounded in the interoceptive system. The latter, as we saw, motivates action so as to maintain the organism within viability limits.

In the second place, the resources that MTE itself provides give no answer to the problem of valence. That is, what is it that makes certain emotions positive and other emotions negative? The agential theory suggested in this Chapter, as we saw, puts valence at the centre of emotion. Emotions are nothing but ways of dealing with valence. In this view, certain emotions are positive (negative) emotions because they contain a bodily representation that informs the organism that the physiological landscape is approaching (deviating form) the expectation of homeostasis.

7.6. Final remarks

In this Thesis I aimed at systematizing, clarifying, and problematizing the interoceptive inference approach to interoception in such a way so as to be able to

make a case for two claims relevant for the philosophy of emotion and the affective sciences.

In the first place, there are different possible ways in which valence can be understood in the PP framework. In some way or another, all these views point towards the same more global PP view on the nature of valence. All these views point towards the claim that valence amounts to an interoceptive percept, formed via interoceptive perceptual inference, that informs about positive and negative homeostatic changes. This is what I called 'the interoceptive theory of valence' (ITV). I defended this claim from non-sensory theories of valence (NSS). The latter is the view that valence does not amount to a sensory phenomenon of any kind, but it rather consists in a non-sensory signal that merely "attaches" to sensory representations. As I argued, besides the fact that ITV can reply to the objections made defenders of NSS to the claim that valence amounts to a sensory phenomenon, NSS fail to show their tenets to be true. This leaves the door open for the view that interoceptive percepts formed via interoceptive perceptual inference can indeed account for affective valence. In fact, several strands of empirical evidence (and some theoretical considerations) suggest that this is the case. ITV can reply to objections, and satisfies agreed desiderata for a theory of valence. Thus, ITV stands out as a promising view on the nature of valence.

In the second place, I discussed and problematized the interoceptive inference view of emotion (IIE). This view holds that emotions *per se* are a forms of perception. More precisely, according to IIE, in direct analogy to visual perception, emotions arise by minimizing interoceptive prediction error (PE) via interoceptive perceptual inference. I argued that this view exhibits a problematic assumption. However, I think that the more general claim put forward by IIE is on track. Emotions do arise by minimizing interoceptive PE. That is why I concluded this Thesis by suggesting one way in which emotions can be understood along the lines of the interoceptive PE minimization approach. I suggested that emotions arise by minimizing interoceptive PE in a manner distinct from the way IIE claims. I concluded by suggesting that emotions are primarily a form of action, not of perception. That is, emotions arise via *external interoceptive*

active inference: by sampling and modifying the external environment in order to change an already formed interoceptive percept. That is, emotions are specific strategies for regulating affective valence. More precisely, I suggested the view that a certain emotion *E* amounts to a specific strategy for minimizing interoceptive PE by way of a specific set of stored knowledge of 'sensorimotor contingencies'. An emotion arises when such knowledge is applied in order to regulate valence.

As I argued in this concluding Chapter, the suggested view turns out to be a promising view, insofar as it avoids the problems of the main kinds of theories of emotion, and it satisfies the desiderata discussed in Chapter 1. Then, the discussion presented in this Chapter sets the basis for an especially dedicated, focused treatment of the view that emotions arise via external interoceptive active inference, together with the development of its philosophical implications. Future research on PP and the generation of emotions *per se* should focus on these affairs. In this Thesis, I focused on the first part of that project: the systematization, clarification, and discussion of the interoceptive inference approach to interoception as applied to valence and emotion. I expect to have shown relevant ways in which the emerging PP account of emotion and valence can be philosophically problematized. I also expect to have made a compelling case for the views of valence and emotions *per se* that emerge from the PP perceptual machinery, and that they constitute themselves as promising views, worthy of further developments and discussion.

References

- Andreatta, M., Mühlberger, A., Yarali, A., Gerber, B., and Pauli, P. (2010). A rift between implicit and explicit conditioned valence after pain-relief learning in humans. *Proceedings Biological Sciences*, *277*, 2411–2416.

- Arnold, M. B. (1960). *Emotion and Personality*, vol. *1*. New York: Columbia University Press.

- Baars, B. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences*, *6*, 47–52.

- Bain, D. (2003). Pain and Intentionalism. *Philosophical Quarterly*, *53*, 502-523.

- Barlassina, L., & Newen, A. (2014). The role of bodily perception in emotion: in defense of an impure somatic theory. *Philosophy and Phenomenological Research*, *89*, 637–678.

- Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion*, *12*, 579-599.

- _____. (2006a). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review, 10*, 20–46.

- _____. (2006b). Are emotions natural kinds? *Perspectives in Psychological Science, 1*, 28–58.

- _____. (2015). Construction as an integrative framework for the science of emotion. In L. F. Barrett and J. A. Russell (Eds.) *The psychological construction of emotion* (p. 448-458). New York: Guilford.

- Barrett, L. F., & Bar, M. (2009). See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), *1325–1334*.

- Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology*, *41*, 167-218.

- Barrett, L.F., Lindquist, K., & Gendron, M. (2007). Language as a context for emotion perception. *Trends in Cognitive Sci*ence, *11*, *327–332*.

- Barrett, L. F., Quigley, K., Bliss-Moreau, E., & Aronson, K. R. (2004). Interoceptive sensitivity and reports of emotional experience. *Journal of Personality and Social Psychology*, *87*, 684-697.

- Barrett, L. F., & Russell, J. A. (1999). Structure of current affect. *Current Directions in Psychological Science, 8,* 10–14.

- Barrett, L. F. & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, *16*, 419-429.

- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

- _____. (2009). Simulation, situated conceptualization, and inference. *Philosophical Transactions of the Royal Society of London: Biological Sciences, 364*, 1281-1289.

- Beauchamp, G. K., & Cowart, B. J. (1985). Congenital and experiential factors in the development of human flavor preferences. *Appetite, 6*(4), 357-372.

- Ben-ze'ev, A. (2010). That Thing Called Emotion: On the Nature of Emotional Experiences. In P. Goldie (Ed.), *Oxford Handbook of Philosophy of Emotion*. Oxford: Oxford University Press.

- Bechara, A. & Damasio, A. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*, *52*, 336-372.

- Bechara, A., Damasio, A.R., Damasio, H., Anderson, S.W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*, 7-15.

- Berman, B.D., Horovitz, S.G., Hallett, M. (2013). Modulation of functionally localized right insular cortex activity using real-time fMRI-based neurofeedback. *Frontiers in Human Neuroscience*, *7*, 638.

- Berntson, G. G., & Cacioppo, J. T. (2007). Integrative physiology: Homeostasis, allostasis and the orchestration of systemic physiology. In Cacioppo, J. T., TassELinary, L. G., &

Berntson, G. G. (Eds.). *Handbook of Psychophysiology*, 3rd edition, (pp. 433-452). Cambridge, UK: Cambridge University Press.

- Berridge, K.C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology, 191*, 391–431.

- _____. (2010). In M.L. Kringelbach & K.C. Berridge (Eds.). *Pleasures of the brain* (p.10). New York: Oxford University Press.

- Berridge, K. C., & Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron*, *86*, 646–664.

- Berridge, K.C., Flynn, F.W., Schulkin, J., Grill, H.J. (1984). Sodium depletion enhances salt palatability in rats. *Behavioral Neuroscience, 98*, 652–660.

- Bleichert, A., Behling, K., Scarperi, M. & Scarperi, S. (1973). Thermoregulatory behavior of man during rest and exercise. *Pflüers. Archive, 338*, 303–312.

- Block, N. (2006). Bodily sensations as an obstacle for representationalism. In M. Aydede (Ed.), *Pain: New essays on its nature and the methodology of its study* (pp. 611-616), Cambridge, MA: MIT Press.

- Bottenberg, E. H. (1975). Phenomenological and operational characterization of factor analytically derived dimensions of emotion. *Psychological reports, 37*, 1253–1254.

- Bradley, M. M. & Lang, P. J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology, 37*, 204–215.

- Büchel, C., Geuter, S., Sprenger, C., Eippert, F. (2014). Placebo analgesia: a predictive coding perspective. *Neuron, 81*, 1223–1239.

- Bush, L.E. (1973). Individual differences multidimensional scaling of adjectives denoting feelings. *Journal of personality and social psychology, 25*(1), 50–57.

- Byrne, A. (2001). Intentionalism Defended. *Philosophical Review, 110*, 199-239.

- Cabanac, M. (1971). Physiological role of pleasure. *Science, 173*(2), 1103–1107.

- _____. (1979). Sensory Pleasure. *The Quarterly Review of Biology*, *54* (1), 1-29.

- Cabanac, M. & Duclaux, R. (1970). Obesity Absence of satiety aversion to sucrose. *Science, 168,* 496-497.

- Cabanac, M., Stolwijk, J. & Hardy, J. (1968). Effect of temperature and pyrogens on single-unit activity in the rabbit's brain stem. *Journal of Applied Physiology*, *24*, 645-651.

- Cacioppo, J., Brentson, G., Andersen, B. (1991). Psychophysiological approaches to the evaluation of psychotherapeutic process and outcome. *Contributions from social psychophysiology. Psychological Assesment A Journal of Consulting and Clinical Psychology*, *3* (3), 321-336.

- Cacioppo, J., Berntson, G., Larsen, J., Poehlmann, K., & Ito, T. (2000). The psychophysiology of emotion. In R. Lewis & J.M. Haviland-Jones (Eds.), *The handbook of emotion* (2nd ed.), (pp. 173–191). New York: Guilford.

- Calejesan, A. A., Kim, S. J., Zhuo, M. (2000). Descending facilitatory modulation of a behavioral nociceptive response by stimulation in the adult rat anterior cingulate cortex. *European Journal of Pain*, *4*, 83–96.

- Cannon, W. (1928). *Bodily changes in pain, hunger, fear and rage: An account of recent researches into the function of emotional excitement*. New York: Appleton-Century

- Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

- Carver, C. S. (2003). Pleasure as a sign you can attend to something else: Placing positive feelings within a general model of affect. *Cognition and Emotion*, *17*, 241–261.

- Carver, C. S., & Scheier, M. F. (1999). Themes and issues in the selfregulation of behavior. In R. S. Wyer, Jr. (Ed.), *Advances in social cognition* (Vol. 12). Mahwah, NJ: Erlbaum.

- Ceunen, E., Vlaeyen, J.W. S., & Van Diest, I. (2016). On the Origin of Interoception. *Frontiers in Psychology*, *7*:*743*.

- Charland, L.C. (2005). The Heat of Emotion: Valence and the Demarcation Problem. *Journal of Consciousness Studies, 12*, 8-10.

- Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.

- _____. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral & Brain Sciences*, *36*(3), 181– 204.

- _____. (2016). *Surfing Uncertainty*. Oxford: Oxford University Press.

- Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.

- Christie, I. C., & Friedman, B. H. (2004). Autonomic specificity of discrete emotion and dimensions of affective space: A multivariate approach. *International Journal of Psychophysiology*, *51*(2), 143–153.

- Cochrane, T. (2009). Eight dimensions for the emotions. *Social Science Information*, *48* (3), 379-420.

- Colman, A. (2015). arousal. In *A Dictionary of Psychology*. Oxford University Press. Retrieved 25 Jul. 2017, from http://www.oxfordreference.com.ezproxy.is.ed.ac.uk/view/10.1093/acref/9780199657681.001.0001/acref-9780199657681-e-635.

- Colombetti, G. (2005). Appraising Valence. *Journal of Consciousness Studies*, *12*, 103–126.

- _____. (2014). *The Feeling Body: Affective Science Meets the Enactive Mind*. Cambridge, MA: MIT Press.

- Corns, J. (2014). Unpleasantness, motivational *oomph*, and painfulness. *Mind and Language*, *29*(2), 238-254.

- Craig, A.D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience*, *3*, 655–666.

- _____. (2003). Interoception: the sense of the physiological condition of the body. *Current Opinion in Neurobiology*, *13*, 500-505.

- _____. (2008). Interoception and emotion: a neuro anatomical perspective. In M.Lewis, J.M. Haviland-Jones, & L.F.Barrett (Eds.). *Handbook of Emotion* (pp. 272–292). New York,NY: Guilford Press.

- _____. (2015). *How do you feel? an interoceptive moment with your neurobiological self*. Princeton University Press.

- Critchley, H.D. (2002). Electrodermal responses: what happens in the brain. *Neuroscientist*, *8*(2), 132-142.

- Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7*, 189–195.

- Cunningham, D. J. & Cabanac, M. (1971). Evidence from behavioral thermoregulatory responses of a shift in set-point temperature related to the menstrual cycle. *Journal of Physiology*, *63*, 236-238.

- Damasio, A. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. New York: Putnam.

- Damasio A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society London B*, *351*, 1413–1420.

- _____. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace.

- _____. (2003). *Looking for Spinoza*. New York: Harcourt Inc.

- Damasio, A., Damasio, H., Tranel, D. (2013). Persistence of feelings and sentience after bilateral damage of the insula. *Cerebral Cortex*, *23*, 833–846.

- D'Andrade, R. G. (1992). Schemas and motivation. In R. G. D'Andrade & C. Strauss (Eds.), *Human motives and cultural models* (pp. 23-44). Cambridge, UK: Cambridge University Press.

- Davidson, R.J. (1993). Cerebral asymmetry and emotion: Conceptual and methodological conundrums. *Cognition and Emotion, 7*, 115–38.

- Darwin, C. (1965). *The expression of the emotions in man and animals*. Chicago: University of Chicago Press. (Original work published 1859).

- Deen, B., Pitskel, N.B., Pelphrey, K.A. (2011). Three systems of insular functional connectivity identified with cluster analysis. *Cerebral Cortex*, *21*, 1498–1506.

- Deigh, J. (2010). Concepts of emotions in modern philosophy and psychology. In Goldie P. (Ed.), *The Oxford handbook of philosophy of emotion* (pp. 17–40). Oxford, UK: Oxford University Press.

- Dennett, D. (2009). The Part of Cognitive Science That Is Philosophy. *Topics in Cognitive Science*, *1*, 231-236.

- Deonna, J., & Teroni, F. (2012). *The Emotions: A Philosophical Introduction*. London: Routledge.

- Devinsky, O., Morrell, M.J., Vogt, B.A. (1995). Contributions of anterior cingulate cortex to behaviour. *Brain*, *118*(1), pp. 279-306.

- Dewey, J. (1895). The theory of emotion. (2) The significance of emotions. *Psychological Review, 2*, 13–32.

- Diesel, D. A., Tucker, A. & Robertshaw, D. (1990). Cold-induced changes in breathing pattern as a strategy to reduce respiratory heat loss. *Journal of Applied Physiology, 69* (6), 1946-1952.

- Dretske, F. (1981). *Knowledge and the flow of information.* Cambridge: MIT Press.

- _____. (1986). Misrepresentation. In R. Bogdan (Ed.), *Belief. Form, content and function* (pp. 17-36). Oxford: Oxford University Press.

- _____. (1995). *Naturalizing the Mind*, Cambridge, MA: Bradford Books / MIT Press.

- Dunn, B.D., Galton, H.C., Morgan, R., et al. (2010). Listening to your heart: how interoception shapes emotion experience and intuitive decision making. *Psychological Science*, *21*(12), 1835–1844.

- Dunn, B.D., Stefanovitch, I., Buchan, K., Lawrence, A.D., Dalgleish, T. (2009). A reduction in positive self-judgment bias is uniquely related to the anhedonic symptoms of depression. *Behavioral Research Therapy*, *47*(5), 374-381.

- Dworkin, B. R. (1993). *Learning and physiological regulation*. Chicago: University of Chicago Press.

- _____. (2007). Interoception. In J. T. Cacciopo, L.G. Tassinary, G.G. Bernston, (Eds.). *Handbook of psychophysiology* (pp. 482–506). Cambridge: Cambridge University Press.

- Dyer, M.G. (1987). Emotions and their computations: Three computer models. *Cognition and Emotion, 1*, 323–347.

- Edworthy, J. & Waring, H. (2006). The effects of music tempo and loudness level on treadmill exercise. *Ergonomics*, *49*(15), 1597-1610.

- Ehlers, A., & Breuer, P. (1992). Increased cardiac awareness in panic disorder. *Journal of Abnormal Psychology, 101*, 371–382.

- Eimer M., Holmes A. (2007). Event-related brain potential correlates of emotional face processing. *Neuropsychologia*, *45*, 15–31.

- Fechner G.T. (1873). *Einige Ideen zur Schöpfungs- und Entwicklungsgeschichte der Organismen*. Leipzig: Breitkopf und Härtel.

- Feldman, L.A. (1995). Valence-focus and arousal-focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, *69*, 153-166.

- Frijda, N. H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

- _____. (2008). The psychologists' point of view. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 68-87). New York: Guilford Press.

- _____. (2009). Emotion experience and its varieties. *Emotion Review, 1*(3), 264-271.

- _____. (2010). Impulsive action and motivation. *Biological Psychology*, *84*, 570–9.

- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *360*, 815– 836.

- _____. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293– 301.

- _____. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.

-Friston, K.J., & Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Networks*, *22*, 1093-1104.

- Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Frontiers Human Neuroscience*, *7*, 598.

- Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*, 417–458.

- Furman D. J., Waugh C. E., Bhattacharjee K., Thompson R. J., & Gotlib I. H. (2013). Interoceptive awareness, positive affect, and decision making in major depressive disorder. *Journal Affective Disorders*, *151*, 780–785.

- Galati, D., Sini, B., Tinti, C. & Testa, S. (2008). The lexicon of emotion in the neo-Latin languages. *Social science information sur les sciences sociales*, *47*(2), 205–220.

- Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., Critchley, H. D. (2015). Knowing your own heart: distinguishing interoceptive accuracy from interoceptive awareness. Biological Psychology, *104*, 65–74.

- Gordon, R. (1987). *The structure of emotions.* Cambridge, UK: Cambridge University Press.

- Gottfried, J.A., O'Doherty, J., Dolan, R.J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science*, *301*, 1104–1107.

- Greene, J.D., R.B. Sommerville, L.E. Nystrom, J.M. Darley, & Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment, *Science*, 293, 2105–2108.

- Green, O. H. (1992). *The Emotions: A Philosophical Theory*. Dordrecht: Kluwer Academic Publishers.

- Grill, J. D., & Coghill, R. C. (2002). Transient analgesia evoked by noxious stimulus offset. Journal of Neurophysiology, *87*(4), 2205–2208.

- Gross J. J. & Barrett L. F. (2011). Emotion generation and emotion regulation: one or two depends on your point of view. *Emotion. Review, 3*, 8–16.

- Gu, X. & FitzGerald, T.H.B. (2014). Interoceptive inference: homeostasis and decision-making. *Trends in Cognitive Science*, *18*, 269-270.

- Guhn M., Hamm A., Zentner M. (2007). Physiological and musico-acoustic correlates of the chill response. *Music Perception*, *24*, 473–483.

- Habibi, A. & Damasio, A. (2014). Music, feelings and the human brain. *Psychomusicology: Music, Mind and Brain*, *24*, 92–102.

- Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, *108*, 814–34.

- Harman, G. (1990). The Intrinsic Quality of Experience. In J. Tomberlin (Ed.), *Philosophical Perspectives* (Volume 4) (pp. 31-52), Atascadero, CA: Ridgeview Publishing Company.

- Harshaw, C. (2015). Interoceptive dysfunction: toward an integrated framework for understanding somatic and affective disturbance in depression. *Psychological Bulletin*, *141*, 311–363.

- Herrmann, D. J. & Raybeck, D. (1981). Similarities and differences in meaning in six cultures. *Journal of cross-cultural psychology*, *12*, 194–206.

- Higgins, E. T. (2007). Value. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 454–472). New York: Guilford Press.

- Hohwy, J. (2011). Phenomenal variability and introspective reliability. *Mind & Language*, *26*(3), 261-286.

- _____. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3*, 96.

- _____. (2013). *The predictive mind*. Oxford: Oxford University Press.

- Holle H., Warne K., Seth A. K., Critchley H. D., Ward J. (2012). The neural basis of contagious itch and why some people are more prone to it. *Proceedings of the National Academy of Science U.S.A.*, *109*, 19816–19821.

- Izard, C.E. (1991). *The Psychology of Emotions*. New York: Plenum.

- Jacobs, K.M., Mark, G.P. & Scott, T.R. (1988). Taste responses in the nucleus tractus solitarius of sodium-deprived rats. *Journal of Physiology*, *406*, 393-410.

- James, W. (1884). What is an emotion? *Mind*, *9*, 188-205.

- Jehee, J. F., & Ballard, D. H. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS Computational Biology*, *5*(5), e1000373.

- Jepma, M., & Wager, T. D. (2013). Multiple potential mechanisms for context effects on pain. *Pain*, *154*, 629-631.

- Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology*, *9*(6), e1003094.

- Johnson, A. K. (2007). The sensory psychobiology of thirst and salt appetite. *Medicine and Science in Sports and Exercise*, *39*, 1388–1400.

- Karageorghis, C., Jones, L., Low, D. (2006). Relationship between Exercise Heart Rate and Music Tempo Preference. *Research Quarterly for Exercise and Sport*, *77* (2), 240-250.

- Katz, L.D. (2016). Preventing bullying. Retrieved from: <https://plato.stanford.edu/archives/win2016/entries/pleasure/>.

- - Khalsa, S.S., Rudrauf, D., Feinstein, J.S., & Tranel, D. (2009). The pathways of interoceptive awareness. *Nature Neuroscience*, *12*(12), 1494–1496.

- Konečni, V. J., Wanic, R. A., & Brown, A. (2007). Emotional and aesthetic antecedents and consequences of music-induced thrills. *The American Journal of Psychology*, *120*(4), 619– 643.

- Kringelbach M.L. & Berridge, K.C. (2010). *Pleasures of the brain*. New York: Oxford University Press.

- Kuppens, P., Tuerlinckx, F., Russell, J. A., and Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, *139*, 917– 940.

- LaBar, K.S., Gitelman, D.R., Parrish, T.B., Kim, Y.H., Nobre, A.C., Mesulam, M.M. (2001). Hunger selectively modulates corticolimbic activation to food stimuli in humans. *Behavioral Neuroscience, 115*, 493–500.

- Lacey, J. I. (1956). The evaluation of autonomic responses: Toward a general solution. *Annals of the New York Academy of Sciences, 67*, 123–164.

- _____. (1967). Somatic response patterning and stress: Some revisions of activation theory. In M. H. Appley & R. Trumbull (Eds.), *Psychological stress: Issues in research* (pp. 4–44). New York: Appleton-Century-Crofts.

- Lane, R.D., Reiman, E.M., Ahern, G.L., Schwartz, G.E., Davidson, R.J. (1997). Neuroanatomical correlates of happiness, sadness, and disgust. *American Journal of Psychiatry, 154*, 926–933.

- Lang, P. J. Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, *30*, 261–273.

- Larsen, J.T., McGraw, A.P., & Cacioppo, J.T. (2001). Can people feel happy and sad at the same time? *Journal of Personality and Social Psychology, 81*, 684–696.

- Lazarus, R. (1991). *Emotion and adaptation*. New York: Oxford University Press.

- LeDoux, J. E. (1996). *The emotional brain.* New York: Simon and Schuster.

- Leknes, S., & Bastian, B. (2014). The Benefits of Pain. *Review of Philosophy and Psychology*, *5*(1), 57-70.

- Levenson, R., Ekman, P., & Friesen, W. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, *27*, 363-384.

- Lewin, K. (1935). *A Dynamic Theory of Personality: Selected Papers* (trans. D.K. Adams & K.E. Zener). New York: McGraw-Hill.

- Lovero, K. L., Simmons, A. N., Aron, J. L., and Paulus, M. P. (2009). Anterior insular cortex anticipates impending stimulus significance. *Neuroimage*, *45*, 976–983.

- Lutz, C. (1982). The domain of emotion words on Ifaluk. *American ethnologist*, *9*, 113–128.

- Lycan, W.G. (1987). *Consciousness*. Cambridge, MA: Bradford Books / MIT Press.

- _____. (1996). *Consciousness and Experience*. Cambridge, MA: MIT Press.

- MacLean, P. D. (1993). Cerebral evolution of emotion. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 67-83). New York: Guilford Press.

- Macpherson, F. (2011). Taxonomising the senses. *Philosophical Studies*, *153*(1), pp. 123-142.

- Matthen, M. (2015). Individuating the Senses. In M. Matthen (ed.) *Oxford Handbook of the Philosophy of Perception* (pp. 567-586). Oxford: Clarendon Press.

- Marks, J. (1982). A theory of emotions. *Philosophical Studies*, *42*, 227–42.

- Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, *23*, 209-237.

- McCaughey, S.A., Forestell, C.A., Tordoff, M.G. (2005). Calcium deprivation increases the palatability of calcium solution in rats. *Physiology and Behavior, 84*, 335–342.

- McCaughey, S.A., & Tordoff, M.G. (2002). Magnesium appetite in the rat. *Appetite, 38*, 29–38.

- Medford, N., & Critchley, H. D. (2010). Conjoint activity of anterior insular and anterior cingulate cortex: Awareness and response. *Brain Structure and Function, 214*, 535– 549.

- Menon V. (2015) Salience Network. In: Arthur W. Toga (Ed.). *Brain Mapping: An Encyclopedic Reference, Vol. 2* (pp. 597-611). Academic Press: Elsevier.

- Millikan, R. (1996). Pushmepullyou Representations. In L. May, M. Friedman and A. Clark (Eds). *Mind and Morals* (pp. 145–62). Cambridge, MA: MIT Press.

- _____. (2004). *Varieties of Meaning: The 2002 Jean Nicod Lectures*. Cambridge, MA: MIT Press.

- Murphy, F.C., Nimmo-Smith, I., & Lawrence, A.D. (2003). Functional neuroanatomy of emotions: a meta-analysis. *Cognitive and Affective Behavioral Neuroscience, 3*, 207–233.

- Naccache, L., & Dehaene, S. (2001). Unconscious semantic priming extends to novel unseen stimuli. *Cognition, 80*(3), 215–229.

- Naqvi, N.H., & Bechara, A. (2010). The insula and drug addiction: an interoceptive view of pleasure, urges, and decision-making. *Brain Structure and Function, 214*, 435–450.

- Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. New York: Oxford University Press.

- Noel, X., Brevers, D., & Bechara, A. (2013). A triadic neurocognitive approach to addiction for clinical interventions. *Frontiers in Psychiatry, 4*, 179.

- Nussbaum, M. (2001). *Upheavals of Thought: The Intelligence of Emotions*. Cambridge: Cambridge University Press.

- Nyklicek, I., Thayer, J. F., & van Doornen, L. J. P. (1997). Cardiorespiratory differentiation of musically induced emotions. *Journal of Psychophysiology*, *11*, 304–321.

- Ohman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General, 130,* 466-478.

- Ondobaka, S., Kilner, J., Friston, K. (2017). The role of interoceptive inference in theory of mind. *Brain and Cognition*, *112*, 64–68.

- Ortony, A., Clore, G.L. & Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

- O'Shaughnessy, B. (1980). *The Will*. Cambridge: Cambridge University Press

- Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York, NY: Oxford University Press.

- _____. (2005). On the embodied neural nature of core emotional affects. *Journal of Consciousness Studies*, *12* (8–10), 158–184.

- Palermo R. & Rhodes G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia*, *45*, 75–92.

- Paulus, M.P., Flagan, T., Simmons, A.N., Gillis, K., Kotturi, S., Thom, N., Johnson, D.C., Van Orden, K.F., Davenport, P.W., Swain, J.L. (2012). Subjecting elite athletes to inspiratory breathing load reveals behavioral and neural signatures of optimal performers in extreme environments. *PLoS One*, *7*, e29394.

- Paulus, M.P., Simmons, A.N., Fitzpatrick, S.N., Potterat, E.G., Van Orden, K.F., Bauman, J., Swain, J.L. (2010). Differential brain activation to angry faces by elite warfighters: neural processing evidence for enhanced threat detection. *PLoS One*, *5*, e10096.

- Pessoa L. (2008). On the relationship between emotion and cognition. Nature Reviews Neuroscience, *9*, 148–158.

- Pezzulo, G. (2014). Why do you fear the Bogeyman? An embodied predictive coding model of perceptual inference. *Cognitive, Affective, and Behavioral Neuroscience, 14*, 902–911.

- Phan, K.L., Wager, T., Taylor, S.F., Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage, 16*, 331-348.

- Picard, R.W. (1997). *Affective Computing.* Cambridge, MA: MIT Press.

- Picciuto, V., & Carruthers, P. (2014). Inner sense. In D. Stokes, M. Matthen, and S. Biggs (Eds.), *Perception and its Modalities* (pp. 277-294). New York: Oxford University Press.

- Pizarro, D. (2000). Nothing more than feelings?: The role of emotions in moral judgment. *Journal for the Theory of Social Behaviour*, *30*, 355-375.

- Ploghaus, A., Tracey, I., Gati, J.S., Clare, S., Menon, R.S., Matthews, P.M., Rawlins, J.N. (1999). Dissociating pain from its anticipation in the human brain. *Science*, *284*, 1979–1981.

- Prinz, J. (2002). *Furnishing the Mind: Concepts and Their Perceptual Basis*, Cambridge, MA.: MIT Press.

- _____. (2004). *Gut Reactions: A perceptual theory of emotion*. Oxford: Oxford University Press.

-_____. (2007). *The Emotional Construction of Morals*. New York, NY: Oxford University Press.

- _____. (2010). For valence. *Emotion Review, 2*, 5–13.

- _____. (2012). *The Conscious Brain*. New York: Oxford University Press.

- Quattrocki, E. & Friston, K. (2014). Autism, oxytocin and interoception. *Neuroscience and Biobehavavioral Reviews, 47*, 410–430.

- Quigley, K. S., & Barrett, L. F. (2014). Is there consistency and specificity of autonomic changes during emotional episodes? Guidance from the Conceptual Act Theory and psychophysiology. *Biological Psychology*, *98*, 82-94.

- Rao, R.P., 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.

- Ritchie, J.B. & Carruthers, P. (2015). The bodily senses. In M. Matthen (Ed.), *The Oxford Handbook of the Philosophy of Perception* (pp. 353-371). Oxford University Press.

- Rolls B. J., Rowe E. A., Rolls E. T., Kingston B., Megson A., Gunary R. (1981). Variety in a meal enhances food intake in man. *Physiology and Behavior*, *26*, 215–221.

- Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*, 329–359.

- Rozin, P. (2003). Introduction: Evolutionary and cultural perspectives on affect. In R.J. Davidson, K.R. Scherer & H.H. Goldsmith (Eds.). *Handbook of Affective Sciences* (pp.839-852). Oxford: Oxford University Press.

- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39,* 1161–1178.

- _____. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*, 145–172.

- Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*, 805-819.

- Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

- Sachs, M. E., Damasio, A., & Habibi, A. (2015). The pleasures of sad music: a systematic review. *Frontiers in Human Neuroscience*, *9*, 404.

- Sarter, M., Bruno, J.P., Givens, B. (2003). Attentional functions of cortical cholinergic inputs: what does it mean for learning and memory? *Neurobiological Learning Memory, 80,* 245–256.

- Scarantino, A. (2009). Core affect and natural affective kinds. *Philosophy of Science, 76*, 940–957.

- _____. (2014). The motivational theory of emotions. In D. Jacobson & J. D'Arms (Eds.), *Moral Psychology and Human Agency* (pp. 156–185), Oxford: Oxford University Press.

- Scarantino, A., & Griffiths, P. (2011). Don't give up on basic emotions. *Emotion Review, 3,* 444–454.

- Schachter, S., & Singer, C. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69,* 379-399.

- Schandry, R. (1981). Heart beat perception and emotional experience. *Psychophysiology, 18,* 483–488.

- Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 293–317). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

- Schneirla, T.C. (1959). An evolutionary and developmental theory of bi-phasic processes underlying approach and withdrawal. In M.R. Jones (Ed.), *Nebraska Symposium on Motivation*. Lincoln: University of Nebraska Press.

- Schroeder, T. (2001). Pleasure, Displeasure, and Representation. *Canadian Journal of Philosophy, 31,* 507–530.

- _____. (2004). *Three Faces of Desire,* New York: Oxford University Press.

- _____. (2006). An Unexpected Pleasure. In L. Faucher & C. Tappolet (Eds.), *Canadian Journal of Philosophy*, supp. Vol. 32, *The Modularity of Emotions* (pp. 255–272).

- Schultz W., Dayan P., Montague R. R. A. (1997). Neural substrate of prediction and reward. *Science, 275*, 1593–1599.

- Seth, A. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Science, 17*(11), 565-573.

- _____. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synaesthesia. *Cognitive Neuroscience, 5*(2), 97-118.

-_____. (2015a). The cybernetic Bayesian brain: from interoceptive inference to sensorimotor contingencies. In T. Metzinger& J.M. Windt (Eds.), *Open MIND*. Frankfurt, Germany: MIND group.

- _____. (2015b) Inference to the best prediction. In Open MIND (eds) T. Metzinger, J.M. Windt, (1–8), Frankfurt, Germany: MIND Group.

- Seth, A. & Critchley, H. (2013) Extending predictive processing to the body: Emotion as interoceptive inference. *Behavioral & Brain Sciences, 36*, 227–228.

-Seth, A. & Friston, K. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B, 3*71, 20160007.

- Seth, A., Suzuki, K., & Critchley, H. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, *2*, 395.

- Seymour, B., O'Doherty, J.P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, *8*, 1234–1240.

- Seymour, B., O' Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston K.J., Frackowiak, R.S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*, 664–667.

- Shaver, P. R., Wu, S., & Schwartz, J. C. (1992). Cross cultural similarities and differences in emotion and its representation: A prototype approach. In M. S. Clark (Ed.), *Review of personality and social psychology* (Vol. 13, pp. 175-212). Newbury Park, CA Sage.

- Sherrington, C.S. (1948). *The Integrative Action of the Nervous System*. Cambridge: University Press.

- Singer, T., Critchley, H. D., and Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Science*, *13*, 334–340.

- Sizer, L. (2000). Towards a Computational Theory of Mood. *British Journal for the Philosophy of Science*, *51* (4), 743-770.

- Skinner, B. F. (1953). *Science and Human Behavior*. New York: Macmillan.

- Solomon, R. C. (1976). *The Passions*. New York: Doubleday

- _____. (2001). Against valence. In R. C. Solomon, *Not passion's slave* (pp. 135–147). New York: Oxford University Press.

- _____. (2001). *Not Passion's Slave*. Oxford: Oxford University Press.

- Srinivasan, M.V., Laughlin, S.B., & Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society London B Biological Sciences*, *216*, 427–459.

- Stemmler, G., Aue, T., & Wacker, J. (2007). Anger and fear: Separable effects of emotion and motivational direction on somatovisceral responses. *International Journal of Psychophysiology*, *66*(2), 141–153.

- Sterling P. (2004). Principles of allostasis. In J. Schulkin (Ed.), *Allostasis, Homeostasis, and the Costs of Adaptation* (pp. 17–64), Cambridge: Cambridge University Press.

- Tolman, E.C. (1932). *Purposive Behavior in Animals and Man*. New York: Century.

-Trappe, H. (2010). Effects of music on cardiovascular system and cardiovascular health. *Heart*, *96* (23), 1868-1871.

- Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: The MIT Press.

- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

- Veldhuizen, M.G., Shepard, T.G., Wang, M.F., Marks, L.E. (2010). Coactivation of gustatory and olfactory signals in flavor perception. *Chemical Senses*, *35*, 121–133.

- Vuilleumier, P. & Pourtois, G. (2007). Distributed and interactive brain mechanisms during emotion face perception: evidence from functional neuroimaging. *Neuropsychologia*, *45*, 174–194.

- Vytal, K. & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: A voxelbased meta-analysis. *Journal of Cognitive Neuroscience, 22,* 2864-2885.

- Watson, J. B. (1919). *Psychology from the Standpoint of a Behaviorist*. Philadelphia: Lippincott.

- Wen, L., Moallem, I., Paller, K. A., & Gottfried, J. A. (2007). Subliminal smells can guide social preferences. *Psychological Science, 18,* 1044-1049.

- Winkielman, P., & Berridge, K. C. (2004). Unconscious emotion. *Current Directions in Psychological Science, 13,* 120-123.

- Winkielman, P., Berridge, K. C., & Wilbarger, J. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality and Social Psychology Bulletin*, *1*, 121–135.

- Wiech, K., Ploner, M., Tracey, I. (2008). Neurocognitive aspects of pain perception. *Trends in Cognitive Science*, *12*, 306–313.

- Wiens S., Mezzacappa, E.S., Katkin, E.S. (2000). Heartbeat detection and the experience of emotions. *Cognition and Emotion*, *14*(3), 417–427.

- Wilson-Mendenhall, C.D., Barret, R. F., Simmons, and Barsalou, L.W. (2011). Grounding emotion in situated conceptualization. *Neuropsychologia*, *49*, 1105–1127.

- Xiang, T., Lohrenz, T., Montague, P.R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, *33*, 1099–1110.